# INTERSPEECH 2018

September 2-6 | Hyderabad, India

Speech Research for Emerging Markets in Multilingual Societies

# Book of Abstracts

Copyright 2018 International Speech Communication Association (ISCA)
http://www.isca-speech.org
secretariat@isca-speech.org

All rights reserved
ISSN: 2308-457X Proceedings USB
ISSN: 1990-9772 Proceedings in ISCA Archive

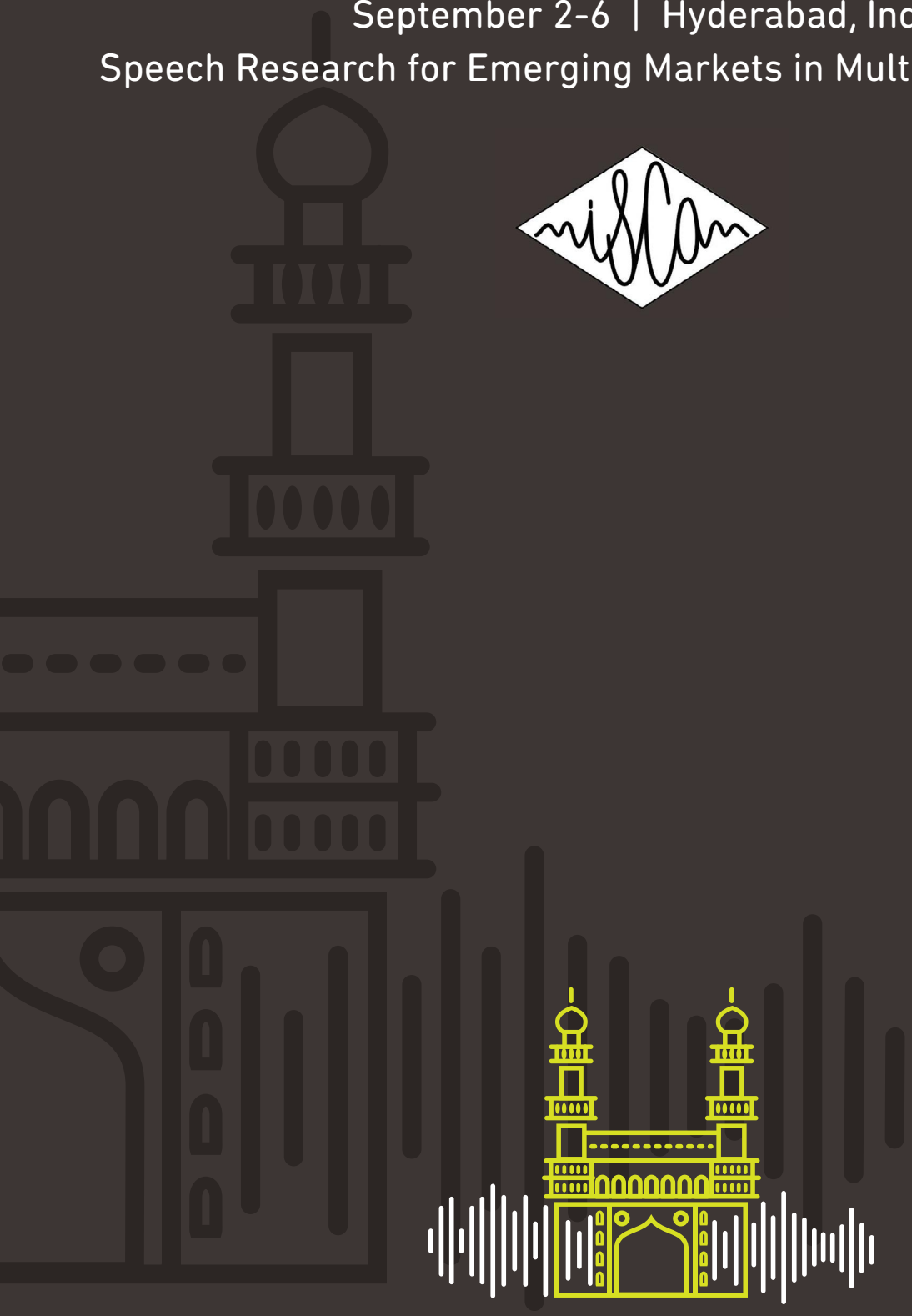
Editors: B. Yegnanarayana, C. Chandra Sekhar, Shrikanth Narayanan, S. Umesh, S. R. M. Prasanna,
Hema A. Murthy, Preeti Rao, Paavo Alku and Prasanta Kumar Ghosh

Design and Production of Abstracts and Proceedings USB by Thefool.in

# INTERSPEECH 2018

September 2-6 | Hyderabad, India

Speech Research for Emerging Markets in Multilingual Societies



## Book of Abstracts

# Table of Contents


1. Exhibitors .....23
2. ISCA .....24
3. Message from ISCA President .....25
4. Message from General Chair and Co-Chairs.....27
5. Message from Technical Program Chairs.....29
6. Speech Research in India.....31
7. Theme Related Events .....33
8. Welcome to Hyderabad.....34
9. General Information.....38
10. Social Events.....42
11. INTERSPEECH 2018 Organizing Team.....45
12. Future INTERSPEECH Conferences.....55
13. Satellite Workshops and Events.....56
14. Tutorials.....59
15. ISCA Medalist.....73
16. Plenary Speakers.....75
17. Perspective Talks.....79
18. Special Sessions.....83
19. Show & Tell.....89
20. Special Events.....90
21. ISCA Meetings.....92
22. ISCA-SAC Events.....94
23. Awards.....97
24. Session Index.....100
25. Abstracts.....104
26. Session Chair Index.....261
27. Author Index.....262
28. Venue Maps.....275
29. Daily Schedule.....277
30. Program at a Glance.....282

## FOUNDING SPONSOR

amazon alexa

## DIAMOND SPONSORS

SAMSUNG

uniphore

JD.COM

mi

## SILVER SPONSORS

Adobe

NVIDIA

## SUPPORTERS

RAMAN IT

idiap RESEARCH INSTITUTE

Sanghavi Impex

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

भारतीय प्रौद्योगिकी संस्थान तिरुपति IIT TIRUPATI

NUANCE

Geethanjali College of Engineering and Technology
An Autonomous Institution
Approved by AICTE and Permanently Affiliated to JNTUH
ISO 9000:2008 Certified

## PLATINUM SPONSORS

Microsoft

Baidu 百度

facebook

## GOLD SPONSORS

Google

QUALCOMM

Sogou 搜狗

DRDO
DEFENCE R&D ORGANISATION
MINISTRY OF DEFENCE

YAHOO! JAPAN

IBM Research AI

## BRONZE SPONSORS

GoVivace
at the pace of technology

VIGNAN'S
Foundation for Science, Technology & Research
(Deemed to be UNIVERSITY)
-Estd. u/s 3 of UGC Act 1956

सत्यमेव जयते
Ministry of Electronics and Information Technology
Government of India

## SUPPORTERS

Sundar Associates

ALPS
World-Class Hearing Aids

Indian Institute of Technology Mandi

DECIBELLS acoustics

ALL INDIA INSTITUTE OF SPEECH AND HEARING
MYSORE

INDIAN INSTITUTE OF SCIENCE
भारतीय विज्ञान संस्थान

INDIAN INSTITUTE OF TECHNOLOGY MADRAS

VISHNU

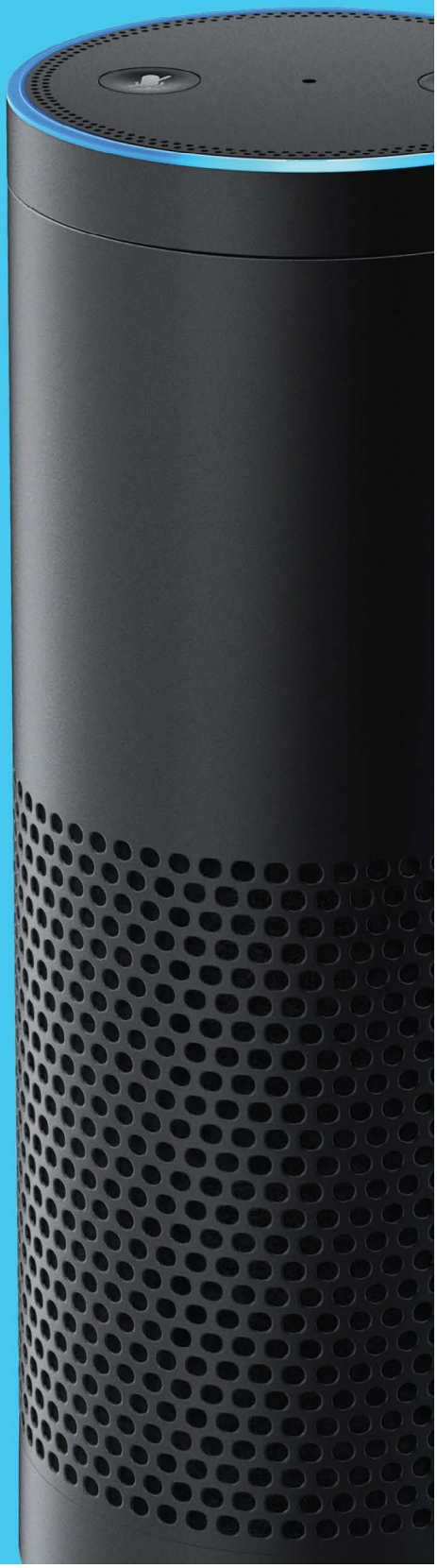
BVRIT
BV Raju Institute of Technology
ESTD 1997
(UGC-Autonomous)

UNIVERSITY of WASHINGTON

amazon alexa

# It's only impossible until it works

**Learn more: amazon.jobs/interspeech2018**
**Email interspeech2018@amazon.com**



Amazon is an Equal Opportunity Employer

Speech and Language is a key part of the Mobile First, Cloud First AI technology that is shaping the future of computing and user experiences. At Microsoft we have openings for Scientists and Engineers who are passionate about the opportunity to work on this next generation of productivity technologies.

We create value for customers with productivity enhancing capabilities that understand *their intent* in *their context*, enabling them to know more, do more and be more. We are the ears and eyes of Cortana that understands our users' intent expressed in spoken & written natural language by engaging them in natural, and meaningful conversation.

To do this we are developing new technologies to make speech recognition, intent understanding, user personalization, dialog modeling, and task definition more scalable than ever before.

The contacts for our teams will help you explore just a few of our opportunities. We welcome inquiries from people with all levels of experience and are particularly interested in people with Machine Learning, Big Data, Language Understanding and Speech backgrounds.

| | | |
|---|---|---|
| Speech & Language | General Inquiries<br>ML & Speech Modeling<br>Dialogue | Xuedong.Huang@microsoft.com<br>Li.Jiang@microsoft.com<br>Elsie.Nallipogu@microsoft.com |
| Microsoft China | TTS / Agents | Michael Zeng nzeng@microsoft.com |
| Microsoft Research Asia | All Inquiries | Frank Soong frankkps@microsoft.com |
| Microsoft India | General Inquiries<br>Speech Modeling | Niranjan Nayak niranjan@microsoft.com<br>Rupesh.mehta@microsoft.com |

Meet us in the exhibit hall at Booth L1 and visit our Interspeech Web at **aka.ms/interspeech-2018**

Microsoft

Baidu 百度
THE WORLD'S LEADING
CHINESE SPEECH RECOGNITION TECHNOLOGY
yuyin.baidu.com
Phoneme based
context-dependent modeling
Advanced deep LSTM models for online search
with large amount of training data
Bidirectional deep LSTM for accurate duration,
F0 and spectrum perdition for text to speech
Discriminative speaker code for
DNN adaptation
Fast one pass decoder
Baidu
mobile search
Baidu IME
Baidu Map
Baidu Translate
Baidu Speech

Facebook is pleased to be a platinum sponsor of Interspeech 2018

# AI is foundational to everything we do at Facebook

AI makes existing products better, creates entirely new experiences, and is fundamental to keeping people safe on our platform.

Through our family of apps and services, we're building a different kind of company that connects billions of people around the world, gives them ways to share what matters most to them, and helps bring people closer together.

To learn more about opportunities within Speech or NLP at Facebook, please contact: **interspeech2018@fb.com**

facebook Artificial Intelligence

VOICE INTELLIGENCE - Come & Experience
Bixby | SAMSUNG
STALL
L23
Smart Access to Content
Entertain
Assist
Know the User
Interact Naturally
Intelligent
Proactive
Interactive
Context Aware (location, time, purpose)
Understand Users world
Predict and Forecast User Needs
Proactive Actions
Spontaneous / real time
Deliver Service Seamlessly
Conversational in nature
More Human, less binary
Friendly and Assistive
Core Technology
Far Field ASR
Personal ASR
Multi Modal Fusion
Contextual NLU
On-Device Voice Solution
Embedded ASR
Embedded NLU & Dialog Mgmt.
Appless Service Integration
Open Ecosystem (Bixby 2.0)
Intelligent Convergence
Voice on DA, VD, Car & IoT Devices
Voice Extended Convergence
Strengthen Data Analysis
Positions open for Speech recognition experts/Data Scientist/ML Experts. Please contact careers.srib@samsung.com
#2870, Phoenix Building, Bagmane ConstellationBusiness Park, ORR, Doddanekundi Circle, Bengaluru 560037
https://research.samsung.com/sri-b

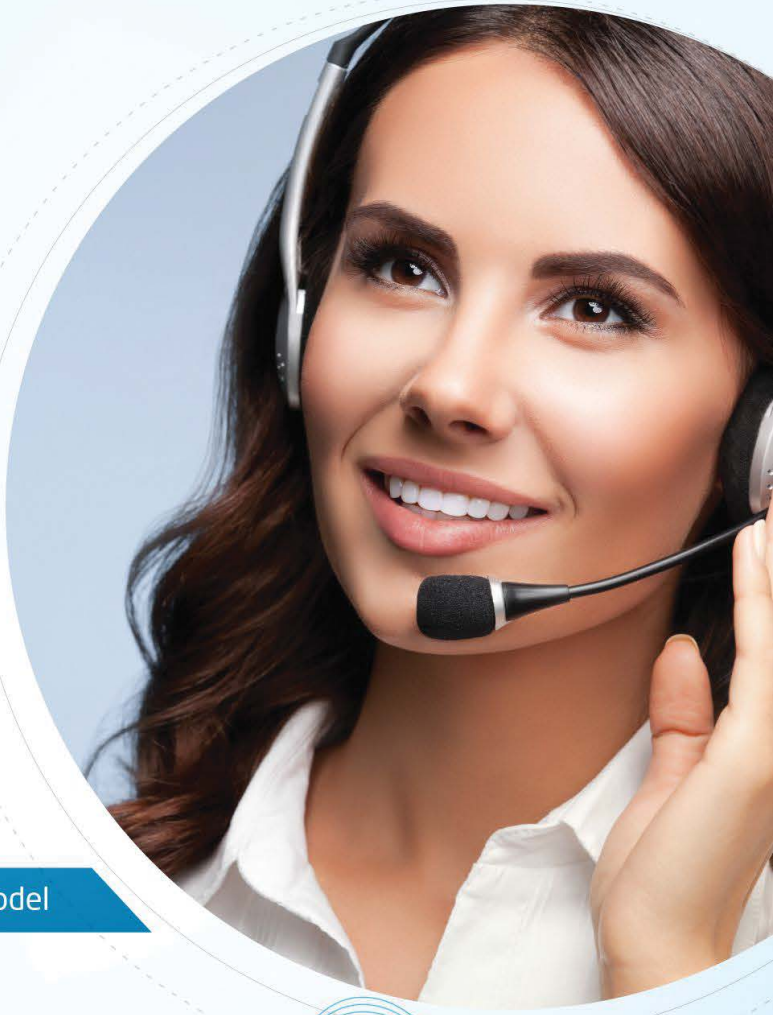
uniphore
Speech Recognition Solutions for Enterprises

auMina™

# SPEECH ANALYTICS

for better business outcomes

- Enhance customer experience, reduce churn
- Increase revenue and optimize ROI
- Improve regulatory compliance
- Automate agent quality monitoring & scoring



100+ Global Languages • SaaS Based Model

Call or write to us for

**A Free Customized Proof of Concept (POC) covering 3 KPIs, 1 Language and 50 hours of Speech Data.**

**GET A POC FOR FREE!**

Write in to **susheel_ext@uniphore.com**

**Uniphore Software Systems**

| Singapore | Bengaluru, India | Chennai, India | USA |
|---|---|---|---|
| +65 65 977044 | +91 80 67276400 | +91 44 30655800 | +1 650 617 3288 |

JD.COM

JD.com

China's Largest Retailer
A Global Fortune 500 Company

JD AI Research

Focusing on cutting edge innovation in AI
Leveraging AI to enhance several facets of JD's business
Offices in China, the US and Europe to attract leading global talent
Collaboration with world-leading academic institutions

Machine Learning Lab

NLP and Speech Lab

Computer Vision Lab

WE ARE HIRING!

contact us: airecruiting@jd.com

# ALWAYS BELIEVE THAT SOMETHING WONDERFUL IS ABOUT TO HAPPEN

Xiaomi (1810.HK) was founded in 2010 by serial entrepreneur Lei Jun based on the vision “innovation for everyone”.
We believe that high-quality products built with cutting-edge technology should be made accessible to everyone.
We create remarkable hardware, software, and Internet services for and with the help of our Mi fans. We incorporate their feedback into our product range, which currently includes Mi and Redmi smartphones, Mi TVs and Mi Ecosystem products including smart home products, wearables and other accessories.
With presence in over 30 countries and regions, Xiaomi is expanding its footprint across the world to become a global brand.

Over a thousand of engineers in Xiaomi are currently working on cloud computing, distributed storage, big data, search, recommendation and AI. Based on Xiaomi’s deep learning platform, we developed advanced technologies of image, speech, natural language understanding and etc.
Speech team of Xiaomi enables the Mi phones, Mi TVs and Ecosystem products to listen and speak to the users. Every day, our users wake up the intelligent assistant “Xiaoai”, experience the voice interaction, access entertainment and acquire information services.

Google

QUALCOMM®

# Application

## Speech Transcription

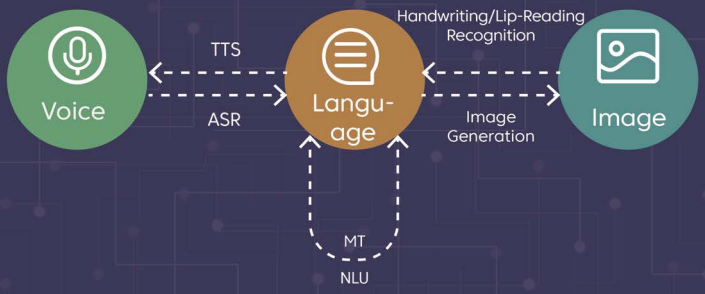
## Speech Translation

| Application Scenarios | |
|---|---|
| Traveling Aboard | International Communication |
| Simultaneous Interpretation | Video Caption |

## Speech Interaction

# Sogou Zhiyin

**voice-based multi-modality natural interaction technology**




## ASR

- Most popular voice input method in China
- Average 400 million voice input requests per day
- 4,629 people are using voice input per second

## Lip-reading Recognition

- The first one to publicly demonstrate Chinese lip-reading system
- Accuracy rate of 90% in vertical scenes

## TTS

- Winner of two subtasks: Speech Pauses and Word Error Rate in Blizzard Challenge 2018
- Technology pioneer of emotion transfer and virtual anchor

## MT

- Winner of Chinese-English translation in WMT2017
- Average 150 million translation requests per day

# Yahoo! JAPAN is Hiring !!

Yahoo! JAPAN is currently looking for Researchers and Engineers. Our mission is to develop speech technologies and offer them for more than 100 internal services. Our big data is appropriate for your brand-new machine learning ideas.

Yahoo! JAPAN | Voice Assistant | Transit | Map | Shopping | Car Navigation

**[Job Description]**

- R&D for state-of-the-art Japanese speech recognition engine
  - Signal processing algorithms for robustness
  - Machine learning algorithms using big data
- Collaborate with various Web Service Teams in Yahoo! JAPAN

**[Qualifications]**

- MS degree in a related technical field or equivalent practical experience
- Experience in speech related technologies:
  - Acoustic modeling, language modeling, ASR decoder (search algorithm)
  - Front-end technologies such as denoising and dereverberation
- Excellent oral and written communication skills in English (Japanese skills are plus)
- Solid programming skills such as C/C++, Matlab, Python

**[Web page]**

https://about.yahoo.co.jp/hr/en/

**[Contact]**

speech-team-hr@mail.yahoo.co.jp

Photo by Aflo

# IBM **Research**

With more than 3,000 researchers, in 12 labs located across six continents, IBM Research is one of the world's largest and most influential corporate research labs.

## Research areas

**Natural language processing**
Machine translation, large scale information extraction, cross-document co-reference, question-answering

**Computer vision and multimedia**
Video content understanding, cloud-based vision services, multimodal comprehension: machine vision

**Machine learning: AI Foundations**
Deep learning, reinforcement learning, brain-inspired, learning to learn, interpretability, optimization, learning with memories

**Knowledge representation and reasoning**
Automated knowledge base creation, symbolic and trainable logic, reusable representations, document understanding

**Novel systems**
Neuromorphic, quantum computing

## At IBM Research, we invent things that matter to the world.

**1954** Mark I and Mark II Automatic Language Translators

**1959** Checkers Player

**1972** Probabilistic Speech Recognition

**1984** First real time demonstration of a Large Vocabulary Speech Recognition Dictation System

**1988** Statistical Translation Between Languages

**1992** TD-Gammon

**1995** Image Retrieval

**1997** Deep Blue Chess

**2000** Statistical machine translation

**2002** BLEU metric

**2011** IBM® Watson™ Jeopardy!

| 6 | 6 | 3 | 5 | #1 | 10 | 1 | 1 |
|---|---|---|---|---|---|---|---|
| Nobel Prizes | Turing Awards | Kavli Prizes | National Medals of Science | Patent Recipient for 24 Years | National Medals of Technology | Presidential Medal of Freedom | Kyoto Prize |

ibm.com/research

DRDO
DEFENCE R&D ORGANISATION
बलस्य मूलं विज्ञानम्
MINISTRY OF DEFENCE
DRDO

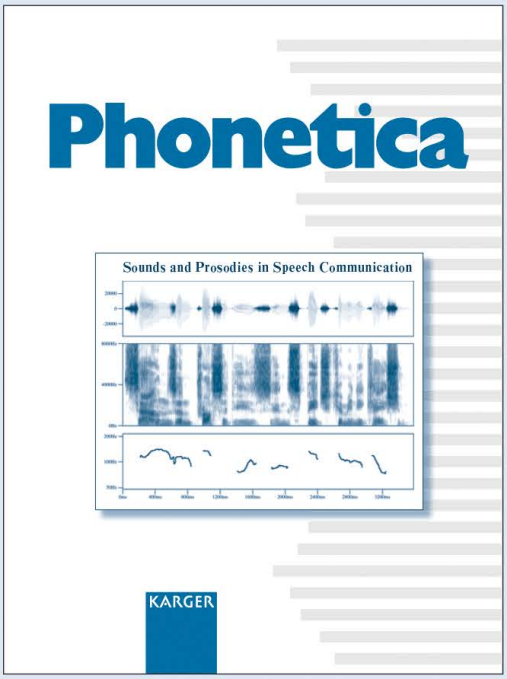
*An interdisciplinary forum for phonetic science research and theory*


## Editor

**C.T. Best**
Bankstown NSW (Australia)

### Associate Editors

**S. Frota**
Lisbon (Portugal)
**W. Gu**
Nanjing (China)
**R. Hayes-Harb**
Salt Lake City, UT (USA)
**A. Jongman**
Lawrence, KS (USA)
**G. Khattab**
Newcastle upon Tyne
(United Kingdom)
**A. Kochetov**
Toronto, ON (Canada)
**M. Pouplier**
Munich (Germany)
**R. Smith**
Glasgow
(United Kingdom)
**P.C.M. Wong**
Hong Kong (China)


### Aims and Scope

Contemporary interdisciplinary research on phonetics employs a wide range of approaches, from instrumental measures to perceptual and neurocognitive procedures, to computational modelling, for investigating the properties and principles of phonetics in communicative settings across the world's languages. It also ranges across styles, types of language users, and communicative modalities (speech, sign, song). Phonetica is an international forum for phonetic science that covers all aspects of the subject matter, from phonetic and phonological descriptions, to articulatory and signal analytic measures of production, to perception, acquisition, and phonetic variation and change. Phonetica thus provides a platform for a comprehensive understanding of producer-perceiver interaction across languages and dialects, and of learning throughout the lifespan and across contexts. Papers published in this journal report expert original work that deals both with theoretical issues, new empirical data, and innovative methods and applications that help to advance the field.

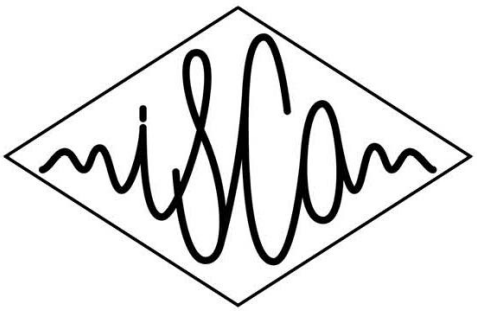

KARGER

Find subscription information and guidelines for authors on
**www.karger.com/pho**

K118252

# EXHIBITORS

Amazon Alexa
Appen
Apple
Behavioral Signal Technologies
Beijing Magic Data Technology Co. Ltd.
Brno University of Technology
Baidu Research
Cobalt Speech and Language
CVTE
Datatang
DefinedCrowd
Dialpad
Facebook
Google
INTERSPEECH 2019
INTERSPEECH 2020
JD.com
Lionbridge
Microsoft
Phonexia
Qualcomm
Springer
Speech Ocean Limited
SpeechX
Slang labs
Samsung
Sogou
Uniphore
Xiaomi
Yahoo Japan



## International Speech Communication Association (ISCA)

## Organizers

ISCA is a non-profit organization. Its original statutes (statutes in French or their translation in English), were deposited on February 23rd at the Prefecture of Grenoble, in France by René CARRÉ and registered on March 27th, 1988.

The association started as ESCA (European Speech Communication Association) and, since its foundation, has been steadily expanding and consolidating its activities. It has offered an increasing range of services and benefits to its members and it has put its financial and administrative functions on a firm professional footing. Indeed, over the ten years of its existence, ESCA has evolved from a small EEC-supported European organisation to a fully-independent and self-supporting international association.

At the General Assembly meeting that took place during the last Eurospeech conference in Budapest (September 1999), ESCA became a truly international association in the global field of speech science and technology, changing its name to ISCA (International Speech Communication Association) and modifying its statutes accordingly. This also marked the unification between ESCA and the PC-ICSLP: Permanent Council of International Conference on Spoken Language Processing, where ESCA and PC-ICSLP joined to form one international organization, ISCA, representing the world's researchers in speech communication.

The purpose of the association is to promote, in an international world-wide context, activities and exchanges in all fields related to speech communication science and technology. The association is aimed at all persons and institutions interested in fundamental research and technological development that aims at describing, explaining and reproducing the various aspects of human communication by speech, that is, without assuming this enumeration to be exhaustive, phonetics, linguistics, computer speech recognition and synthesis, speech compression, speaker recognition, aids to medical diagnosis of voice pathologies. The main objectives of ISCA are to stimulate scientific research and education, to organize conferences and workshops, to encourage the study of different languages to promote relation between Public and Private and between Science and Technology.

# Message from ISCA President

**John H.L. Hansen**

Welcome to Hyderabad — the first time INTERSPEECH takes place in India! With a country of 1.3 billion people and 22 official languages, but more than 800 languages spoken, India represents one country where advancements in speech and language technology, as well as speech/linguistic sciences, are converging to improve the capabilities of communications between people-to-people and people-to-machines. We are pleased to have you join us for this gathering at our annual INTERSPEECH conference during Sept. 2-6, 2018 in Hyderabad, India. This year's Interspeech is centered around the theme of "Speech Research for Emerging Markets in Multilingual Societies" – a clear reason why India is ideal to host our meeting. As the world continues to advance with applications based on computational modeling and technology inspired by recent advances in artificial intelligence (AI) and machine learning (ML), it is important to remember and recognize that many of these advanced ML concepts/paradigms were first conceived of and deployed within speech-based applications. As such, speech science, linguistics, acoustic phonetics, signal processing, and speech processing are undeniably at the heart of multimodal AI and ML advancements.

ISCA MEDAL for SCIENTIFIC ACHIEVEMENT: The ISCA Medal for Scientific Achievement recognizes and honors an individual each year who has made extraordinary contributions to the field of speech communication science and technology. We are pleased to announce that at this year's INTERSPEECH-2018, we will honor Dr. Bishnu S. Atal, for "Pioneering contributions to speech coding and speech analysis." A number of his papers have been landmark contributions to the field of speech modeling, and formed the basis for many of today's speech coding and communication systems.

ISCA MEDAL SPECIAL SERVICE AWARD: This year, ISCA will recognize the extraordinary service contributions of Prof. Christian Wellekens, who has been providing support for ISCA since the organization was established! Of the many areas he has championed, the ISCApad has been the primary mechanism for information dissemination for the ISCA community.

We will also celebrate our members' achievements by recognizing six ISCA Fellows for 2018, who include: Denis Burnham, Carol Espy-Wilson, Keikichi Hirose, Simon King, Haizhou Li, and Richard Schwartz. Please join me in extending our warm congratulations! INTERSPEECH 2018 marks the 19th Annual Conference of ISCA, which continues the success of recent events. It is a privilege for me to be a part of the conference preparation as ISCA President in the first year of my term. This year's INTERSPEECH Organizing Committee was truly remarkable in their scope, depth, and attention to details – overseeing all the logistics of paper submission, reviews, and final program preparation. Under the leadership of the four technical program

chairs (Hema A. Murthy, Preeti Rao, Paavo Alku, and Prasanta Kumar Ghosh), a total of 48 Area Chairs oversaw the logistics of a 1668 paper submissions, 1319 reviewers, and 5235 reviews! A total of 749 papers were accepted this year. The technical program committee deserves our gratitude for putting an immense amount of work to prepare a quality technical program that covers the latest advancement of speech science and technology. To ensure high quality in the paper review process, this year's INTERSPEECH-2018 ensures that on average each submission was evaluated by approximately four reviews. We extend our sincere thanks to all of our colleagues for completing reviews and ensuring a high quality technical program! Organizing an INTERSPEECH event takes enormous courage, endurance and dedication. I would like to express my gratitude and appreciation to the General Chair, B. Yegnanarayana, as well as the four General Co-Chairs (C. Chandra Sekhar, Shrikanth Narayanan, S. Umesh and S. R. M. Prasanna) who led an experienced organization team to bring INTERSPEECH to India for the first time. Finally, I do hope that you have an enjoyable and productive time in Hyderabad, that you have some time to see and experience the wonderful culture, people and food, and that you will leave with fond memories of INTERSPEECH 2018. With my best wishes for a successful conference!

**John H.L. Hansen**
ISCA President

# Message from General Chair and Co-Chairs

## General Chair

**B. YEGNANARAYANA**
International Institute of Information Technology Hyderabad

## General Co-Chairs

**C. CHANDRA SEKHAR**
Indian Institute of Technology Madras

**SHRIKANTH NARAYANAN**
University of Southern California, Los Angeles

**S. UMESH**
Indian Institute of Technology Madras

**S. R. M. PRASANNA**
Indian Institute of Technology Dharwad

Interspeech is the world's largest and most comprehensive conference on the science and technology of speech and spoken language processing. The conference emphasizes interdisciplinary approaches, addressing all aspects of speech science and technology, ranging from basic theories to advanced applications.

Interspeech 2018 is the 19th annual conference of the International Speech Communication Association (ISCA). We are honoured and delighted to host this year's conference in Hyderabad, the capital city of the newly formed state of Telangana in India. The venue of Interspeech 2018 is the Hyderabad International Convention Centre (HICC), a state-of-the-art facility to host conferences for over 10000 delegates. The attached Novotel hotel is a modern five star hotel with 287 guest rooms. The surrounding HiTech city area has numerous options for transport, accommodation, food, shopping and entertainment, all within a radius of about 10 kilometres from HICC. The city of Hyderabad is over four hundred years old, and has numerous attractions for visitors. There is something for everyone – historical monuments, museums, palaces, handicraft emporiums,

antiques stores, jewellery shops, leather shops, restaurants serving several gastronomic delights, and yoga ashrams. The HICC is 40 minute drive from the Rajiv Gandhi international airport, which is rated among the best in the world for airport service quality.

India is a land of rich history and diversity. With a population of over 1.3 billion, and a geographical area of over 3 million square kilometres, India's rich diversity includes languages, script, history, arts, culture, clothing, cuisine, customs, hospitality, climate, terrain, flora and fauna. This makes communication among people, especially in spoken form, unique and challenging for scientific research and exploration. The theme of Interspeech 2018 "Speech Research for Emerging Markets in Multilingual Societies" is chosen in view of the focus of voice-based mobile services and emerging AI applications by industries across the world. Interspeech 2018 will give a rare opportunity to experience first hand, the challenges of providing technology services to a large multilingual population, with code switching in day-to-day conversations.

The technical program of the Interspeech 2018 conference starts with a talk by the ISCA medalist of this year, Dr. Bishnu S Atal, who happens to hail from India, and educated at the famous Indian Institute of Science in Bangalore. The technical program committee has worked tirelessly to come up with an excellent 5-day program, which includes plenary talks, oral and poster presentations, tutorials, special sessions and events, show & tell sessions, exhibits and social events. In addition to attending technical sessions, we hope that participants will be able to take some time off to enjoy and experience the wide variety of options for food, shopping and historical monuments that the city offers.

While the Indian Institute of Technology (IIT) Madras is the main organizer of the conference, the following institutes in India actively participated and supported this effort: IIIT Hyderabad, IISc Bangalore, IIT Bombay, IIT Mandi, IIT Tirupati, Vignan University, AIISH Mysore, BVRIT Hyderabad and Geetanjali College of Engineering and Technology Hyderabad. The Government of India, especially DRDO and MEiTY, have also generously supported Interspeech 2018. We are very grateful to receive support in the form of sponsorship from many companies and organizations, which include: Amazon, Apple, Baidu, Microsoft, Facebook, Samsung, Uniphore, JD.com, Xiaomi, Qualcomm, Google, IBM Research, Yahoo Japan, Sogou, Nvidia, Adobe, GoVivace, Nuance and the University of Washington.

The organizing committee would like to acknowledge the excellent support from the team led by Arjun Narne, KW Conferences. The organizing committee would also like to thank the ISCA advisory committee, reviewers, area chairs, session chairs and volunteers for their dedicated effort.

Many participants of Interspeech 2018 are going to be the first time visitors to India. India means different things to different people. Come and participate in Interspeech 2018, and discover for yourself what India means for you. We look forward to welcoming you to Hyderabad, and to India.

**B. Yegnanarayana**
General Chair

**C. Chandra Sekhar  Shrikanth Narayanan  S. Umesh  S. R. M. Prasanna**
General Co-Chairs

# Message from the Technical Program Chairs

**HEMA A. MURTHY**
Indian Institute of Technology Madras

**PREETI RAO**
Indian Institute of Technology Bombay

**PAAVO ALKU**
Aalto University, Aalto

**PRASANTA KUMAR GHOSH**
Indian Institute of Science Bangalore

We are very proud to have had the privilege to put together the technical program for Interspeech 2018, in the Indian sub-continent for the first time ever! Over the past decade or more, a growing number of Indian academics and scientists have been participating in Interspeech in its different exciting locations across the globe. We knew we had a huge challenge on our hands, as Technical Program Chairs, to get this edition of the conference in India to live up fully to the long-standing reputation of Interspeech as the most topical, while also comprehensive, speech research conference.

We were truly gratified by the numbers that greeted us on submission closing date. We received 1668 contributed papers, and would like to express our heartfelt thanks to all the authors across 63 countries. The weeks following the submission deadline were filled with organising review assignments and delegating papers for review with the help of a very able team of 48 Area Chairs across the 12 technical areas. We deeply appreciate the promptness with which the Area Chairs responded to our (all too frequent) requests for updates and their commitment to ensuring the timely completion of reviews and the subsequent decision making. Above all, our large team of 1319 reviewers deserves huge thanks for high-quality reviews that made a great technical program possible. We are happy to mention that over 85% of the submissions received four or more reviews. The reviews were indicative of the overall high quality of the submissions, and resulted in the acceptance of a total of 749 papers (acceptance rate = 54%). The ensuing TPC meeting in Hyderabad in May 2018 saw the vibrant exchange of ideas leading to the crystallization of the technical program. We gratefully acknowledge the significant support of the entire ISCA Board in the planning of the program as we now see it. We would like to thank ISCA President John Hansen and the entire team for their valuable advice on various aspects, and for being so encouraging about the new initiatives proposed by the Interspeech India organising team.

Now, for a quick glimpse of the technical program - we have 9 technical sessions each having regular oral, poster, and Show & Tell sessions. We have a glittering line-up of Plenary Talks. In addition, for the first time this year, we shall have Perspective Talks that will offer a bird's-eye view of research in a specific topic. These talks will serve as introduction to the uninitiated, while offering potentially new insights to experts. Interspeech 2018 received several excellent special session proposals including those relating to the theme, such as Indian language speech technology, and technologies for code-switching. The Special Sessions received a large number of submissions and constitute a prominent component of the technical program. We also have a parallel stream for Show & Tell demonstrations throughout the program. The regular technical program is preceded by the conference tutorials, judiciously selected by the 4-member Tutorials Committee from a number of submitted proposals. Several exciting Satellite Events are slated for days both before and after Interspeech in Hyderabad and at other locations in India.

India is a land of many languages, and this linguistic diversity presents many a challenge in spoken language technology, making it the perfect backdrop for Interspeech. We look forward to welcoming you to Hyderabad, and making sure that you have an interesting, enjoyable and productive conference experience!

**Hema A. Murthy** **Preeti Rao** **Paavo Alku** **Prasanta Kumar Ghosh**
Technical Program Chairs

# Speech Research in India

Speech research is today an integral part of academic activity in several prominent institutes across the country, primarily in the departments of Electrical Engineering, Electronics and Communication, and Computer Science. With its origins nearly four decades ago in fundamental studies in Acoustics, Linguistics and Electronics, research focusing on speech signals emerged first in the context of telecommunications applications for bandwidth compression. Almost simultaneous was the inception of interest in human-machine interaction with its naturally high place for voice based communication. The development of academic coursework around the time in the then emerging field of digital signal processing was a natural impetus for signal based research towards the above two large problems. Certain institutions and individuals stand out clearly in the journey of speech research in India from its beginnings in the 1970s. At Indian Institute of Science (IISc) in Bangalore, Prof. B.S. Ramakrishna was known for his work in Acoustics (Architectural, Wave propagation, Music instruments, SAW devices), followed by several brilliant scholars in the same institute who took up speech processing. Dr. Bishnu Atal, Dr. Man Mohan Sondhi, Dr. Nikhil Jayanth and Prof. B. Yegnanarayana all belong to this impressive league. Similarly, at the Tata Institute of Fundamental Research (TIFR), Mumbai, a Physics and Mathematics research Institute, Computer Science was established in the 1960s under the leadership of Prof. R. Narasimhan and Prof. P.V.S. Rao. The closely related research topics of speech, image, script and language processing occupied a prominent place. Advanced laboratories were developed around the Kay Sonograph and other large data processing facilities based on CDC-3600 and DEC-10 computers. Prof. Kuldip Paliwal is another prominent researcher who did seminal work in TIFR. The first dedicated symposium on speech processing was held in TIFR in 1978, with the participation of the top scientists of the time, Prof. Gunnar Fant, KTH, Sweden and Prof. Alvin Liberman, Haskins Labs, USA.

While centres of speech research in India existed in relatively few places until the 1980s, a major spurt occurred soon after with the growing number of doctorates awarded in the area by the aforementioned institutes. With the available larger pool of academics and researchers, several of the Indian Institutes of Technology (IITs) set up speech research labs. This was followed by several private universities who started fruitful collaborations with IIT departments. The traditional emphasis in the early years was on speech analyses including notable work on glottal source signal estimation and group-delay processing. Over the past two decades, large research and development efforts in speech technology have taken root, generously funded by the Department of Electronics and Information Technology, Government of India, as well as some of India's defence research labs. The funding has helped propel the much needed effort of spoken language corpora building for major Indian languages. Collaborative research consortia that have come up across institutes worked to develop limited-domain automatic speech recognition systems for Interactive Voice Response (IVR) as well as text-to-speech systems for Hindi, Marathi, Tamil, Bengali and Assamese among other languages. The challenges posed by the dialect variation and environmental conditions served as important research topics for academic studies and student theses in the member institutions. The growing activity in speech research in India is evident in the prominent place that speech research occupies in major India-based conferences such as the annual National Conference on Communications (NCC) and the biennial International Conference on Signal Processing and Communications (SPCOM). Further, engagement with global leaders in speech research is an important component of the thematic series of winter schools, the Winter School on Speech and Audio Processing (WiSSAP), first envisioned and convened by Prof. T.V.

Sreenivas of IISc Bangalore in 2006, and now much appreciated by the entire speech research community in India. There are also several other prominent speech workshops held annually in India such as the Summer School on Speech Signal Processing and Workshop on Image and Speech Processing.

Currently, the speech research community in India is a very vibrant one and spread across TIFR, IISc, IITs, International Institutes of Information Technology (IIITs), National Institutes of Technology (NITs), and numerous private universities and institutes. Organizations such as the All India Institute of Speech and Hearing (AIISH), Mysore also render clinical services related to speech disorders and hearing impairment in addition to carrying out academic research. There has also been a huge surge in the number of startup companies focusing on speech based applications. Several multinational companies also have a strong presence in all major cities in the country. Going forward, we see an increasing emphasis on data science perspectives in speech research, sophisticated experimental set-ups and the growth of interdisciplinary research.

**Preeti Rao**
Indian Institute of Technology Bombay

**T. V. Sreenivas**
Indian Institute of Science Bangalore

# Theme Related Events

The theme of INTERSPEECH 2018 is "Speech Research for Emerging Markets in Multilingual Societies." In keeping with the theme, various sessions have been planned. These include code-switching in speech synthesis, recognition, low-resource speech recognition, low-resource ASR challenges, language diarization, accent and language adaptation for speaker recognition, and dialog design for multilingual societies. A session-wise listing of the theme related contributions is given below.

**(Wed-P-2-3) Second Language Acquisition and Code-switching:** This poster session has several papers on L1, L2 acquisition and interaction.

**(Wed-O-3-3) Dialectal Variation:** This oral session has contributions related to code switching, regional variations, and multilingual studies.

**(Wed-O-1-2) Language Identification:** This session has papers on language diarization – equivalent to code-switching in spoken dialog.

**(Wed-P-2-2) Speech Synthesis Paradigms and Methods:** This session has papers on cross-lingual studies and code switching.

**(Wed-P-2-1) Adjusting to Speaker, Accent and Domain:** This session has papers on cross-lingual and multilingual aspects in automatic speech recognition (ASR).

**(Wed-O-3-1) Zero-resource Speech Recognition:** This session has papers that deal with languages that do not have too many digital resources.

**(Wed-P-2-4) Topics in Speech Recognition:** This session has important papers on code-switched ASR.

**(Mon-P-2-1) Spoken Dialog Systems and Conversational Analysis:** This session has papers focused on cross-lingual analysis in the context of dialogue.

**(Tue-SS-1-1) Speech Recognition for Indian languages:** This session is completely devoted to ASR for multiple Indian languages with very low resources.

**(Thu-SS-1-2) Low-resource Speech Recognition Challenge for Indian Languages:** This session has papers related to the low-resource challenge that was organised by Microsoft India, Hyderabad.

**(Wed-SS-1-2) Speech Technologies for Code-switching in Multilingual Communities:** This is again a session that is fully devoted to code-switching in multilingual societies.

In addition, we have the following satellite events related to the theme:
**6th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU'18),** New Delhi, India
August 29-31, 2018
URL: http://www.mica.edu.vn/sltu2018/

**English in India and Indian Englishes: New Horizons in the Study of Phonetics and Phonology,**
University of Hyderabad
September 07, 2018
URL: https://phinde1.wordpress.com/blog/phonology-of-indian-englishes-2018/

# Welcome to Hyderabad

Hyderabad is the capital city of the Indian states of Telangana and Andhra Pradesh, in the region of Telangana.

It lies on the Deccan Plateau, 541 meters (1776 ft) above sea level, over an area of 625 $km^2$. The city has an estimated population of around 8 million, making it the 4th largest city in India, while the population of the metropolitan area is estimated to be 9 million. Hyderabad is home to several religions and cultures co-existing in harmony.

| | |
|---|---|
| Name of the city | Hyderabad |
| Name of the state | Telangana |
| Area | 625 $km^2$ (241 sq mi) |
| Time Zone | IST (UTC+5:30) |
| Distance(s) | From Delhi - 1499 km |
| | From Mumbai - 711 km |
| | From Bangalore - 562 km |
| | From Chennai - 688 km |
| | From Kolkata - 1516 km |
| Region | Telangana |
| Founded | 1591 |
| Founder | Muhammad Quli Qutb Shah |
| | Fifth Qutb Shahi Ruler |
| Country code of India | 0091 or +91 |
| City code / STD code of Hyderabad | 040 |
| Languages | Urdu, Telugu, Hindi and English |
| Currency of Hyderabad India | Indian Rupees (INR) |

Hyderabad city is known for its rich history, food and its multilingual culture, both geographically and culturally. The city was founded in the year 1591 by the fifth Qutb Shahi Ruler Muhammad Quli Qutb Shah. Hyderabad was founded on the banks of river Musi, now known as the historic old city which is home to the Charminar, Falaknuma Palace, Chowmahallah Palace and Mecca Masjid.

Hyderabad and Secunderabad are twin cities, separated by the Hussain Sagar (bound by the 'Tank Bund'), an artificial lake made during the time of Ibrahim Qutub Shah in 1562.

## People

What makes Hyderabad such a special destination is its multicultural mix of people.

## Languages

There are four languages in Hyderabad: Hindi, Deccani-Urdu, Telugu and English. English is the language for business and administration, and is widely spoken and understood. Most Hyderabadis are bilingual, and speak their mother tongue as well as English.

## Religions

Hyderabad is also a mixture of religions. Hyderabad's skyline boasts the distinctive minarets of

**Charminar**

mosques, the intricate figurines of Hindu temple gods, beautiful Churches and the distinctive architecture of Gurudwaras. The main religions are Hinduism, Islam, Christianity and Sikhism.

## Security & Safety

Hyderabad has a well-founded reputation as one of the safest cities in the world. Crime rates are very low and the streets are safe to walk at any time of the day or night. However, we do recommend that visitors to the city take the normal precaution of keeping their valuables in a safe place at all times.

The Telangana Government has been vigilant in ensuring that Hyderabad remains safe. It has stepped up security measures at key installations and other sensitive places.

## Wine & Dine

Hyderabad's multicultural population is reflected in the wide variety of excellent restaurants offering a diversity of cuisines from around the world.

## Driving

Minimum age for driving is 18 years with a valid state/provincial driver's license. An international driver's license is required for visits beyond one month. Cars follow the right-hand drive and have to be driven on the left-hand side of the road.

## Passport

By law, you are required to carry your passport with you at all times. You would also need your passport to check-in to your hotel. The police personnel have the authority to ask a foreigner to produce his/her passport.

## What to Wear

Hyderabad has a moderate temperature with 15 to 35 degrees centigrade during September. Light and summer clothing made from natural fabrics like cotton is best for everyday wear. Casual dress is acceptable for most situations and occasions but some establishments may require a more formal dress code. It is always advisable to check beforehand on dress regulations. If you are well dressed, you will usually be treated better, especially when you are doing official business. When going to religious places, you should dress conservatively.

## Credit Cards
American Express, Diners Club, Mastercard and Visa are widely accepted.

## Banking & Money Changing
In general, banks open at 9.30 am and close at 3.30 pm Monday to Friday, and 9.30 am to 12.30 am on Saturday. The second and fourth Saturdays are holidays. Automated teller machines (ATM) are open round-the-clock, offering the MAESTRO or CIRRUS network system. Apart from the convenience of exchanging money at all banks and hotels, money can also be exchanged wherever the sign "Licensed Money Changer" is displayed. Most shopping complexes have a licensed money changer. Visitors are advised not to change money with unlicensed money changers.

## Office Hours
Business and government offices are usually open from Monday to Saturday, 9 am - 6 pm. Offices and shops are closed on Sunday.

## Shopping Hours
Usually shops are open from 10 am to 9 pm.

## Communications / Mobile Phones
Hyderabad is the center of telecommunications in India with 24-hour telex, international direct dial (IDD) telephone, telegram and facsimile services. Telephone country code is +91 and city code is 040. There are many telephone service providers – AirTel, Idea, Reliance, Vodafone and BSNL.

## Travel and Health Insurance
It is strongly recommended that you carry an insurance policy to cover medical and travel expenses.

## Drug Abuse
Illicit trafficking in narcotic drugs and psychotropic substances are strictly prohibited by law.

## Hitech City
Hyderabad is an emerging information technology (IT) and biotechnology hub of India. Hyderabad has witnessed a remarkable growth in the real estate business, thanks to a predominantly information technology-driven boom in the 1990s and the retail industry growth over the last few years which has spurred hectic commercial activity. A number of mega malls have come up in the city.

The rapid growth of the city, along with the growth of Secunderabad and neighbouring municipalities has resulted in a large metropolitan area. Hyderabad is the financial and economic capital of the state. The city is the largest contributor to the state's gross domestic product, state tax and excise revenues. The infrastructural facilities for basic research in Hyderabad are some of the best in the country, hosting a large academic population from all over the country and beyond.

## Health
Please consult your physician regarding health precautions prior to visiting India. If you are coming from or have recently visited Africa, South America or an area infected with Yellow Fever within five days prior to your arrival in India, proof of vaccination against Yellow Fever is required.

Drinking bottled water at all times is strongly recommended. Avoid eating salads or cut fruits on the roadside.

## Electricity

The voltage in India is 230 volts. You will require a voltage converter if you are carrying a device that does not support 230 volts. Wall sockets accept plugs with two or three round pins with a separation of 2 cm between the two smaller round pins. Should you wish to use appliances of 110 volts, most conference hotels will provide adapters on request but it may be useful to carry your own for convenience. It is also advisable to carry a good flashlight.

## Local Time

Indian Standard Time (IST) is 5 hours 30 minutes ahead of GMT. During the congress, IST is 4 hours 30 minutes ahead of London, 3 hours 30 minutes ahead of Paris and 9 hours 30 minutes ahead of New York and 2 hours 30 minutes behind Singapore. We do not change the time during summer. All of India has just one time zone.

## Alcohol Consumption

There are plenty of shops selling liquor of all varieties where the price will be less than the same brand served in the hotel.

As a result of current legislation in the state of Telangana, it is illegal to serve alcohol in public places after 10.30 pm.

# General Information

## Conference Venue

Hyderabad International Convention Centre (HICC)
P O Bag 1101 Cyberabad Post Office, 500081
Kondapur, Hyderabad
India
Tel: +91-40-66824422
Fax: +91-40-66844422

| | |
|---|---|
| Plenary Sessions: | Hall 3 |
| Oral Sessions: | Hall 1, Hall 2, Hall 3, MR 1.01-1.02, MR G.01-G.02, MR G.03-G.04, |
| Poster sessions: | Hall 4 |
| Tutorials: | MR 1.01-1.02, MR G.01-G.03, MR 1.03, MR 1.04 |
| Show & Tell: | MR G.05-G.06 |

## Registration Desk

### Location

Ground Floor, HICC

### Timings

Sunday, September 2, 2018: 08:00 - 18.00
Monday, September 3, 2018: 09:00 - 18.00
Tuesday, September 4, 2018: 08:00 - 18.00
Wednesday, September 5, 2018: 08:00 - 18.00
Thursday, September 6, 2018: 08:00 - 18.00

## Speaker Information

Oral presenters will have 15 minutes for presentation, followed by 4 minutes for summary and question & answer (Q&A), and 1 minute for speaker change. The timing will be strict, and session chairs have been requested to stop speakers from exceeding the 19-minute presentation slot. Presenters should introduce themselves to the session chairs during the break before the start of their oral session.

## Speaker Check-in Room

Location: MR 1.07
Presentations must be submitted on a USB stick at the Speaker Check-In Desk (at least an hour before the scheduled session) during the designated hours to ensure that the talk is accompanied by slides. The operating hours for speaker check-in room are as follows:

| | |
|---|---|
| Monday, September 3, 2018: | 09:00 - 18:30 |
| Tuesday, September 4, 2018: | 08:00 - 17:00 |
| Wednesday, September 5, 2018: | 08:00 - 19:00 |
| Thursday, September 6, 2018: | 08:00 - 16:30 |

## Presentation Format and Room Facilities

Speakers are required to use the computers provided by the conference organizers for their oral presentations. Personal laptops may not be used. The laptops provided by the organizers are equipped with the following:

- Windows i5 Laptops with OS - Windows 7
- Acrobat version - Adobe 11
- PowerPoint Version - MS Office 2013
- Video player - VLC player
- Browser - Google Chrome

PowerPoint or Adobe PDF are the accepted file formats.

The aspect ratio of projectors will be 4:3.

It is recommended that multimedia, sound, or video files are embedded in the presentation. If, for some reason, this is not possible, then they must be provided at the speaker check-in together with the presentation file and clear instructions about which presentation file they belong to.

## Speaker Tips

Please arrive at your presentation room at least 10 minutes prior to the session start time in order to familiarize yourself with the stage set-up (e.g. laser pointer, remote slide advancing etc). Please occupy the front-row seats, which are reserved for the presenters.

The Session Chair will introduce you as well as monitor the length of the presentation. Mobile phones must be turned off while you are presenting. Mobile phones even in silent mode might cause feedback with the microphones. A laser pointer and slide advancer will be available at the podium for your use.

An illustration of how the timer works is given below.

## Poster Information

The poster boards are of size width 72" x height 48" (width 182 cm x height 122 cm). Material to affix your poster to the board will be available on site.

Presenters are advised to mount their posters 15 minutes before the start of the session, and remove it after the session ends. Posters have to be removed after the end of the session. Conference organizers will not be responsible for posters left behind after the session ends.

At least one of the authors must be present at the poster during the designated session. This will help the session chair to identify the presenter and seek any clarifications should the need arise.

## Suggested Poster Format

- Posters should be readable from a distance of 5 feet (1.5 meters).
- A poster printed on one large sheet of matte or semi-gloss paper is preferred. Please avoid

pinning up A4 or A3 sheets of paper.
- Include the title of the presentation in large font, the authors' names, and the institution(s) where the work was carried out at the top of the poster.
- Avoid lengthy paragraphs, provide information in distinct clear and short bullet points.
- Poster numbers will be provided on the poster boards. You do not have to provide space for the poster number on your poster.
- It is advisable to have an introduction and summary/conclusions section.
- Format text with double spacing between lines.

## Poster Design Recommendations

**Title:** no smaller than 72 points
**Author names**: no smaller than 36 points
**Affiliations**: including city and country, no smaller than 26 points
**Fonts**: Preferably sans serif such as Helvetica or Arial
**Abstract**: The abstract presents the main results succinctly.

**Body:** The body of the poster will describe background, methodology, results, and interpretation. The text in the body should be no smaller than 16 points and preferably sans serif such as Helvetica or Arial.

**Graphics:** Graphics are essential for an effective poster and should be accompanied by captions. The captions should be no smaller than 14 points and preferably sans serif.

**Conclusions:** Conclusions should be succinct and provide a clear "take-home" message.

## Recommended Printer for Posters and Other Print Material

If you wish to print your poster in Hyderabad, we recommend that you contact the printer service:
M/s. Shakthikala Printers, Hyderabad
Email: avinash@shakthikala.co.in
Phone: +91 99493 51777

Shakthikala offers poster printing and delivery at Posters Help Desk in Halls 4, 5, and 6 (Exhibition Hall) at HICC (Conference venue). Printing orders can be placed via email on or before August 31, 2018. The accepted file format is a high-resolution PDF/CDR in ready-to-print format. Payment to be made in advance through a bank transfer, the details of which will be provided along with the invoice on the receipt of the order. Below are the printing options and costs.
Poster size 33" (H) X 46" (W)
- Printing on non-tearable material or cloth – INR 1,600 + 18% GST
- Printing on 100 GSM paper – INR 950 + 18% GST

On-site Printing Orders: Printing orders can be placed on-site at the Posters Help Desk by making the payment either by cash or by card. It is advisable that the orders be placed one day in advance of the scheduled poster presentation.

## Show-and-Tell Presentation

The authors of Show-and-Tell presentations are expected to make a poster presentation showcasing their contribution. Poster boards similar to those provided in regular poster sessions will be available. Every Show-and-Tell presentation must be accompanied by a demonstration. A table will be provided in front of every poster board to setup the demo. Electric plug sockets will also be provided to charge computers or laptops or any other equipment needed for demonstration. Wi-Fi will be available in the Show-and-Tell room.

## Emergency Contact Numbers

- Police 100
- Fire 101
- Ambulance 102

These are the numbers in Hyderabad, but they are usually the same in other cities of India.

## Liability

The conference organizers cannot accept liability for injuries or losses arising from accidents or other situations during or as a consequence of the conference.

## Smoking

Smoking is prohibited in public places such as buses, elevators, theatres, cinemas, air-conditioned restaurants, shopping centres and government offices.

## Phones

All mobile phones must be turned off while presenting. Mobile phones kept in silent mode will cause feedback with the microphones.

## Lost & Found

A lost & found counter will be located at the Registration Desk of the conference venue.

## Name Badge

Access to conference events will not be granted without the conference name badge.

## Disabilities

Wheelchair-accessible transportation, reserved seating are available if requested in advance. Should you require assistance onsite, please visit the Registration Desk at the conference venue.

# Social Events

Interspeech 2018 is happy to announce exciting social events.

## 1. Cultural Program (September 3, 2018)

Venue: Hall 3, HICC, Hyderabad
Time: 18:30-19:30

**"Bahurang" - Celebrating Colours, Culture & Creativity**
**by Nirupama and Rajendra of the Abhinava Dance Company, Bangalore**

As the name indicates, Bahurang opens into a blaze of swirling colours, highlighting the beauty and glory of Indian culture, art and philosophy through exquisitely choreographed sequences, high in energy and enthusiasm, and innovative in theme.

## 2. Welcome Reception (September 3, 2018)

Venue: Novotel Gardens, HICC, Hyderabad
Reception Period: 19:30-21:30

## 3. Students' Reception (September 4, 2018)

Location: Club House, Jesmin, Ellaa Hotel, Gachibowli, Hyderabad
Vehicles start at HICC at 19:00
Reception Period: 20:00 - 22:00
Entertainment: DJ Zamaika

## 4. Reviewers' Reception (September 4, 2018)

Location: Kebab Pavilion, Ellaa Hotel, Gachibowli, Hyderabad
Vehicles start at HICC at 19:00
Reception Period: 20:00 - 22:00
Entertainment: Live band - Niraval

# 5. Conference Banquet (September 5, 2018)

Location: HICC, Hyderabad
Reception Period: 19:00 - 21:30
Entertainment: Live band - Threeory
Threeory started with three musicians: Mark - The Pianist, Dattasai - The Violinist, jamming together with a drummer. Today, it is one of the most sought-after bands in Hyderabad and has been awarded Best Live Act award at the India Nightlife Convention and Awards. The initial objective of the band was to collaborate with other artists from the country and explore various genres. They were joined by Syntyche Mongro - The Vocalist later in mid 2014, Sentilong on the Guitars, and Tarun Vishal - The Drummer joined the band later and made the band complete.

# INTERSPEECH 2018 Organizing Team

### General Chair
B. Yegnanarayana, International Institute of Information Technology Hyderabad

### General Co-Chairs
C. Chandra Sekhar, Indian Institute of Technology Madras
Shrikanth Narayanan, University of Southern California, Los Angeles
S. Umesh, Indian Institute of Technology Madras
S. R. M. Prasanna, Indian Institute of Technology Dharwad

### Technical Program Chairs
Hema A. Murthy, Indian Institute of Technology Madras
Preeti Rao, Indian Institute of Technology Bombay
Paavo Alku, Aalto University, Aalto, Finland
Prasanta Kumar Ghosh, Indian Institute of Science Bangalore

### Local Organizing Committee
Suryakanth V. Gangashetty, International Institute of Information Technology Hyderabad
V. Kamakshi Prasad, Jawaharlal Nehru Technological University, Hyderabad
C. Krishna Mohan, Indian Institute of Technology Hyderabad

### Publicity Committee
Chandra Sekhar Seelamantula, Indian Institute of Science Bangalore
A. G. Ramakrishnan, Indian Institute of Science Bangalore
T. V. Sreenivas, Indian Institute of Science Bangalore
Mathew Magimai Doss, Idiap Research Institute, Martigny
Andreas Stolcke, Microsoft Research, USA

### Sponsorship Committee
Sriram Ganapathy, Indian Institute of Science Bangalore
Prem Natarajan, University of Southern California, Los Angeles
Rohit Prasad, Amazon, USA
Lin-shan Lee, National Taiwan University, Taipei
Ajish Abraham, All India Institute of Speech and Hearing, Mysore

### Tutorials
V. Ramasubramanian, International Institute of Information Technology Bangalore
Yannis Stylianou, University of Crete, Crete
G. Panayiotis Georgiou, University of Southern California, Los Angeles
Israel Cohen, Technion Israel Institute of Technology, Technion City

## Satellite Workshops

Mark Hasegawa-Johnson, University of Illinois at Urbana-Champaign
Gerard Bailly, GIPSA Lab, CNRS
Hemant A. Patil, Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar

## Finance

K. Sreenivasa Rao, Indian Institute of Technology Kharagpur
A. D. Dileep, Indian Institute of Technology Mandi

## Publications

K. Samudravijaya, Indian Institute of Technology Guwahati
T. Nagarajan, SSN College of Engineering, Chennai
Samuel Thomas, IBM Watson Research Center, USA
D. R. Sanand, Toshiba Research, UK

## Special Events

Preethi Jyothi, Indian Institute of Technology Bombay
Pratibha Moogi, Samsung R&D, Bangalore
Dinesh Babu, International Institute of Information Technology Bangalore
Deepu Vijayasenan, National Institute of Technology Karnataka, Surathkal

## Students Affairs and Grants

John H. L. Hansen, University of Texas-Dallas, USA
Priyankoo Sarmah, Indian Institute of Technology Guwahati

## Web, Tools and APPs

Chandra Sekhar Seelamantula, Indian Institute of Science Bangalore
Srikanth Madikeri, IDIAP, Martigny
Harshavardhan Sundar, Johns Hopkins University, Baltimore
Ajay Srinivasamurthy, Amazon, Bangalore
Anil Kumar Vuppala, International Institute of Information Technology Hyderabad

## Industry Liaison

Radhakrishnan Srikanth, Microsoft, Hyderabad
Jithendra Vepa, Samsung R&D, Bangalore
Kavita Reddi, Voxta Communication Solutions Pvt. Ltd., Hyderabad

## Social Events

R. Padmanabhan, Indian Institute of Technology Mandi
K. S. Vijaya Sekhar, International Institute of Information Technology Hyderabad
V. V. Ramana, International Institute of Information Technology Hyderabad

## Exhibits

K. Sri Rama Murty, Indian Institute of Technology Hyderabad
Vinay Kumar Mittal, Ritwik Software Technologies Pvt. Ltd., Hyderabad
Sunayana Sitaram, Microsoft India, Bangalore

## Advisory Committee

D. Raj Reddy, Carnegie Mellon University, Pittsburgh
Bhaskar Ramamurthi, IIT Madras

S. S. Agrawal, KIIT Group of Colleges, Gurgaon
Keikichi Hirose, University of Tokyo Tokyo
Roger K. Moore, University of Sheffield, Sheffield
Isabel Trancoso, INESC-ID / IST / UTL Spoken Language Processing
Douglas O'Shaughnessy, INRS-EMT, Montreal
Abeer Alwan, University of California, Los Angeles
John H.L. Hansen, The University of Texas at Dallas
Hynek Hermansky, Johns Hopkins University, Baltimore
Haizhou Li, National University of Singapore, Singapore
G. Satheesh Reddy, Scientific Advisor to Raksha Mantri, Government of India
Swaran Lata, TDIL Programme, DEITY, New Delhi
Rajeev Sangal, Director, IIT (BHU) Varanasi
David Nahamoo, IBM Research, USA

# Technical Program Committee

## Technical Program Chairs

Hema A. Murthy, Indian Institute of Technology Madras
Preeti Rao, Indian Institute of Technology Bombay
Paavo Alku, Aalto University, Aalto
Prasanta Kumar Ghosh, Indian Institute of Science Bangalore

## Show & Tell Chairs

Ananth Sankar, LinkedIn USA
Kalika Bali, Microsoft Research Lab India

## Area Chairs

**Speech Perception, Production, and Acquisition**
Martin Cooke, University of the Basque
Olov Engwall, KTH Royal Institute of Technology
Tan Lee, Chinese University of Hong Kong
Odette Scharenborg, Delft University of Technology
T. V. Ananthapadmanabha, Voice and Speech Systems, Bangalore

**Phonetics, Phonology, and Prosody**
Julia Hirschberg, Columbia University
Jacques Koreman, NTNU
Priyankoo Sarmah, Indian Institute of Technology Guwahati
Mariapaola D'Imperio, Aix Marseille University

**Analysis of Paralinguistics in Speech and Language**
Chi-Chun Lee, National Tsing Hua University, Taiwan
Elizabeth Shriberg, International Computer Science Institute, Berkeley
Björn W. Schuller, Imperial College London & University of Augsburg
Denis Jouvet, LORIA – INRIA Nancy – Grand-Est.

**Speaker and Language Identification**
Haizhou Li, National University of Singapore
Najim Dehak, Johns Hopkins University
Petr Motlicek, Idiap Research Institute
Luciana Ferrer, National Council of Scientific and Technical Research, Argentina
Ming Li, Duke Kunshan University

**Analysis of Speech and Audio Signals**
Mark Hasegawa-Johnson, Beckman Institute, University of Illinois at Urbana-Champaign
Koichi Shinoda, Tokyo Institute of Technology
K. Sri Rama Murty, Indian Institute of Technology Hyderabad
P. Vijayalakshmi, SSN College of Engineering

**Speech Coding and Enhancement**
Rajesh M. Hegde, Indian Institute of Technology Kanpur
Sebastian Möller, Technische Universität Berlin
Tom Bäckström, Aalto University
Gernot Kubin, Graz University of Technology

**Speech Synthesis and Spoken Language Generation**
Heiga Zen, Google Inc., U.K.
K. Sreenivasa Rao, Indian Institute of Technology Kharagpur
Raul Fernandez, Human Language Technologies, IBM

**Speech Recognition–Signal Processing, Acoustic Modeling, Robustness, and Adaptation**
Jean-Luc Gauvain, LIMSI-CNRS
Bhiksha Raj, Carnegie Mellon University
Thomas Hain, University of Sheffield
Rohit Sinha, Indian Institute of Technology Guwahati
Martin Karafiat, Brno University of Technology

**Speech Recognition–Architecture,Search, and Linguistic Components**
Murat Saraclar, Boğaziçi University
Preethi Jyothi, Indian Institute of Technology Bombay
Michiel Bacchiani, Google Inc., New York

**Speech Recognition–Technologies and Systems for New Applications**
Sri Garimella, Amazon, Bangalore
Mathew Magimai Doss, Idiap Research Institute
Torbjørn Svendsen, Norwegian University of Science and Technology

**Spoken Dialog Systems and Analysis of Conversations**
Joakim Gustafson, Royal Institute of Technology (KTH), Sweden
Alexandros Potamianos, National Technical University of Athens
Isabel Trancoso, University of Lisbon, Chair of Scientific Council of INESC-ID
Helen Meng, The Chinese University of Hong Kong

**Spoken Language Processing–Translation, Information Retrieval, Summarization, Resources and Evaluation**
Tanja Schultz, University of Bremen
Arne Jönsson, Linköping University
Lori Lamel, French National Centre for Scientific Research

**Special Sessions**
Panayiotis Georgiou, University of Southern California
Brian Kingsbury, IBM Research, USA
Shiva Sundaram, Amazon, USA
Bhuvana Ramabhadran, Google Inc., USA

# Scientific Review Committee

Aaron Lawson
Aasish Pappu
Abdellah Kacha
Abhinav Misra
Abualsoud Hanani
Adam Janin
Adam Vogel
Adamantios Gafos
Adriana Vasilache
Agnieszka Wagner
Agustin Gravano
Ahilan Kanagasundaram
Ahmed Ali
Ahmed Hussen Abdelaziz
Aijun Li
Ajay Srinivasamurthy
Akio Kobayashi
Akira Sasou
Alan McCree
Alberto Abad
Alejandrina Cristia
Alessandro Delfino
Alessandro Moschitti
Alessio Brutti
Alex Acero
Alexander Lodermeyer
Alexander Rudnicky
Alexander Sorin
Alexandra Jesse
Alexandre Allauzen
Alexandros Lazaridis
Alexei Kochetov
Alexey Karpov
Alexsandro Meireles
Alfonso Ortega
Ali Orkan Bayer
Allison Wetterlin
Alok Parlikar
Amandine Michelas
Amaro Lima
Amelie Rochet-Capellan
Amir Hossein Harati Nejad Torbati
Amparo Varona
Ana Isabel Mata
anand darji
Ananth Sankar
Ananthakrishnan Gopal
Ananya Misra
Anastasios Alexandridis
Anastassia Loukina
Anders Lindström
Anders Lofqvist
Andre Adami
Andre Coy
Andrea DeMarco
Andreas Maier
Andreas Tsiartas
Andrej Zgank
Andrew Breen
Andrew Plummer
Andrew Rosenberg
Angel Gomez
Angelika Braun
Angeliki Metallinou
Anil Kumar Vuppala
Anita Wagner
Ankur Gandhe
Anne Bonneau
Annemie Van Hirtum
Anoop Deoras
Anssi Rämö
Ante Jukić
Anthony Larcher
Antje Schweitzer
Antoine Laurent
Anton Batliner
Anton Ragni
Antonio Bonafonte
Antonio M. Peinado
Antonio Miguel
Antonio Origlia
Antonio Satue Vilar
António Teixeira
Arivudai Nambi
Ariya Rastrow
Arnab Ghoshal
Arnaud Delhay
Arseniy Gorin
Arun Narayanan
Arvi Tavast
Åsa Abelin
Ascension Gallardo-Antolin
Asger Heidemann Andersen
Ashish Panda
Ashtosh Sapru
Asli Celikyilmaz
Asterios Toutios
Astik Biswas
Asuncion Moreno
Aswin Shanmugam Subramanian
Atef Ben Youssef
Athanasios Katsamanis
Athanasios Lykartsis
Athanasios Mouchtaris
Atsunori Ogawa
Atsushi Nakamura
Attila Máté Tóth
Axel Roebel
Bartosz Ziolko
Bassam Jabaian
Batuhan Gundogdu
Bayya Yegnanarayana
Ben Milner
Benjamin Elie
Benjamin Lecouteux
Benjamin Munson
Benjamin Weiss
Benoit Fauve
Benoit Favre
Berlin Chen
Bernardt Duvenhage
Bernd Möbius
Bernd T. Meyer
Bettina Braun
Bharathi B
Bidisha Sharma
Bin Gao
Bishnu Atal
Björn Granström
Blaise Potard
Bo Li
Bo Xiao
Bogdan Vlasenko
Bonny Sands
Branislav Gerazov
Brian Hutchinson
Brian Mak
Bruce Denby
Bruce Gerratt
Byeongchang Kim
Calbert Graham
Camille Fauth
Carl Quillen
Carl Vogel
Carlo Drioli
Carlos Busso
Carlos Ferrer
Carlos Gussenhoven
Carlos Ishi
Carlos Lima
Carlos Ribeiro
Carlos Segura
Carlos Teixeira
Carlos Vivaracho-Pascual
Carlos-D. Martínez-Hinarejos
Carme de-la-Mota
Carmen Peláez-Moreno
Carol Espy-Wilson
Caroline Wiltshire
Cassia Valentini-Botinhao
Catalin Zorila
Catharine Oertel
Catherine Lai
Catherine Watson
Cemal Hanilci
Cenk Demiroglu
Chai Wutiwiwatchai
Chan Chun-an
Chandra Sekhar Chellu
Chandra Sekhar Seelamantula
Chandramohan Senthilkumar
Chang Yoo
Changhuai You
Changxue Ma
Chao Weng
Chao Zhang
Chaojun Liu
Cheeran Alice
Chen-Yu CHIANG
Cheol Jae Seong
Cheolwoo Jo
Chi (Leo) Zhang
Chien-Lin Huang
Ching-Feng Yeh
Chin-Hui Lee
Chloé Clavel
Chongjia Ni
Chris Davis
Christian Gollan
Christian Kroos
Christian Plahl
Christian Raymond
Christian Ritz
Christian Wellekens
Christina Bergmann
Christina Esposito
Christina Hagedorn
Christina Tånnander
Christophe Cerisara
Christophe Coupe
Christophe Couvreur
Christophe d'Alessandro
Chung-Hsien Wu
Chunlei Zhang
Chunyang Wu
Ciprian Chelba
Claudio Vair
Coriandre Vilain
Corinne Fredouille
Dafydd Gibbon
Daisuke Saito
Damian Cetnarowicz
Damien Lolive
Dan Povey
Dani Byrd
Daniel Duran
Daniel Ellis
Daniel Erro
Daniel Garcia-Romero
Daniel Kohlsdorf
Daniel Olson
Daniel Ramos
Daniel Tihelka
Daniel Willett
Dante Barone
Daryush Mehta
Dat Tran
David Braude
David Burshtein
David Doukhan
David Griol
David Guennec
David Harwath
David House
David Imseng
David Janiszek
David Looney
David Marks
David Martínez González
David Martins de Matos
David Novick
David Rybach
David Schlangen
David Suendermann-Oeft
Dávid Sztahó
David van Leeuwen
Dayana Ribas Gonzalez
Deborah Dahl
Deepak Baby
Deepak K T
Deepu Vijayasenan
Delphine Charlet
Dhananjaya Gowda
Diamantino Caseiro
Diego Castan
Dietrich Klakow
Dileep Aroor Dinesh
Dilek Hakkani-Tur
Dimitra Vergyri
Dipankar Das
Dirk Van Compernolle
Dokmanic Ivan
Dominic Telaar
Donald Williamson
Dong Wang
Dong Yu
Dong Yu
Dongmei Wang
Dongsuk YOOK
Dongyan Huang
Donna Erickson
Dorothea Kolossa
Douglas O'Shaughnessy
Douglas Reynolds
DoYeong Kim
Duc Le
Ebru Arisoy
Eduardo Coutinho
Eduardo Lleida Solano
Eduardo Rodriguez Banga
Elias Iosif
Eliathamby Ambikairajah
Elie Khoury
Elif Bozkurt

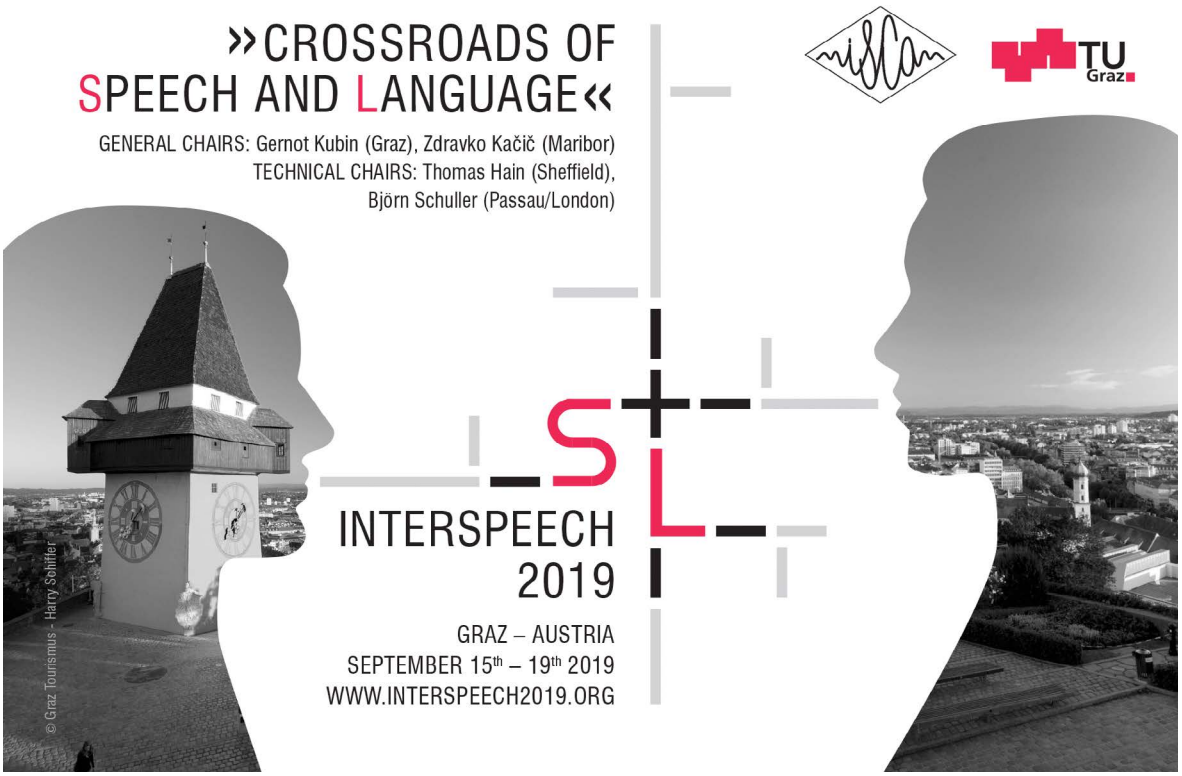
Elizabeth Sherly
Ellen Marklund
Elliot Moore
Elliot Saltzman
Elliot Singer
Elmar Nöth
Emilia Parada-Cabaleiro
Emilio Sanchis
Emily Mower Provost
Emma Jokinen
Emmanuel Vincent
Emre Yilmaz
Encarna Segarra
Eng Siong Chng
Engin Erzin
Enrico Zovato
Eric Fosler-Lussier
Eric Hansen
Erica Gold
Erik Edwards
Erik Marchi
Erik McDermott
Erinç Dikici
Esther Janse
Esther Klabbers
Ethan Selfridge
Ethan Sherr-Ziarko
Eugene Weinstein
Eva Hajicova
Eva Navas
Eva Reinisch
Eva Szekely
Evandro Gouvea
Evgeny Stepanov
Ewen MacDonald
Fabien Ringeval
Fabio Tamburini
Fabrice Lefevre
Fabrice Malfrère
Falavigna Daniele
Fang Chen
Fanny Meunier
Faris Mona
Febe De Wet
Fei Chen
Felicitas Kleber
Felix Albu
Felix Burkhardt
Felix Putze
Fernando Batista
Fernando de la Calle Silos
Fernando Fernández-Martínez
Fernando García
Fernando Gil Vianna Resende Junior
Fernando Perdigão
Fernando Villavicencio
Ferran Pla
Fil Alleva
Firoj Alam
Flavio Simoes
Florian Hintz
Florian Hönig
Francesca Bonin
Francesco Cutugno
Francis Grenez
Francisco Casacuberta
Francisco J Valverde-Albacete
François Yvon
Frank Kurth
Frank Rudzicz
Frantisek Grezl
Franz Pernkopf
Fred Cummins
Fred Richardson
Frederic Bechet
Frederic Berthommier
Fredrik Karlsson
Friedrich Faubel
Friedrich Neubarth
Fuming Fang
Gábor Gosztolya
Gabriel Hernandez-Sierra
Gabriel Skantze
Gakuto Kurata
Galina Lavrentyeva
Gang Liu
Ganna Raboshchuk
Gary Leung
Gautam Mantena
Geoffrey Zweig
Georg Stemmer
George Kafentzis
George Saon
Georges Linares
Gerald Friedland
Gerald Penn
Géraldine Damnati
Gérard Bailly
Gerasimos Potamianos
Germán Sanchis-Trilles
Gerry Docherty
Géza Németh
Giacobello Daniele
Giampiero Salvi
Gilles Adda
Gilles Boulianne
Gilles Degottex
Gina-Anne Levow
Gintare Grigonyte
Giuseppe Di Fabbrizio
Giuseppe Riccardi
Gláucia Laís Salomão
Gokhan Tur
Golan Pundak
Govind D
Grace Kuo
Graham Wilcock
Grazyna Demenko
Gregory Sell
Gregory Senay
Grzegorz Chrupała
Guangpu Huang
Gueorgui Pironkov
Guillaume Fuchs
Gunilla berndtsson
Guoguo Chen
Guoli Ye -
Guruprasad Seshadri
Gustav Eje Henter
Gwénolé Lecorvé
György Szaszák
Hagai Aronowitz
Haifeng Li
Haizhou Li
Hakan Erdogan
Hannah Muckenhirn
Hanseok Ko
Hans-Guenter Hirsch
Hao Tang
Hao Wang
Harald Höge
Hari Krishna Maganti
Harinath Garudadri
Harish Arsikere
Harishchandra Dubey
Harshavardhan Sundar
Hartmut Helmke
Haws David
Héctor Delgado
Heidi Christensen
Helena Moniz
Hélène Maynard
Helmer Strik
Hemant Patil
Heng Lu
Henk van den Heuvel
Henning Schepker
Heriberto Cuayahuitl
Herman Kamper
Herve Bourlard
Hervé Bredin
Heysem Kaya
Hideki Kawahara
Hinako Masuda
Hiroaki Kato
Hirokazu Kameoka
Hiroki Mori
Hiroki Tanaka
Hiromitsu Nishizaki
Hiroshi G. Okuno
Ho-hsien Pan
Hoi Rin Kim
Homayoon Beigi
Hong Kook Kim
Hong-Goo Kang
Hongwei Ding
Horacio Franco
Houwei Cao
Hsiao-Chuan Wang
Hsin-Min Wang
Hsin-Te Hwang
Hua Xing
Hugo Cordeiro
Hugo Van hamme
hui feng
Hui Jiang
Huiqun Deng
Hung-yi Lee
Huy Phan
Hynek Boril
Hynek Hermansky
Hyung Soon Kim
Ian Howard
Ian Maddieson
Ian McLoughlin
Ian Stavness
Ibon Saratxaga
I-Fan Chen
Igor Szoke
ilija zeljkovic
Ilya Oparin
Imed Laaridh
Imed Zitouni
Ina Kodrasi
Ingo Siegert
Inma Hernaez
Ioana Vasilescu
Irina Illina
Irina Kipyatkova
Iris-Corinna Schwarz
Isabel Falé
Isabelle Ferrané
Itshak Lapidot
Ivan Himawan
Ivan Kraljevski
Iván López-Espejo
Jacques Duchateau
Jaewon Lee
James Hieronymus
James M Scobbie
James Wayman
Jan Alexandersson
Jan Černocký
Jan Nouza
Jan Silovsky
Jan Skoglund
Jan Švec
Jan Trmal
Jan van Doorn
Jan van Santen
Jane Stuart-Smith
Jane Wottawa
Jangwon Kim
Jared Bernstein
Jarek Krajewski
Jaromir Tovarek
Jasha Droppo
Jason Pelecanos
Javier Ferreiros
Javier Franco-Pedroso
Javier Hernando
Javier Latorre
Javier Macias-Guarasa
Javier Perez
Jayanna H S
Jayashree Bhat
Jean Schoentgen
Jean-Francois Bonastre
Jeanin Jügler
Jean-Luc Rouas
Jeesun Kim
Jeff Ma
Jen-Chun Lin
Jennifer Cole
Jens Edlund
Jen-Tzung Chien
Jeremie Lecomte
Jesper Jensen
Jesús B. Alonso-Hernández
Jesus Villalba
Jeung-Yoon Elizabeth Choi
Jia Cui
Jian Cheng
Jian Gong
Jianguo Wei
Jianhua Tao
Jianming Liu
Jianwu Dang
Jilt Sebastian
Jimmy Kunzmann
Jindrich Matousek
Jinfu Ni
Jinsong Zhang
Jinyu Li
Jithendra Vepa
Jiyoung Shin
Joan Andreu Sánchez
Joao Cabral
João Paulo Teixeira
Jody Kreiman
Johan Boye
Johan Frid
Johan Rohdin
Johan Wouters
John Bridle
John H.L. Hansen
John Henderson
John Kane
John Pitrelli
Jon Barker
Jon Gudnason
Jon Sanchez
Jonas Beskow
Jonas Borgstrom
Jonathan Chevelu
Jonathan Ginzburg
Jonathan Harrington
Jonathan Le Roux
Jordi Adell
Jorge Baptista
Jorge Lucero
Jorn Anemuller
Jose A. Gonzalez
José A. R. Fonollosa
José L. Pérez-Córdoba
José Lopes
Jose Luis Blanco Murillo
Jose Mario De Martino
Jose Miguel Benedi
José Pardo
José Portêlo
Jose Ramon Calvo de Lara
Joseph Campbell
Joseph Polifroni
Joseph Tepperman
Jouni Pohjalainen

Herman Kamper
Herve Bourlard
Hervé Bredin
Heysem Kaya
Hideki Kawahara
Hinako Masuda
Hiroaki Kato
Hirokazu Kameoka
Hiroki Mori
Hiroki Tanaka
Hiromitsu Nishizaki
Hiroshi G. Okuno
Ho-hsien Pan
Hoi Rin Kim
Homayoon Beigi
Hong Kook Kim
Hong-Goo Kang
Hongwei Ding
Horacio Franco
Houwei Cao
Hsiao-Chuan Wang
Hsin-Min Wang
Hsin-Te Hwang
Hua Xing
Hugo Cordeiro
Hugo Van hamme
hui feng
Hui Jiang
Huiqun Deng
Hung-yi Lee
Huy Phan
Hynek Boril
Hynek Hermansky
Hyung Soon Kim
Ian Howard
Ian Maddieson
Ian McLoughlin
Ian Stavness
Ibon Saratxaga
I-Fan Chen
Igor Szoke
Ilija Zeljkovic
Ilya Oparin
Imed Laaridh
Imed Zitouni
Ina Kodrasi
Ingo Siegert
Inma Hernaez
Ioana Vasilescu
Irina Illina
Irina Kipyatkova
Iris-Corinna Schwarz
Isabel Falé
Isabelle Ferrané
Itshak Lapidot
Ivan Himawan
Ivan Kraljevski
Iván López-Espejo
Jacques Duchateau
Jaewon Lee
James Hieronymus
James M Scobbie
James Wayman
Jan Alexandersson
Jan Černocký
Jan Nouza
Jan Silovsky
Jan Skoglund
Jan Švec
Jan Trmal
Jan van Doorn
Jan van Santen
Jane Stuart-Smith
Jane Wottawa
Jangwon Kim
Jared Bernstein
Jarek Krajewski
Jaromir Tovarek
Jasha Droppo
Jason Pelecanos
Javier Ferreiros
Javier Franco-Pedroso
Javier Hernando
Javier Latorre
Javier Macias-Guarasa
Javier Perez
Jayanna H S
Jayashree Bhat
Jean Schoentgen
Jean-Francois Bonastre
Jeanin Jügler
Jean-Luc Rouas
Jeesun Kim
Jeff Ma
Jen-Chun Lin
Jennifer Cole
Jens Edlund
Jen-Tzung Chien
Jeremie Lecomte
Jesper Jensen
Jesús B. Alonso-Hernández
Jesus Villalba
Jeung-Yoon Elizabeth Choi
Jia Cui
Jian Cheng
Jian Gong
Jianguo Wei
Jianhua Tao
Jianming Liu
Jianwu Dang
Jilt Sebastian
Jimmy Kunzmann
Jindrich Matousek
Jinfu Ni
Jinsong Zhang
Jinyu Li
Jithendra Vepa
Jiyoung Shin
Joan Andreu Sánchez
Joao Cabral
João Paulo Teixeira
Jody Kreiman
Johan Boye
Johan Frid
Johan Rohdin
Johan Wouters
John Bridle
John H.L. Hansen
John Henderson
John Kane
John Pitrelli
Jon Barker
Jon Gudnason
Jon Sanchez
Jonas Beskow
Jonas Borgstrom
Jonathan Chevelu
Jonathan Ginzburg
Jonathan Harrington
Jonathan Le Roux
Jordi Adell
Jorge Baptista
Jorge Lucero
Jorn Anemuller
Jose A. Gonzalez
José A. R. Fonollosa
José L. Pérez-Córdoba
José Lopes
Jose Luis Blanco Murillo
Jose Mario De Martino
Jose Miguel Benedi
José Pardo
José Portêlo
Jose Ramon Calvo de Lara
Joseph Campbell
Joseph Polifroni
Joseph Tepperman
Jouni Pohjalainen
Juan Ignacio Godino llorente
Juan Rafael Orozco-Arroyave
Judith Bishop
Juergen Schroeter
Juergen Trouvain
Julian David Arias Londoño
Julie Berndsen
Julie Valérie Belião Beliao
Julien Epps
Julien Pinquier
Jun Wang
Jun Wang
Junfeng Li
Junichi Yamagishi
Junru Wu
Jun-Won Suh
Juraj Šimko
JV Satyanarayana
K Samudravijaya
K V Vijay Girish
Kaavya Sriskandaraja
Kai Yu
Kailash Patil
Kalika Bali
Kallirroi Georgila
Kamel Smaili
Kanako Tomaru
Karan Nathwani
Karen Mulak
Karthika Vijayan
Kartik Audhkhasi
Katarzyna Klessa
Kate Knill
Kaustubh Kalgaonkar
Kazuhiro Kobayashi
Kazuhiro Nakadai
Kazumasa Yamamoto
Kazunori Komatani
Kazuyo Tanaka
Kee-Ho Kim
Keelan Evanini
Kei Hashimoto
Keiichi Funaki
Keiichi Yasu
Keiichiro Oura
Keikichi Hirose
Keisuke Kinoshita
Kenneth Church
Kevin Farrell
Kevin Tang
Khalid Choukri
Khalid Daoudi
Khiet Truong
Kimiko Tsukada
Klara Vicsi
Koichiro Yoshino
Koji Iwano
Kong Aik Lee
Konstantin Simonchik
Korbinian Riedhammer
Korin Richmond
Kornel Laskowski
Kou Tanaka
Kris Demuynck
Kristiina Jokinen
Kristin Precoda
Kristina Lundholm Fors
Kristina Nilsson Björkenstam
Krzysztof Marasek
Krzysztof Wolk
Kshitiz Kumar
Kuan-Yu Chen
Kun Qian
Kuniko Nielsen
Kyle Gorman
Kyu Han
Ladan Baghai-Ravary
Lahiru Samarakoon
Lakshmi Saheer
Lars Bo Larsen
Lasguido Nio
László Tóth
Laura Docio-Fernandez
Laura Koenig
Laura Tomokiyo
Laurence Bruggeman
Laurens van der Werff
Laurent Besacier
Laurent Girin
Lee Seunghun
Leena Mary
Lei He
Lei Miao
Lei Xie
Leonardo Badino
Levent Arslan
Liang Lu
Lijuan Wang
Lin-shan Lee
Lionel Feugère
Liping Chen
Lirong Dai
Lisa Gustavsson
Liu, Jia -
Lluís-F. Hurtado
Loredana Sundberg Cerrato
Lori Lamel
Lorin Wilde
Louis Goldstein
Louis ten Bosch
Luc Ardaillon
Ludek Muller
Ludo Max
Luis Caldas de Oliveira
Luis Fernando D'Haro
Luis Hernandez-Gomez
Luis Javier Rodriguez-Fuentes
Luis M.T. Jesus
Luísa Coheur
Lukas Burget
Lukas Latacz
Lyan Verwimp
Maarten Van Segbroeck
Madina Hasan
Maeva Garnier
Magdalena Igras-Cybulska
Mahesh Kumar Nandwana
Mahsa Yarmohammadi
Maja Taseska
Malcolm Slaney
Man-hung Siu
Manny Rayner
Manu Airaksinen
Manwai Mak
Marc Delcroix
Marc Ferras
Marc Swerts
Marcel Kockmann
Marcin Wlodarczak
Marco Matassoni
Margaret Zellers
Margarita Kotti
Maria Jose Castro-Bleda
Maria Josep Solé
Maria Koutsogiannaki
Maria Uther
Maria Wolters
Marianne Pouplier
Marie Meteer
Marie Tahon
Marija Tabain
Mariko Kondo
Marina Kalashnikova
Mario Refice
Marion Dohen
Mark Fuhs

Mark Gales
Mark Hasegawa-Johnson
Mark Huckvale
Mark Steedman
Mark Tiede
Markku Turunen
Markus Müller
Markus Nussbaum-Thom
Martha Larson
Martin Cmejrek
Martin Graciarena
Martin Heckmann
Martine Adda-Decker
Martti Vainio
Mary Beckman
maryam najafian
Marzena Zygis
Marzieh Razavi
Masafumi Nishida
Masafumi Nishimura
Masakiyo Fujimoto
Masami Akamine
Masanori Morise
Masashi Unoki
Masato Akagi
Masato Ishizaki
Masayuki Nishiguchi
Masayuki Suzuki
Massimiliano Todisco
Mathieu Chollet
Mats Wirén
Matt Seigel
Matt Shannon
Matthew Henderson
Matthias Paulik
Matthias Sjerps
Matthias Wolff
Maulik Madhavi
Mauro Falcone
Mauro Nicolao
Md Jahangir Alam
Md Sahidullah
Melanie Weirich
Melissa A. Redford
Melvyn Hunt
Mengyue Wu
Michael Deisher
Michael Heck
Michael Johnson
Michael Levit
Michael Picheny
Michael Proctor
Michael Riley
Michael Seltzer
Michael Tjalve
Michael Tyler
Michael Wagner
Michel Hoen
Michel Simard
Michelina Savino
Mikael Roll
Mikko Kurimo
Milan Legát
Milos Cernak
Miloslav Konopik
Milton Sarria-Paja
Min Ma
Ming Li
Ming Sun
Ming Sun
Ming-Hsiang Su
Minghui Dong
Minho Jin
Mireia Diez
Mireia Farrús
Mitchell McLaren
Mohamed Afify
Mohamed Faouzi Benzeghiba
Mohamed Morchid
Mojtaba Farmani
Molly Babel
Monojit Choudhury
Mukund Padmanabhan
Munir Georges
Muralishankar R
Murtaza Bulut
Myoung-Wan Koo
Na Li
Nagaraj Adiga
Najmeh Sadoughi
Nancy Chen
Naohiro Tawara
Naomi Harte
Naoyuki Kanda
Nassima Abdelli-Beruh
Natalia Tomashenko
Nathalie Camelin
Nathalie Henrich
Naveen Kumar
Naveen Parihar
Navid Shokouhi
Neeraj Sharma
Nelly Barbot
Nestor Becerra Yoma
Ngoc Thang Vu
Nicholas Cummins
Nicholas Evans
Nick Campbell
Nicolas Morales
Nicolas Obin
Nicolas Scheffer
Nicole Beringer
Nigel Ward
Nikolay Gaubitch
Nikoletta Basiou
Nina Dethlefs
Ning Ma
Nitendra Rajput
Noam Amir
Nobuaki Minematsu
Nobuyuki Nishizawa
Noel Nguyen
Norbert Reithinger
Norihide Kitaoka
Odette Scharenborg
Oh-Wook Kwon
Okko Räsänen
Oldrich Plchot
Olga Dmitrieva
Oliver Jokisch
Oliver Niebuhr
Oliver Watts
Olivier Crouzet
Olivier Deroo
Olivier Galibert
Olivier Perrotin
Olivier Rosec
Olivier Siohan
Ondřej Dušek
Ondrej Glembek
Oscar Saz
Ove Andersen
Oytun Turk
Ozlem Kalinli
Pablo Riera
Padmanabhan Rajan
Paolo Baggia
Parham Mokhtari
Pascal Perrier
Patricia Keating
Patrick Cardinal
Patrick Nguyen
Patrick Paroubek
Patrick Wambacq
Patti Price
Paul Heisterkamp
Paula Lopez-Otero
Pavel Golik
Pavel Ircing
Pavel Kral
Pavel Matejka
Pavol Partila
Pedro Gómez-Vilda
Pedro Torres-Carrasquillo
Pejman Mowlaee
Peng Shen
Peng Song
Pengfei Liu
pengyuan zhang
Penny Karanasou
Peter Assmann
Peter Bell
Peter Birkholz
Peter Cahill
Peter Mihajlik
Phil Green
Phil Woodland
Philip Harding
Philip N. Garner
Philippe Boula de Mareüil
Philippe Gournay
Phillip DeLeon
Phu Le
Piero Cosi
Pierre Badin
Pierre Lanchantin
Pierre Lison
Pierre-Edouard Honnet
Pierre-Michel Bousquet
Pietro Laface
Plinio Barbosa
Polina Drozdova
Pongtep Angkititrakul
Po-Sen Huang
Pranay Dighe
Prasanta Ghosh
Prathosh A P
Pratibha Moogi
Prem C. Pandey
Priyank Thakkar
Priyankoo Sarmah
Qiang Chen
Qiang Huang
Qin Jin
Qinghua Sun
Qingyang Hong
Qiong Hu
R Kumaraswamy
Rachel E. Bouserhal
Rafid Sukkar
Raheel Qader
Rahim Saeidi
Rahul Gupta
Rainer Jäckel
Rainer Martin
Rajbabu Velmurugan
Ralf Schlüter
Ralph Rose
Rama Rao
Rama Sanand Doddipatla
Ramón Astudillo
Ramon Lopez-Cozar
Ramya Rasipuram
Ranniery Maia
Rasmus Dall
Rathinavelu Chengalvarayan
Ratree Wayland
Ravichander Vipperla
Raymond Slyh
Raymond W. M. Ng
Reinhold Haeb-Umbach
Renato de Mori
Ricard Marxer
Ricardo Cordoba
Ricardo Ribeiro
Richard Dufour
Richard Hendriks
Richard Rose
Rita Singh
Rivka Levitan
Rob van Son
Robert A. Fox
Robert Clark
Robert Eklund
Robert Fuchs
Robert Östling
Robert Whitman
Roberto Barra-Chicote
Roberto Gemello
Roberto Pieraccini
Roberto Togneri
Robin Lickley
Rodrigo Guido
Roger Moore
Rogier van Dalen
Rohan Kumar Das
Rohit Prabhavalkar
Rohit Sinha
Roland Maas
Ron Hoory
Ronald Böck
Rongqing Huang
Rosa González Hautamäki
Rosario Signorello
Rubén Pérez Ramón
Rubén San Segundo Hernández
Rubén Solera-Ureña
Ruediger Hoffmann
Ruhi Sarikaya
Rui Xia
Rui Zhao
Rupayan Chakraborty
Ryan Price
Ryo Masumura
Ryuichiro Higashinaka
S R Mahadeva Prasanna
Sabato Marco Siniscalchi
Sabrina Stehwien
Saeed Bagheri Sereshki
Sakriani Sakti
Salil Deena
Salvatore Anzalone
Samantha Gordon Danner
Sameer Maskey
Samuel Kim
Samuel Thomas
Sandro Cumani
Sanghun Kim
Sanjeev Gurugopinath
Sanjika Hewavitharana
Sankaran Panchapagesan
Sara Myrberg
Sarala Padi
Sarangarajan Parthasarathy
Sascha Disch
Satoshi Imaizumi
Satoshi Kobashikawa
Satoshi Nakamura
Saurabh Sahu
Sayaka Shiota
Sayan Ghosh
Sayoko Takano
Scott Wisdom
Sebastian Stüker
Sébastien Le Maguer
Seiichi Nakagawa
Seokhwan Kim
Seongkyu Mun
Serdar Yildirim
Sergey Novoselov
Serguei Pakhomov
Seyed Omid Sadjadi
Seyedmahdad Mirsamadi
Shahina A
Shakuntala Mahanta
Shammur Absar Chowdhury

Shaofei Xue
Sharon Gannot
Shashidhar G Koolagudi -
Sheng Li
Shinichi Kawamoto
Shinichi Tokuma
Shinji Takaki
Shinji Watanabe
Shi-wook Lee
Shiyin Kang
SHK (Hari) Parthasarathi
Shmulik Markovich-Golan
Shoko Araki
Shrikant Joshi
Shu-Chuan Tseng
shyam agrawal
Silen Hanna
Simon Dobnik
Simon Dobrišek
Simon Hammond
Simon Keizer
Simon King
Simon Stone
Slava Shechtman
Snezhina Dimitrova
Sofia Strömbergsson
Sofoklis Kakouros
Solomon Teferra Abate
Sophie Dufour
Sophie Rosset
Spyros Matsoukas
Srikanth Madikeri
Sriram Ganapathy
Staffan Larsson
Stan Salvador
Štefan Beňuš
Stefan Hahn
Stefan Steidl
Stefan Ultes
Stefanie Jannedy
Stephan Kanthak
Stephen Shum
Stephen Voran
Stephen Zahorian
Steve Beet
Steve Renals
Steven Sandoval
Sumam David
Sunayana Sitaram
Sundararajan Srinivasan
Sungbok Lee
Sunil Kumar Kopparapu
Sunil Sivadas
Suryakanth V Gangashetty
Susann Luperfoy
Susanne Burger
Susanne Fuchs
Suwon Shon
Suzanne Boyce
Sven Nordholm
Sven Schmeier
Sven Shepstone
Syed Shahnawazuddin
Symeon Delikaris Manias
Szu-Chen Stan Jou
Szymon Drgas
T. J. Hazen
Takaaki Hori
Takafumi Koshinaka
Takahiro Shinozaki
Takahito Shinya
Takanobu Nishiura
Takao Kobayashi
Takashi Fukuda
Takashi Masuko
Takashi Nose
Takayuki Arai
Takehiro Moriya
Takemi Mochida
Takuya Hiraoka
Tamás Gábor Csapó
Tanaya Guha
Tanel Alumäe
Taniya Mishra
Tanvina Patel
Tara Sainath
Tatsuya Kawahara
Taufiq Hasan
Tekla Etelka Graczi
Teresa Lopez-Soto
Tetsuji Ogawa
Tetsuya Shimamura
Thangavelu Nagarajan
Theodora Chaspari
Thiago Fraga da Silva
Thierry Dutoit
Thomas Drugman
Thomas Fang Zheng
Thomas Hueber
Thomas Kleinbauer
Thomas Merritt
Thomas Niesler
Thomas Pellegrini
Thomas Quatieri
Thomas Schaaf
Thomsen Nicolai Bæk
Tiago Falk
Tien Dung Tran
Tien Ping Tan
Tilman Becker
Tim Fingscheidt
Tim Juergens
Tim Polzehl
Timo Baumann
Timo Gerkmannn
Tine Mooshammer
Ting Wang
Tino Haderlein
Tobias Bocklet
Tofigh Naghibi
Tom Bäckström
Tomáš Bořil
Tomi Kinnunen
Tomohiro Nakatani
Tomoki Koriyama
Tomoki Toda
Tomoko Matsui
Tony Robinson
Torsten Dau
Toru Nakashika
Toru Takahashi
Toshiko Isei-Jaakkola
Toshio Irino
Tran-Huy Dat
Tuomas Virtanen
Tuomo Raitio
Turaj Shabestary
Udo Zoelzer
Umit Yapanel
Unto K. Laine
V Ramasubramanian
Vaclav Eksler
Valentín Cardeñoso-Payo
Vataya Chunwijitra
Veena Thenkanidiyoor
Venki Nagesha
Vered Silber-Varod
Veronique Delvaux
Vesa Ruoppila
Victor Sorokin
Victoria Sanchez
Vidhyasaharan Sethu
Vijayaditya Peddinti
Vikas Joshi
Vikram Ramanarayanan
Vikramjit Mitra
Vikrant Singh Tomar
Viktor Rozgic
Ville Hautamaki
Vinay Kumar Mittal
Vinayak Abrol
Vincent Barriac
Vincent Wan
vishwa gupta
Vivek Kumar Rangarajan Sridhar
Volker Fischer
Volker Hohmann
Waad Ben Kheder
Wai Lok Woo
Walter Andrews
Wei Rao
Wei Zhang
Wendy Holmes
Wenju Liu
Wentao Gu
William Chan
William Hartmann
William Thorpe
William Yang Wang
Wim van Dommelen
Wissam Jassim
Wolfgang Hess
Wooil Kim
Xavier Anguera
Xiaodan Zhuang
Xiaodong Cui
Xiaoyun Wang
Xie Chen
Xie Sun
Xing Fan
Xinhui Zhou
Xiong Xiao
Xueliang Zhang
Xugang Lu
Xunying Liu
Ya Li
Yamato Ohtani
Yanhua Long
Yanmin Qian
Yannick Estève
Yannis Stylianou
Yanzhang He
Yao Qian
Yasuaki Shinohara
Yasuharu Den
Yasunari Obuchi
Yasuo Ariki
Yi Liu
Yi Su
Yi-Cheng Pan
Yifan Gong
Yiheng Huang
Yik-Cheung Tam
Yoichi Yamashita
Yonghong Yan
Yongqiang Wang
Yoshiko Arimoto
Yoshinori Sagisaka
Yoshinori Shiga
yosuke sugiura
Young-Sun Yun
Yu Tsao
Yu Zhang
Yue Pan
Yuexian Zou
Yuki Mitsufuji
Yuma Koizumi
Yun Tang
Yun-Nung Chen
Yunxin Zhao
Yu-Ren Chien
Yuri Matveev
Yusuke Hioka
Yusuke Ijima
Yutaka Kamamoto
Yuuki Tachioka
Yuya Akita
Yuzong Liu
Zdravko Kacic
Zheng-Hua Tan
Zhen-Hua Ling
Zhijian Ou
Zhiyong Wu
Zinny Bond
Zixing Zhang
Zofia Malisz
Zoltán Tüske
Zoraida Callejas

# Future INTERSPEECH Conferences



# Satellite Workshops and Events

## SATELLITE EVENTS

### 1. Speech Processing in Challenging Environments (SPICE)

September 8, 2018, IISc Bangalore, India
URL: https://spiceworkshop2018.wixsite.com/spice2018

**Organizers**

Anil Kumar Chilli, CAIR, DRDO
Anupam Mandal, CAIR, DRDO
Prasanna Kumar K R, CAIR, DRDO
Sriram Ganapathy, IISc, Bangalore
Prasanta Kumar Ghosh, IISc, Bangalore
Dr. S. Sethu Selvi, Electronics and Communications Department of MSRIT

### 2. Summer School on Speech Production

September 9-11, 2018, DA-IICT Gandhinagar
URL: https://sites.google.com/site/s4p2018daiict/home

**Organizers**

Hemant A. Patil, DA-IICT Gandhinagar
Suman K. Mitra, DA-IICT Gandhinagar
Prasenjit Majmuder, DA-IICT Gandhinagar
Yash Vasavada, DA-IICT Gandhinagar
Soman Nair, DA-IICT Gandhinagar

# SATELLITE WORKSHOPS

## 1. 6th International Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU'18)

August 29-31, 2018, New Delhi, Gurgaon, India
URL: http://www.mica.edu.vn/sltu2018/

**Organizers**

Shyam S Agarwal, KIIT, India
Laurent Besacier, LIG, France
Sakriani Sakti, NAIST, Japan
Neelima Kamrah, KIIT, Gurgaon, India
Hemant A. Patil, DA-IICT Gandhinagar
Sunayana Sitaram, Microsoft India

## 2. Speech, Music and Mind 2018 (SMM18): Detecting and Influencing Mental States with Audio

September 1, 2018, IIIT Hyderabad, India
URL: http://smmw.iiit.ac.in/

**Organizers**

Venkata Subramanian Viraraghavan, TCS Research and Innovation
Suryakanth V Gangashetty, IIIT Hyderabad
Balamuralidhar P, TCS Research and Innovation
Sunil Kopparapu, TCS Research and Innovation

## 3. English in India and Indian Englishes: New Horizons in the Study of Phonetics and Phonology

September 7, 2018, University of Hyderabad.
URL: https://phinde1.wordpress.com/blog/phonology-of-indian-englishes-2018/

**Organizers**

Robert Fuchs, University of Hamburg, Germany
Pingali Sailaja, University of Hyderabad, India

## 4. Workshop on Machine Learning in Speech and Language Processing (MLSLP)

September 7, 2018, Google, Hyderabad, India
URL: https://sites.google.com/view/mlslp/home

**Organizers**

Preethi Jyothi, Indian Institute of Technology Bombay
Rohit Prabhavalkar, Google Inc.
Liang Lu, Microsoft Inc.
Tara Sainath, Google Inc.
Negar Saei, Google Inc.

## 5. The 5th International Workshop on Speech Processing in Everyday Environments (CHiME 2018)

September 07, 2018, Microsoft Hyderabad campus
URL: http://spandh.dcs.shef.ac.uk/chime_workshop/

**Organizers**

Jon Barker, Univ. of Sheffield, UK
Shinji Watanabe, Johns Hopkins University, USA
Emmanuel Vincent, INRIA, France

## 6. Workshop on Speech Processing for Voice, Speech and Hearing Disorders

September 8-9, 2018, All India Institute of Speech and Hearing, Mysore, India
URL: http://wspd.co.in/

**Organizers**

Ajish K. Abraham, All India Institute of Speech & Hearing, Mysore
S R M Prasanna, IIT Dharwad, India
Rainer Martin, Institute of Communication Acoustics, Ruhr University, Germany
Yannis Stylianou, University of Crete, Greece
K. T. Deepak, IIIT Dharwad, India

## 7. Blizzard Challenge Workshop 2018

September 8, 2018, Microsoft, Hyderabad, India
URL: https://www.synsig.org/index.php/Blizzard_Challenge_2018

**Organizers**

Alan Black, Carnegie Mellon University, USA
Keiichi Tokuda, Nagoya Institute of Technology, Japan
Simon King, University of Edinburgh, UK

# Tutorials

The INTERSPEECH 2018 Tutorial committee comprising V. Ramasubramanian, Yannis Stylianou, G. Panayiotis and Israel Cohen is pleased to announce the following eight pre-conference tutorials. They will be offered on Sunday, September 2, 2018. All tutorials will be of 3.5 hours duration including a 30 minute break.

## September 2, 2018
## FORENOON TUTORIALS

(Parallel sessions; 09:00 - 12.30 ; Coffee/tea break: 10.30 - 11.00)

### Tutorial 1 (T1): Deep Learning Based Speech Separation

DeLiang Wang, The Ohio State University

**Abstract:** Speech separation is the task of separating target speech from background interference. In contrast to the traditional signal processing perspective, speech separation can be formulated as a supervised learning problem, where the discriminative patterns of speech, speakers, and background noise are learned from training data. The recent introduction of deep learning to supervised speech separation has dramatically accelerated progress and boosted separation performance, including the first demonstration of substantial speech intelligibility improvements by hearing impaired listeners in noisy environments. This tutorial provides a comprehensive overview of the research on deep learning based supervised speech separation in the last several years. We systematically introduce three main components of supervised separation: learning machines, training targets, and acoustic features. Much of the tutorial will be on separation algorithms where we describe monaural methods, including speech enhancement (speech-nonspeech separation), speaker separation (multi-talker separation), and speech dereverberation, as well as multi-microphone techniques. In addition, we discuss a number of conceptual issues, including what constitutes the target source.

**DeLiang Wang** received the B.S. degree in 1983 and the M.S. degree in 1986 from Peking (Beijing) University, Beijing, China, and the Ph.D. degree in 1991 from the University of Southern California, Los Angeles, CA, all in computer science. Since 1991, he has been with the Department of Computer Science & Engineering and the Center for Cognitive and Brain Sciences at Ohio State University, Columbus, OH, where he is currently a Professor and University Distinguished Scholar. He also holds a visiting appointment at the Center of Intelligent Acoustics and Immersive Communications, Northwestern Polytechnical University, Xi'an, China. He has been a visiting scholar to Harvard University, Oticon A/S (Denmark), and Starkey Hearing Technologies. Wang's research interests include machine perception and deep neural networks. Among his recognitions are the Office of Naval Research Young Investigator Award in 1996, the 2005 Outstanding Paper Award from IEEE Transactions on Neural Networks, and the 2008 Helmholtz Award from the International Neural Network Society. He serves as Co-Editor-in-Chief of Neural Networks, and on the editorial boards of several journals. He is an IEEE Fellow.

## Tutorial 2 (T2): End-To-End Models for Automatic Speech Recognition

Rohit Prabhavalkar, Google Inc., USA
Tara Sainath, Google Inc., USA

**Abstract:** Traditional automatic speech recognition (ASR) systems are comprised of a set of separate components, namely an acoustic model (AM); a pronunciation model (PM); and a language model (LM). The AM takes acoustic features as input and predicts a distribution over subword units, typically context-dependent phonemes. The PM, which is traditionally a hand-engineered lexicon maps the sequence of subword units produced by the acoustic model to words. Finally, the LM assigns probabilities to various word hypotheses. In traditional ASR systems, these components are trained independently on different datasets, with a number of independence assumptions which are made for tractability.

Over the last several years, there has been a growing interest in developing end-to-end systems, which attempt to learn these separate components jointly in a single system. Examples of such system include attention-based models [1, 6], the recurrent neural transducer [2, 3], the recurrent neural aligner [4], and connectionist temporal classification with word targets [5]. A common feature of all of these models is that they are composed on a single neural network, which when given input acoustic frames directly outputs a probability distribution over graphemes or word hypotheses. In fact, as has been demonstrated in recent work, such end-to-end models can surpass the performance of a conventional ASR systems [6].

In this tutorial, we will provide a detailed introduction to the topic of end-to-end modeling in the context of ASR. We will begin by charting out the historical development of these systems, while emphasizing the commonalities and the differences between the various end-to-end approaches that have been considered in the literature. We will then discuss a number of recently introduced innovations that have significantly improved the performance of end-to-end models, allowing these to surpass the performance of conventional ASR systems. The tutorial will then describe some of the exciting applications of this research, along with possible fruitful directions to explore.

Finally, the tutorial will discuss some of the shortcomings of existing end-to-end modeling approaches and discuss ongoing efforts to address these challenges.

[1] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, Attend and Spell," in Proc. ICASSP, 2016.
[2] A. Graves, "Sequence transduction with recurrent neural networks," in Proc. of ICASSP, 2012.
[3] K. Rao, H. Sak, and R. Prabhavalkar, "Exploring Architectures, Data and Units For Streaming End-to-End Speech Recognition with RNN-Transducer," in Proc. ASRU, 2017.
[4] H. Sak, M. Shannon, K. Rao, and F. Beaufays, "Recurrent neural aligner: An encoder-decoder neural network model for sequence-to-sequence mapping," in Proc. Interspeech, 2017.
[5] H. Soltau, H. Liao, and H. Sak, "Neural speech recognizer: acoustic-to-word LSTM model for large vocabulary speech recognition," in Proc. of Interspeech, 2017.
[6] C.C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, N. Jaitly, B. Li, J. Chorowski and M. Bacchiani, "State-of-the-art Speech Recognition With Sequence-to-Sequence Models," in Proc. ICASSP, 2018.

**Rohit Prabhavalkar** received the B.E. degree in computer engineering from the University of Pune, India, in 2007 and the M.S. and Ph.D. degrees in computer science and engineering from The Ohio State University, USA, in 2012 and 2013, respectively. He is currently a research scientist at Google, having joined in 2013, where his research focuses on various aspects of acoustic modeling with the goal of improving automatic speech recognition technology. His other research interests include deep neural networks, natural language processing, and machine learning.

**Tara Sainath** received her PhD in Electrical Engineering and Computer Science from MIT in 2009. The main focus of her PhD work was in acoustic modeling for noise robust speech recognition. After her PhD, she spent 5 years at the Speech and Language Algorithms group at IBM T.J. Watson Research Center, before joining Google Research. She has served as a Program Chair for ICLR in 2017 and 2018. Also, she has co-organized numerous special sessions and workshops, including Interspeech 2010, ICML 2013, Interspeech 2016 and ICML 2017. In addition, she is a member of the IEEE Speech and Language Processing Technical Committee (SLTC) as well as the Associate Editor for IEEE/ACM Transactions on Audio, Speech, and Language Processing. Her research interests are mainly in acoustic modeling, including deep neural networks, sparse representations and adaptation methods.

## Tutorial 3 (T3): Spoofing Attacks in Automatic Speaker Verification: Analysis and Countermeasures

Haizhou Li, National University of Singapore, Singapore
Hemant A. Patil, Dhirubhai Ambani Institute of Information and Communication Technology, India
Nicholas Evans, EURECOM, France

**Abstract:** Speech is the most natural means of communication between humans. Speech signals carry various levels of information, such as linguistic content, emotion, the acoustic environment, language, the speaker's identity and their health condition, etc. Automatic speaker recognition technologies aim to verify or identify a speaker using recordings of his/her voice. In practice, automatic speaker verification (ASV) systems should be robust to nuisance variation such as differences in the microphone and transmission channel, intersession variability, acoustic noise, speaker ageing, etc. Significant effort invested over the last three decades has been tremendously successful in developing technologies to compensate for such nuisance variation, thereby improving the reliability of ASV systems in a multitude of diverse application scenarios. In a number of these, specifically those relating to authentication applications, reliability can still be compromised as a result of spoofing attacks whereby fraudsters can gain illegitimate access to protected resources or facilities through the presentation of specially crafted speech signals that reflect the characteristics of another, enrolled person's voice. ASV systems should be resilient to such malicious spoofing attacks. This tutorial presents a treatment of the issues concerning the robustness and security of an ASV system in the face of spoofing attacks. We also discuss current research trends and progress in developing anti-spoofing countermeasures to protect against attacks derived from voice conversion, speech synthesis, replay, twins (which has more malicious nature in attacking ASV systems and also called as twin's fraud in biometrics literature) and professional mimics. The tutorial will give an overview of the risk and technological challenges associated with each form of attack in addition to an overview of the two internationally competitive ASVspoof challenges held as special sessions at INTERSPEECH 2015 and INTERSPEECH 2017. The tutorial will conclude with a summary of the current state-of-the-art in the field and a discussion of future research directions.

**Core Topics Covered**

- ASV design cycle, research issues and technological challenges
- Different spoofing attacks: voice conversion, speech synthesis, replay and twins, professional mimicry
- ASVspoof 2015 Challenge, INTERSPEECH 2015 – Objective, Database and Results
- ASVspoof 2017 Challenge, INTERSPEECH 2017 – Objective, Database and Results
- ASVspoof 2019 Roadmap
- Technological Challenges in Replay Spoof Speech Detection (SSD)
- Strategies to combined spoofing countermeasures with ASV

**Haizhou Li** is currently a Professor at the Department of Electrical and Computer Engineering, National University of Singapore. Professor Li's research interests include speech information processing and natural language processing. He is currently the Editor-in-Chief of IEEE/ACM Transactions on Audio, Speech and Language Processing (2015-2018). Professor Li was the President of the International Speech Communication Association (2015-2017), the President of Asia Pacific Signal and Information Processing Association (2015-2016). Professor Li is a Fellow of the IEEE. He was a recipient of the President's Technology Award 2013 in Singapore.

**Hemant A. Patil** is a Professor at DA-IICT Gandhinagar, India. Prof. Patil's research interests include speaker recognition, spoofing attacks, TTS, infant cry analysis, etc. He developed Speech Research Lab at DA-IICT, which is recognized as ISCA speech labs. Dr. Patil is member of IEEE, IEEE Signal Processing Society, IEEE Circuits and Systems Society, International Speech Communication Association (ISCA), EURASIP and an affiliate member of IEEE SLTC. He visited department of ECE, University of Minnesota, Minneapolis, USA (May-July, 2009) as short-term scholar. Dr. Patil has taken a lead role in organizing several ISCA supported events at DA-IICT. Dr. Patil has supervised 03 doctoral and 37 M.Tech. theses. Recently, he offered a joint tutorial with Prof. Haizhou Li during Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) 2017. He has been selected as APSIPA Distinguished Lecturer (DL) for 2018-2019. He delivered APSIPA Distinguished Lecture at University of Calgary, Canada, SJTU, Shanghai, China, KIIT Gurgaon, India and SGGSIE&T Nanded, India.

**Nicholas Evans** is a Professor at EURECOM, France where he heads research in Speech and Audio Processing within the Biometrics and Digital Media Research Group of the Digital Security Department. His current research aims to develop new countermeasures to protect automatic speaker verification technology from the threat of spoofing. He is among the co-founders of the ASVspoof initiative and evaluation series (2015 & 2017) and was Lead Guest Editor for the IEEE Transactions on Information Forensics and Security special issue in Biometrics Spoofing and Countermeasures, Lead Guest Editor for the IEEE SPM special issue on Biometric Security and Privacy and Guest Editor for the IEEE JSTSP special issue on Spoofing and Countermeasures for Automatic Speaker Verification. He delivered as an invited speaker a talk on Spoofing and Countermeasures for Automatic Speaker Verification at the IEEE SPS Winter School on Security and Privacy Issues in Biometrics and contributed to a tutorial on Spoofing and Anti-Spoofing: A Shared View of Speaker Verification, Speech Synthesis and Voice Conversion at APSIPA ASC 2015. He is currently co-editing the 2nd edition of the Handbook of Biometric Anti-spoofing to be published in 2018.

## Tutorial 4 (T4): Spoken Dialog Technology for Education Domain Applications

Vikram Ramanarayanan, Keelan Evanini and David Suendermann-Oeft
Educational Testing Service R&D, San Francisco, USA

**Abstract:** This workshop will introduce participants to the basics of designing conversational applications in the educational domain using spoken and multimodal dialog technology. The increasing maturation of automated conversational technologies in recent years holds much promise towards developing intelligent agents that can guide one or multiple phases of student instruction, learning, and assessment. In language learning applications, using spoken dialogue systems (SDS) could be an effective solution to improving conversational skills, because an SDS provides a convenient means for people to both practice and obtain feedback on different aspects of their conversational skills in a new language. These allow learners to make mistakes without feeling incompetent, empowering them to improve their skills for when they do speak with native speakers. From the assessment perspective, well-designed dialog agents have potential to elicit and evaluate the full range of English speaking skills (such as turn taking abilities, politeness strategies, pragmatic competence) that are required for successful communication. Such technologies can potentially personalize education to each learner, providing a natural and practical learning interface that can adapt to their individual strengths and weaknesses in real time so as to increase the efficacy of instruction.

The workshop will assume no prior knowledge of dialog technology or intelligent tutoring systems and will demonstrate the use of open-source software tools in building conversational applications. The first part of the tutorial will cover the state of the art in dialog technologies for educational domain applications, with a particular focus on language learning and assessment. This will include an introduction to the various components of spoken dialog systems and how they can be applied to develop conversational applications in the educational domain, as well as some advanced topics such as methods for speech scoring. The final part of the tutorial will be specifically dedicated to a hands-on application building session, where participants will have a chance to design and deploy their own dialog application from scratch on the HALEF cloud-based dialog platform using open-source OpenVXML design toolkit, which will allow a better understanding how such systems can potentially be designed and built. Participants are required to bring their own laptops for this portion of the workshop (with Windows or Mac operating systems installed; Linux not currently supported). Additional software installation instructions will be sent prior to the workshop.

**Vikram Ramanarayanan** is a Research Scientist at Educational Testing Service's R&D division in San Francisco and also holds an Assistant Adjunct Professor appointment in the Department of Otolaryngology — Head and Neck Surgery at the University of California, San Francisco. Vikram's research interests lie in applying scientific knowledge to interdisciplinary engineering problems in speech, language and vision and in turn using engineering approaches to drive scientific understanding. He has authored over 70 papers in peer-reviewed journals and conference proceedings. His work on speech science and technology has won 2 Best Paper Awards, an Editor's Choice Award and an ETS Presidential Award. He holds M.S and Ph.D. degrees in Electrical Engineering from the University of Southern California, Los Angeles. Personal Webpage: http://vikramr.com

**Keelan Evanini** is a Research Director at Educational Testing Service in Princeton, NJ. His research interests include automated assessment of non-native spoken English, automated feedback in computer assisted language learning applications, and spoken dialog systems. He leads the research team that develops SpeechRater, the ETS capability for automated spoken language assessment. He also leads the team of research engineers that work on applying state-of-the-art natural language processing, speech processing, and

machine learning technology to a wide range of research projects aimed at improving assessment and learning. He received his Ph.D. from the University of Pennsylvania in 2009, has published over 70 papers in peer-reviewed journals and conference proceedings, has been awarded 9 patents, and is a senior member of the IEEE. Personal Webpage: http://www.evanini.com/keelan.html

**David Suendermann-Oeft** is director of the Dialog, Multimodal, and Speech (DIAMONDS) research center and manager of the San Francisco Office at Educational Testing Service (ETS). He is also the director, co-founder, and chief scientist of EMR.AI Inc., a company providing AI solutions to the medical sector, headquartered in San Francisco. Throughout his career, he has been working in the field of spoken language processing, machine learning, and artificial intelligence at multiple industrial and academic institutions including Siemens (Munich and Mexico City), RWTH (Aachen), UPC (Barcelona), USC (Los Angeles), Columbia University (New York), SpeechCycle (New York), Synchronoss (New York), and ICSI (Berkeley) and has also held appointments as tenured professor of computer science and department head at DHBW Stuttgart. David has authored over 140 publications and patents including two books and ten book chapters and serves as evaluator, rapporteur, and innovation expert for the European Commission. Personal Webpage: http://suendermann.com

## September 2, 2018
## AFTERNOON TUTORIALS

(Parallel sessions; 14:00 - 17:30; Coffee/tea break: 15:30 – 16:00)

### Tutorial 5 (T5): Information Theory of Deep Learning: What do the Layers of Deep Neural Networks Represent ?

*(Interactive Video Mode)*

Naftali Tishby, Hebrew University of Jerusalem

**Abstract:** In this tutorial I will present a novel comprehensive theory of large scale learning with Deep Neural Networks, based on the correspondence between Deep Learning and the Information Bottleneck framework. The new theory has the following components: (1) rethinking Learning theory; I will prove a new generalization bound, the input-compression bound, which shows that compression of the representation of input variable is far more important for good generalization than the dimension of the network hypothesis class, an ill-defined notion for deep learning. (2) I will prove that for large scale Deep Neural Networks the mutual information on the input and the output variables, for the last hidden layer, provide a complete characterization of the sample complexity and accuracy of the network. This makes the information Bottleneck bound for the problem as the optimal trade-off between sample complexity and accuracy with ANY learning algorithm. (3) I will show how Stochastic Gradient Descent, as used in Deep Learning, achieves this optimal bound. In that sense, Deep Learning is a method for solving the Information Bottleneck problem for large scale supervised learning problems. The theory provides a new computational understating of the benefit of the hidden layers and gives concrete predictions for the structure of the layers of Deep Neural Networks and their design principles. These turn out to depend solely on the joint distribution of the input and output and on the sample size.

Based partly on joint works with Ravid Shwartz-Ziv, Noga Zaslavsky, and Amichai Painsky.

**Naftali Tishby** is a professor of Computer Science, and the incumbent of the Ruth and Stan Flinkman Chair for Brain Research at the Edmond and Lily Safra Center for Brain Science (ELSC) at the Hebrew University of Jerusalem. He is one of the leaders in machine learning research and computational neuroscience in Israel, and his numerous former students serve in key academic and industrial research positions all over the world. Tishby was the founding chair of the new computer-engineering program, and a director of the Leibnitz Center for Research in Computer Science at Hebrew University. Tishby received his PhD in theoretical physics from Hebrew University in 1985 and was a research staff member at MIT and Bell Labs from 1985 to 1991. Tishby has been a visiting professor at Princeton NECI, the University of Pennsylvania, UCSB, and IBM Research.

## Tutorial 6 (T6): Articulatory Representations: Measurement, Estimation, and Application to State-of-the-Art Automatic Speech Recognition

Carol Espy-Wilson, University of Maryland, USA
Mark Tiede, Haskins Lab, USA
Hosung Nam, Korea University, Seoul
Vikramjit Mitra, Apple Inc., USA
Ganesh Sivaraman, Pindrop, Atlanta, USA

**Abstract:** Articulatory representations have been studied and applied to various speech technologies for many years. One of the persisting challenges of articulatory research has been the paucity of reliable articulatory data and non-scalability to large scale subject-independent applications. The aim of this tutorial is to present an overview of research in articulatory representations for large scale applications. This tutorial will discuss best practices in articulatory data collection and synthesis, present recent developments in subject independent acoustic-to-articulatory speech inversion and describe state-of-the-art Convolutional Neural Network (CNN) architectures for large vocabulary continuous speech recognition incorporating both articulatory and acoustic features. This tutorial combines experts from scientific and engineering backgrounds to present a concise tutorial about articulatory features, their measurement, synthesis and application to state-of-the-art Automatic Speech Recognition.

**Carol Espy-Wilson,** the lead organizer is a Professor in the Department of Electrical and Computer Engineering and the Institutes for Systems Research at the University of Maryland, College Park. She has more than 3 decades of leading research endeavors in fundamental speech acoustics, vocal tract modeling, speech and speaker recognition, speech segregation, speech enhancement, and more recently speech inversion. One of the tools coming out of Dr. Espy-Wilson's lab is a vocal tract modeling tool, VTAR that many scientists and engineers have downloaded to use for research and teaching. Dr. Espy-Wilson has advised a number of PhD and MS students and published many papers in reputed academic conferences and journals.

**Mark Tiede** is a Senior Scientist at Haskins Laboratories active in speech production research, particularly the study of dyadic interaction. He has more than 20 years experience in the use of point source tracking methods such as electromagnetic articulography (EMA) and is the author of a widely used tool for the use of analyzing such data (mview).

**Hosung Nam** is an Assistant Professor in the Department of English Language and Literature at Korea University, Seoul, South Korea. He received the M.S. and Ph.D. degrees from the Department of Linguistics at Yale University, New Haven, CT, in 2007. He is a linguist who is an expert in the field of articulatory phonology. His research emphasis is on the link between speech perception and production, speech error, automatic speech recognition, sign language, phonological development, and their computational modeling. He has been a Research Scientist at Haskins Laboratories, New Haven, since 2007.

**Vikramjit Mitra** is a Research Scientist at Apple Inc. He received his Ph.D. in Electrical Engineering from University of Maryland, College Park; M.S. in Electrical Engineering from University of Denver, B.E. in Electrical Engineering from Jadavpur University, India. His research focuses on signal processing for noise/channel/reverberation, speech recognition, production/perception-motivated signal processing, information retrieval, machine learning and speech analytics. He is a senior member of the IEEE, an affiliate member of the SLTC and has served on NSF panels.

**Ganesh Sivaraman** is a Research Scientist at Pindrop. He received his M.S. and Ph.D. in Electrical Engineering from University of Maryland College Park, B.E. in Electrical Engineering from Birla Institute of Technology and Science, Pilani, India. His research focuses on speaker independent acoustic-to-articulatory inversion, speaker adaptation, speech enhancement and robust speech recognition.

## Tutorial 7 (T7): Vocal Music Processing for Music Information Retrieval (MIR)

Preeti Rao, Indian Institute of Technology Bombay, Mumbai
Hema Murthy, Indian Institute of Technology Madras, Chennai

**Abstract:** The singing voice is the most flexible of musical instruments and, unsurprisingly, vocal performance has dominated the music of many cultures. Information extraction from the singing voice requires dealing with the structure and semantics of music which are quite distinct from that of spoken language. Like speech, MIR is an interdisciplinary field where signal processing and computing experts benefit from interaction with musicologists and psychologists working in music cognition. Extracting musically useful information from vocal music recordings benefits real-world applications such as music recommendation, search and navigation, musicology studies and evaluation of singing skill. The relevant signal-level tasks include vocal activity detection and separation in polyphonic music, song segmentation, extracting voice features related to melody, rhythm and timbre and establishing models for perceived similarity across the possibly culture-specific musical dimensions. In this tutorial, we will present the audio signal processing and machine learning methods that underlie a variety of MIR applications with examples drawn from Western popular and Indian classical genres. The goal of the tutorial is to show speech researchers how they can contribute fruitfully to MIR problems which are considered topical and rewarding.

**Preeti Rao** has been on the faculty of the Department of Electrical Engineering, I.I.T. Bombay since 1999. She currently serves as H.A.L. R&D Chair Professor. She received the B.Tech. EE degree from I.I.T. Bombay in 1984, and Ph.D. from the University of Florida, Gainesville in 1990. After post-doctoral stints at the University of Illinois at Urbana-Champaign and Hitachi Research Labs in Tokyo, she joined the EE department at I.I.T. Kanpur in 1994, before moving to I.I.T. Bombay in 1999. Her research interests lie in speech and audio signal processing. She previously worked on low bit rate speech compression algorithms for the sub 1 kbps rates. More recently, she has been involved in computer aided spoken language training, prosody modeling for Indian languages and multichannel speech enhancement. She took up research in music computing about a decade ago and this effort received a big boost through engagement with the ERC funded CompMusic project led by UPF Barcelona in 2011-2016, towards culture-specific music information extraction with a focus on Indian classical genres. Her team of students also participated successfully in melody extraction and vocal separation challenges hosted by MIREX (MIR Evaluations) during this period. She is currently on the Editorial Board of the Journal of New Music Research (JNMR). She has been actively involved in the development of systems for Indian music and spoken language learning applications. She co-founded SensiBol Audio Technologies, a start-up focusing on music learning applications, with her Ph.D. and Masters students in 2011.

**Hema Murthy** is a Professor in the Department of Computer Science and Engineering, I.I.T Madras. She received the B.E (Electronics and Communication Engineering) from Osmania University (1980), M.Eng (Electrical and Computer Engineering) from McMaster University in 1986, and Ph.D (Computer Science and Engineering) from I.I.T Madras in 1992. She worked as Scientific Officer at the Tata Institute of Fundamental Research from 1980-83 where she worked in computer graphics. She joined as faculty in the Department of Computer Science and Engineering in 1988 as a lecturer. Her primarily line of work is in "signal processing directed machine learning," where the objective is to understand a given domain, apply appropriate signal processing

techniques to enable faster convergence of machine learning algorithms. She has worked in various domains including education, text, networks, brain signals, speech and music. Her major focus currently is on speech, music and brain signals. Her involvement in the ERC funded project on CompMusic where she was responsible for the analysis of Carnatic music gave an impetus to the effort on Carnatic music. The shifting tonic, and significant extempore improvisation in Indian music were some of the genre specific challenges that were addressed.She also led a consortium effort on text to speech synthesis where the objective was to build text to speech synthesis for 13 Indian languages.This technology has been transferred to a number of companies.

## Tutorial 8 (T8): Multimodal Speech and Audio Processing in Audio-Visual Human-Robot Interaction

Petros Maragos, School of E.C.E., National Technical University of Athens, Athens 15773, Greece
Athanasia Zlatintsi, Athena Research Center, Robot Perception and Interaction Unit, Greece

**Abstract:** The goal of this tutorial is to provide a concise overview of ideas, methods and research results in multimodal speech and audio processing, spatio-temporal sensory processing, perception and fusion, with applications in Human-Robot Interaction. Nowadays, most data are multimodal, thus there is the emergent need of developing multimodal methodologies, taking also into account the visual modality so as to enhance and assist the audio/speech modality. This tutorial will present state-of-the-art work for the major application area, which is Human-Robot Interaction, for social, edutainment and healthcare applications, including among others audio-gestural recognition for natural communication with the robotic agent and audio-visual speech synthesis for assistance and maximization of the naturalness of the interaction. Established results and recent advances from our research in various EU projects concerning the above areas as well as for the purposes of distant-speech interaction for robust home applications will also be discussed. Additionally, it will present a secondary application area that also relies on audio-visual processing, including in this case methodologies for saliency detection and automatic summarization of mono-modal or multimodal data (i.e., audio or video) and for the development of virtual interactive environments, where human body motion or hand gestures are used for audio-gestural music synthesis. Related papers and current results can be found in http://cvsp.cs.ntua.gr and http://robotics.ntua.gr

**Petros Maragos** received the M.Eng. Diploma in E.E. from the National Technical University of Athens (NTUA) in 1980 and the M.Sc. and Ph.D. degrees from Georgia Tech, Atlanta, in 1982 and 1985. In 1985, he joined the faculty of the Division of Applied Sciences at Harvard University, Boston, where he worked for eight years as professor of electrical engineering, affiliated with the Harvard Robotics Lab. In 1993, he joined the faculty of the School of ECE at Georgia Tech, affiliated with its Center for Signal and Image Processing. During periods of 1996-98 he had a joint appointment as director of research at the Institute of Language and Speech Processing in Athens. Since 1999, he has been working as professor at the NTUA School of ECE, where he is currently the director of the Intelligent Robotics and Automation Lab. He has held visiting positions at MIT in 2012 and at UPenn in 2016. His research and teaching interests include signal processing, systems theory, machine learning, image processing and computer vision, audio-speech & language processing, and robotics. In the above areas he has published numerous papers, book chapters, and has also co-edited three Springer research books, one on multimodal processing and two on shape analysis. He has served as: Member of the IEEE SPS committees on DSP, IMDSP and MMSP; Associate Editor for the IEEE Transactions on ASSP and the Transactions on PAMI, as well as editorial board member and guest editor for several journals on signal processing, image analysis and vision; Co-organizer of several conferences and workshops, including ECCV 2010 (Program Chair), IROS 2015 Workshop on Cognitive Mobility Assistance Robots, and EUSIPCO 2017 (General Chair). He has also served as member of the Greek National Council for Research and Technology. He is the recipient or co-recipient of several awards for his academic work, including: a 1987-1992 US NSF Presidential Young Investigator Award; the 1988 IEEE ASSP Young Author Best Paper Award; the 1994 IEEE SPS Senior Best Paper Award; the 1995 IEEE W.R.G. Baker Prize for the most outstanding original paper; the 1996 Pattern Recognition Society's Honorable Mention best paper award; the best paper award from the CVPR-2011 Workshop on Gesture Recognition. In

1995, he was elected IEEE Fellow for his research contributions. He received the 2007 EURASIP Technical Achievement Award for contributions to nonlinear signal processing, systems theory, image and speech processing. In 2010 he was elected Fellow of EURASIP for his research contributions. He has been elected IEEE SPS Distinguished Lecturer for 2017-2018.

**Athanasia Zlatintsi** is a Senior Researcher at the School of Electrical and Computer Engineering, National Technical University of Athens (NTUA), Greece. She received the Ph.D. degree in Electrical and Computer Engineering from NTUA in 2013 and her M.Sc. in Media Engineering from Royal Institute of Technology (KTH, Stockholm, Sweden) in 2006. Since 2007 she has been a researcher at the Computer Vision, Speech Communication & Signal Processing Group (CVSP) at NTUA participating in different projects, funded by the European Commission and Greek Ministry of Education in the areas of music and audio signal processing and specifically the analysis of music signals using computational methods aiming in their automatic recognition. Her research interests include signal processing and analysis of musical signals, analysis and processing of monomodal and multimodal signals for the extraction of robust representations for detection of perceptually salient events, speech/multimodal processing for human-computer interaction, and machine learning. Her research contributions include the development of efficient algorithms (using also nonlinear methods) for the analysis of the structure and the characteristics of musical signals, the detection of saliency in audio signals in general, the creation of multimodal, audio and music summaries, as well as the development of multimodal human-computer interaction systems.

# ISCA Medalist

**Bishnu S. Atal**
Department of Electrical Engineering
University of Washington, Seattle, WA 98195
https://www.ee.washington.edu/people/bishnu-atal/

## From Vocoders to Code-Excited Linear Prediction: Learning How We Hear What We Hear

**Abstract**
It all started almost a century ago, in 1920s. A new undersea transatlantic telegraph cable had been laid. The idea of transmitting speech over the new telegraph cable caught the fancy of Homer Dudley, a young engineer who had just joined Bell Telephone Laboratories. This led to the invention of Vocoder - its close relative Voder was showcased as the first machine to create human speech at the 1939 New York World's Fair. However, the voice quality of vocoders was not good enough for use in commercial telephony. During the time speech scientists were busy with vocoders, several major developments took place outside speech research. Norbert Wiener developed a mathematical theory for calculating the best filters and predictors for detecting signals hidden in noise. Linear Prediction or Linear Predictive Coding became a major tool for speech processing. Claude Shannon established that the highest bit rate in a communication channel in presence of noise is achieved when the transmitted signal resembles random white Gaussian noise. Shannon's theory led to the invention of Code-Excited Linear Prediction (CELP). Nearly all digital cellular standards as well as standards for digital voice communication over the Internet use CELP coders. The success in speech coding came with understanding of what we hear and what we do not. Speech encoding at low bit rates introduce errors and these errors must be hidden under the speech signal to become inaudible. More and more, speech technologies are being used in different acoustic environments raising questions about the robustness of the technology. Human listeners handle situations well when the signal at our ears is not just one signal, but also a superposition of many acoustic signals. We need new research to develop signal-processing methods that can separate the mixed acoustic signal into individual components and provide performance similar or superior to that of human listeners.

**Biography**
Bishnu S. Atal is an Affiliate Professor in the Electrical Engineering Department at the University of Washington, Seattle, WA. Born in India, Atal received his bachelor's degree in physics from the University of Lucknow, Diploma of the Indian Institute of Science, Bangalore, and a Ph.D. in electrical engineering from Brooklyn Polytechnic Institute. He joined Bell Laboratories in

1961, where he researched speech and acoustics until retiring in 2002. Atal holds more than 16 patents. Inspired by the high cost of long-distance phone calls to his family in India when he first moved to the U.S., Atal's research led to the invention of efficient digital speech coders and standards that lie at the heart of practically every mobile phone in use today. His work has enabled wireless networks to use less spectrum space and fewer towers, enabling even countries without substantial fiber-optic infrastructures to join the mobile revolution. He is a member of the U.S. National Academy of Sciences and National Academy of Engineering. His many honors include the IEEE Jack S. Kilby Signal Processing Medal (2013), the Benjamin Franklin Medal in Electrical Engineering (2003), the Thomas Edison Patent Award (1994), the New Jersey Hall of Fame Inventor of the Year Award (2000), and the IEEE Morris N. Liebmann Memorial Field Award (1986). Bishnu resides in Mukilteo, Washington. He has two daughters, Alka and Namita, two granddaughters, Jyotica and Sonali, and two grandsons, Ananth and Niguel.

# Plenary Speakers

**Jacqueline Vaissière**
Professor Emeritus, Université Sorbonne Nouvelle, France
http://www.univ-paris3.fr/vaissiere-jacqueline-29931.kjsp

## Universal Tendencies for Cross-Linguistic Prosodic Tendencies : A Review and Some New Proposals

**Abstract**
The present talk aims first to review the literature on similar tendencies regularly observed in typologically unrelated languages. The tendencies concern the use of fundamental frequency (F0, including declination line as the reference line, the top-line, up-stepping, down-stepping, register change and range widening-reducing), lengthening-shortening maneuvers and strengthening-weakening phenomena at the glottal and supraglottic levels, for instantiating acoustically the syllable, the word, the minor and major phrases, and the utterance. Our presentation concerns only attitudinally and emotionally neutral utterances. The second part of the talk will present particular aspects: 1) the different centers of articulatory "effort" at the syllable level; 2) the suggestion of the existence of an unmarked strong-long pattern, neither trochaic nor iambic, at the word level in languages where natives don't have the consciousness of a "lexical stress," or don't agree on its existence or position; 3) the regrouping of one or more words into a prosodic phrase by the application of two established principles: a) the "hat-pattern" principle (t'Hart) favoring initial high-rising and final low-falling F0, and b) the intensive or the temporal rhythmic basic tendencies (Woodrow, Fraisse) favoring a more intense, stronger, more precisely articulated beginning and a lengthened ending; 4) the existence of a multilayer rhythm at the utterance level composed by the repetition/alternation of integrated Gestalts at the levels of the syllable, word, and phrases. One or two Gestalts will prevail perceptually depending on a) the language, b) the style, and c) the rate of speech. The impressionistic evidence of a particular type of language-dependent "rhythm" depends on the listener's expectations, related to his maternal language and the languages he already masters, and up to a certain extent, to his pre-existing theoretical beliefs.

**Biography**
After a thesis in 1970 on French prosody synthesis at the IBM Research Center, La Gaude (France), and the Centre d'études pour la Traduction Automatique, Grenoble, Jacqueline Vaissière joined the Speech Communication Group at MIT, directed by Ken Stevens, as a visiting scientist, where she acquired a specialization in acoustic phonetics. In 1975, she joined the Centre National d'Etudes des Télécommunications (France), where she worked for 15 years on automatic speech recognition and automatic directory services. When the speech processing community moved towards black-box models for recognition and synthesis, she chose to become a professor at the Université Sorbonne Nouvelle in 1990 and became the director of Laboratoire de phonétique

et de Phonologie, associated with the Centre National de la Recherche Scientifique (CNRS). She was the director of 125 masters and 34 doctoral theses, with students of different backgrounds (medical doctors, engineers, and linguistics native of a large variety of languages). She was awarded a CNRS silver medal in 2009. Since 2010, she has been the principal coordinator of the 10-year project « Laboratoire d'Excellence « Empirical Foundations of Linguistics ». She is currently developing methods and applications for acquiring or improving pronunciations, based on visualization of spectrograms, and F0 curves (CleanAccent) and give courses decoding segmental and suprasegmental cues from spectrograms, F0 and intensity curves in different languages. Jacqueline Vaissière is a Fellow of ISCA.

**Hervé Bourlard**
Idiap Research Institute and EPFL, Switzerland
http://people.idiap.ch/bourlard

## Evolution of Neural Network Architectures for Speech Recognition

**Abstract**
Over these last few years, the use of Artificial Neural Networks (ANNs), now often referred to as deep learning or Deep Neural Networks (DNNs), has significantly reshaped research and development in a variety of signal and information processing tasks. While further boosting the state-of-the-art in Automatic Speech Recognition (ASR), recent progresses in the field have also allowed for more flexible and faster developments in emerging markets and multilingual societies (e.g., under-resourced languages).

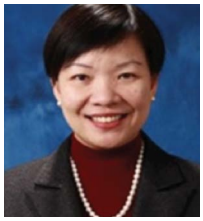
In this talk, we will provide a historical account of ANN architectures used for ASR since the mid-1980's, and now used in most ASR and spoken language understanding applications. We will start by recalling/revisiting key links between ANNs and statistical inference, discriminant analysis, and linear/nonlinear algebra. Finally, we will briefly discuss more recent trends towards novel DNN-based ASR approaches, including complex hierarchical systems, sparse recovery modeling, and "end-to-end systems."

However, and in spite of the recent progress in the area, we still lack basic understanding of the problems in hands. Although more and more tools are now available, in association with basically "unlimited" processing and data resources, we still fail in building principled ASR models and theories. Alternatively, we are still relying on "ignorance-based" models, often exposing limitations of our understanding, rather than enriching the field of ASR. Discussion of these limitations will underpin all of our overview.

**Biography**
Hervé Bourlard is Director of the Idiap Research Institute, Full Professor at the Swiss Federal Institute of Technology Lausanne (EPFL), and Founding Director of the Swiss NSF National Centre of Competence in Research on "Interactive Multimodal Information Management (IM2)" (2001-2013). He is also an External Fellow of the International Computer Science Institute (ICSI), Berkeley, CA.

His research interests mainly include statistical pattern classification, signal processing, multi-channel processing, artificial neural networks, and applied mathematics, with applications to a wide range of Information and Communication Technologies, including spoken language processing, speech and speaker recognition, language modeling, multimodal interaction, and augmented multi-party interaction.

H. Bourlard is the author/co-author/editor of 8 books, and over 330 reviewed papers (including one IEEE paper award). He is a Fellow of IEEE and ISCA, and a Senior Member and Member of the European Council of ACM. He is the recipient of several scientific and entrepreneurship awards.



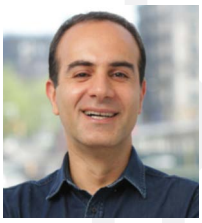
**Helen Meng**
Department of Systems Engineering and Engineering Management
Chinese University of Hong Kong
http://www.se.cuhk.edu.hk/people/hmmeng.html

## Speech and Language Processing for Learning and Wellbeing

**Abstract**
Spoken language is a primary form of human communication. Spoken language processing techniques must incorporate knowledge of acoustics, phonetics and linguistics in analyzing speech. While great strides have been made in the community in general speech recognition, reaching human parity in performance, our team has been focusing on the problems of recognizing and analyzing non-native, learners' speech for the purpose of mispronunciation detection and diagnosis in computer-aided pronunciation training. In order to generate personalized, corrective feedback, we have also developed an approach that uses phonetic posterior-grams (PPGs) for personalized, cross-lingual text-to-speech synthesis given arbitrary textual input, based on voice conversion techniques. We have also extended our work to disordered speech, focusing on automated distinctive feature (DF)-based analyses of dysarthric recordings. The analyses are intended to inform intervention strategies. Additionally, voice conversion is further developed to restore disordered speech to normal speech. This talk will present the challenges in these problems, our approaches and solutions, as well as our ongoing work.

**Biography**
Helen Meng is Patrick Huen Wing Ming Professor and Chairman of the Department of Systems Engineering & Engineering Management, Chinese University of Hong Kong (CUHK). She is the Founding Director of the CUHK Ministry of Education (MoE)-Microsoft Key Laboratory for Human-Centric Computing and Interface Technologies, Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, and CUHK Stanley Ho Big Data Decision Analytics Research Center. She has also established the CAS-CUHK Shenzhen Institute of Advanced Technology Ambient Intelligence and Multimodal Systems Laboratory and served as its Director between 2007 and 2011. Previously, she has served as CUHK Faculty of Engineering's Associate Dean (Research), Editor-in-Chief of the IEEE Transactions on Audio, Speech and Language Processing, ISCA Board Member, Member of the IEEE SPS Board of Governors and Hong Kong-Guangdong ICT Expert Group member. Presently, she is serving as ISCA International Advisory Council Member, elected Chairperson of ISCA's Special Interest Group on Chinese Spoken Language Processing (since 2014) and elected Standing Committee Member of the China Computer Federation Task Force on Speech, Dialogue and Auditory Processing. Her appointments by the Hong Kong SAR Government (HKSARG) include Research Grants Council Member, eHealth Record Sharing Steering Committee Member, and Chairlady of the Working Party for the Manpower Survey of the Innovation & Technology Sector. She was APSIPA's inaugural Distinguished Lecturer 2012-2013 and ISCA Distinguished Lecturer 2015-2016. Her awards include the Ministry of Education Higher Education Outstanding Scientific Research Output Award 2009, Hong Kong Computer Society's inaugural Outstanding ICT Woman Professional Award 2015, Microsoft Research Outstanding Collaborator Award 2016, IEEE ICME 2016 Best Paper Award, IBM Faculty Award 2016, HKPWE Outstanding Women Professionals and Entrepreneurs Award 2017 and Hong Kong ICT Award 2018 Silver Award for Smart Inclusion. Helen received all her degrees from MIT and is a Fellow of HKCS, HKIE, IEEE and ISCA.

# Perspective Talks

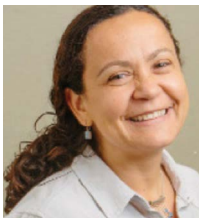


**Nima Mesgarani**
Electrical Engineering Department
Columbia University in the City of New York
http://nima.ee.columbia.edu/

## Speech Processing in the Human Brain Meets Deep Learning

**Abstract**
Speech processing technologies have seen tremendous progress since the advent of deep learning, where the most challenging problems no longer seem out of reach. In parallel, deep learning has advanced the state-of-the-art in processing the neural signals to speech in the human brain. This talk reports progress in three important areas of research: I) Decoding (reconstructing) speech from the human auditory cortex to establish a direct interface with the brain. Such an interface not only can restore communication for paralyzed patients, but also has the potential to transform human-computer interaction technologies, II) Auditory Attention Decoding, which aims to create a mind-controlled hearing aid that can track the brain-waves of a listener to identify and amplify the voice of the attended speaker in a crowd. Such a device could help hearing-impaired listeners communicate more effortlessly with others in noisy environments, and III) More accurate models of the transformations that the brain applies to speech at different stages of the human auditory pathway. This is achieved by training deep neural networks to learn the mapping from sound to the neural responses. Using a novel method to study the exact function learned by these neural networks has led to new insights on how the human brain processes speech. On the other hand, these new insights motivate distinct computational properties that can be incorporated into the neural network models to better capture the properties of speech processing in the human auditory cortex.

**Biography**
Nima Mesgarani is an associate professor at Electrical Engineering Department and Mind-Brain-Behavior Institute of Columbia University in the City of New York. He received his Ph.D. from the University of Maryland and was a postdoctoral scholar in Center for Language and Speech Processing at Johns Hopkins University and the Neurosurgery Department of University of California San Francisco. He has been named a Pew Scholar for Innovative Biomedical Research, and has received several distinctions including the National Science Foundation Early Career Award and Kavli Institute for Brain Science Award. His interdisciplinary research combines theoretical and experimental techniques to model the neural mechanisms involved in human speech communication which critically impacts research in modeling speech processing and speech brain-computer interface technologies.



**Dilek Hakkani-Tűr**
Research Scientist, Amazon, USA
https://scholar.google.com/citations?user=GMcL_9kAAAAJ&hl=en

## Deep Learning Based Situated Goal-oriented Dialogue Systems

**Abstract**
Interacting with machines in natural language has been a holy grail since the beginning of computers. Given the difficulty of understanding natural language, only in the past couple of decades, we started seeing real user applications for targeted/limited domains. More recently, advances in deep learning based approaches enabled exciting new research frontiers for end-to-end goal-oriented conversational systems. In this talk, I'll review end-to-end dialogue systems research, with components for situated language understanding, dialogue state tracking, policy, and language generation. The talk will highlight novel approaches where dialogue is viewed as a collaborative game between a user and an agent in the presence of visual information, and will aim to summarize challenges for future research.

**Biography**
Dilek is a research scientist at Amazon and has previously held research scientist positions at Google, Microsoft Research, ICSI, and AT&T Labs – Research. She is a fellow of the IEEE and of ISCA. Her research interests include conversational AI, natural language and speech processing, spoken dialogue systems, and machine learning for language processing.

**Sriram Ganapathy**
Department of Electrical Engineering
Indian Institute of Science Bangalore, India
http://leap.ee.iisc.ac.in/sriram/

## Speaker and Language Recognition – From Laboratory Technologies to the Wild

**Abstract**
Detecting the paralinguistic components of speech like speaker and language is of substantial interest for many commercial, surveillance and security applications. The problem is at least three decades old with some of the early techniques based on simple Gaussian mixture models. A significant advancement in this area came about a decade ago with the advent of joint factor analysis and i-vector models. The last couple of years have seen further breakthroughs with deep embeddings and end-to-end models based on deep learning. With these improvements in modeling speaker and language, the application of the technology has also moved from clean controlled speech data to telephone channel recordings, far-field microphones and more recently to multi-speaker conversations in the wild. In the talk, I will provide a prospective view of the broad research directions in the field of speaker and language recognition. I will also highlight some of the recent advancements from our work on hierarchical end-to-end approaches with relevance modeling.

**Biography**
Sriram Ganapathy is a faculty member at the Department of Electrical Engineering, Indian Institute of Science Bangalore. Previously, he was a research staff member at the IBM Watson Research Center. He received his PhD from the Center of Language and Speech Processing, Johns Hopkins University. His research interests include signal processing and machine learning applied to speech recognition, speaker recognition and auditory neuroscience. He is a member of the ISCA and a senior member of the IEEE.

**Bhuvana Ramabhadran**
Google Inc., USA
https://scholar.google.com/citations?user=jecE00EAAAAJ&hl=en

## Open Problems in Speech Recognition

**Abstract**
In this talk, I will focus on the evolution of ideas in speech recognition over the last couple of decades, with emphasis on the key breakthroughs over the last ten years, its impact across spoken language processing in several languages, recent trends and open challenges that remain to be addressed. One such breakthrough is the use of several neural network model variants, which has had an enormous impact on the performance of state-of-the-art large vocabulary speech recognition systems. They have also had impact on keyword search which is the task of localizing an orthographic query in a speech corpus, and is typically performed through analysis of automatic speech recognition (ASR). Using the recently concluded IARPA funded Babel program as an example of a well-benchmarked task that focussed on the rapid development of speech recognition capability for keyword search in a previously unstudied language, I will present the successes and challenges that persist with limited amounts of transcription. Interpreting and understanding the hidden representations of various models remains a challenge today. I will also discuss current research taking advantage of such interpretations to improve robustness to noisy environments, speaker/domain adaptation algorithms, and dialects/accents. I will conclude with relevant metrics to measure speech recognition performance today that include and ignore the bigger picture of end-to-end user experience.

**Biography**
Bhuvana Ramabhadran (IEEE Fellow, 2017, ISCA Fellow 2017) currently leads a team of researchers at Google, focussing on multilingual speech recognition and synthesis. Previously, she was a Distinguished Research Staff Member and Manager in IBM Research AI, at the IBM T. J. Watson Research Center, Yorktown Heights, NY, USA, where she led a team of researchers in the Speech Technologies Group and coordinated activities across IBM's worldwide laboratories in the areas of speech recognition, synthesis, and spoken term detection. She was the elected Chair of the IEEE SLTC (2014–2016), Area Chair for ICASSP (2011–2018) and Interspeech (2012–2016), was on the editorial board of the IEEE Transactions on Audio, Speech, and Language Processing (2011–2015), and is currently an ISCA board member. She has served on the editorial board of T-ASLP (2012-2016), technical area chair for ICASSP (2011-2017), Interspeech (2012, 2014-2016), and was one of the lead organizers and technical chair of IEEE ASRU 2011, She has given tutorial and keynote presentations at several international conferences and served as an adjunct professor in Columbia University, where she co-taught a course in speech recognition. She has published over 150 papers and been granted over 40 U.S. patents. She was named a Master Inventor twice by IBM. She is a reviewer for ICASSP, Inerspeech, NAACL, ACL, EMNLP and serves on student dissertation committees and NSF review panels. Her research interests include speech recognition and synthesis algorithms, statistical modeling, signal processing, and machine learning.

# Special Sessions

## 1. INTERSPEECH 2018 Computational Paralinguistics ChallengE (ComParE): Atypical and Self-Assessed Affect, Crying & Heart Beats

September 3, 2018, 14:00-16:00 and 16:30-18:30, MR 1.01-1.02, HICC, Hyderabad

**Organizers**
Björn Schuller (bjoern.schuller@imperial.ac.uk)
Stefan Steidl (stefan.steidl@fau.de)
Anton Batliner (batliner@cs.fau.de)
Peter Marschik (peter.marschik@medunigraz.at)
Harald Baumeister (harald.baumeister@uni-ulm.de)
Fengquan Dong (fengquan.dong@foxmail.com)

The Interspeech 2018 Computational Paralinguistics ChallengE (ComParE) is an open Challenge dealing with states and traits of speakers as manifested in their speech signal's acoustic properties. There have so far been nine consecutive Challenges at INTERSPEECH since 2009 (cf. the Challenge series' repository at http://www.compare.openaudio.eu), but there still exists a multiplicity of not yet covered, but highly relevant paralinguistic phenomena. Thus, in this year's 10th anniversary edition, we introduce four new tasks. The following Sub-Challenges are addressed:

- In the Atypical Affect Sub-Challenge, emotion of disabled speakers is to be recognised.
- In the Self-Assessed Affect Sub-Challenge, self-assessed affect shall be determined.
- In the Crying Sub-Challenge, mood-related types of infant vocalisation have to be classified.
- In the Heart Beats Sub-Challenge, types of Heart Beat Sounds need to be distinguished.

All Sub-Challenges allow contributors to find their own features with their own machine learning algorithm. However, a standard feature set and tools will be provided that may be used. Participants will have to stick to the definition of training, development, and test sets as given. They may report results obtained on the development sets, but have only five trials to upload their results on the test set per Sub-Challenge, whose labels are unknown to them. Each participation has to be accompanied by a paper presenting the results that undergoes the normal Interspeech peer-review and has to be accepted for the conference in order to participate in the Challenge. The organisers preserve the right to re-evaluate the findings, but will not participate themselves in the Challenge.

In these respects, the INTERSPEECH 2018 Computational Paralinguistics challengE (ComParE) shall help bridging the gap between excellent research on paralinguistic information in spoken language and low compatibility of results. We encourage both – contributions aiming at highest performance w.r.t. the baselines provided by the organisers, and contributions aiming at finding new and interesting insights w.r.t. these data. Overall, contributions using the provided or equivalent data are sought for (but not limited to):

• Participation in a Sub-Challenge
• Contributions focussing on Computational Paralinguistics centred around the Challenge topics
The results of the Challenge will be presented at Interspeech 2018 in Hyderabad, India. Prizes will be awarded to the Sub-Challenge winners. If you are interested and planning to participate in INTERSPEECH 2018 ComParE, or if you want to be kept informed about the Challenge, please send the organisers an e-mail (steidl@xxxxxxxxx) to indicate your interest and visit the homepage: http://emotion-research.net/sigs/speech-sig/is18-compare

## 2. The First DIHARD Speech Diarization Challenge

September 5, 2018, 17:00-19:00, MR 1.01-1.02, HICC, Hyderabad

**Organizers**

Kenneth Church (kenneth.ward.church@gmail.com)
Christopher Cieri (ccieri@ldc.upenn.edu)
Alejandrina Cristia (alecristia@gmail.com)
Jun Du (jundu@ustc.edu.cn)
Sriram Ganapathy (sriram.iisc@gmail.com)
Mark Liberman (markyliberman@gmail.com)
Neville Ryant (nryant@ldc.upenn.edu)

DIHARD is a new annual challenge focusing on "hard" diarization; that is, speech diarization for challenging corpora where there is an expectation that the current state-of-the-art will fare poorly, including, but not limited to:
• clinical interviews
• extended child language acquisition recordings
• web videos
• "speech in the wild" (e.g., recordings in restaurants)

Because performance of diarization is highly dependent on the quality of the speech activity detection (SAD) system used, the challenge will have two tracks:

Track 1: diarization beginning from gold speech segmentation
Track 2: diarization from scratch

For more details, please see the challenge site at https://coml.lscp.ens.fr/dihard/index.html

## 3. Novel Paradigms for Direct Synthesis based on Speech-Related Biosignals

September 6, 2018, 10:00-12:00, MR G.03-G.04, HICC, Hyderabad

**Organizers:**

Lorenz Diener (lorenz.diener@uni-bremen.de)
Jose Gonzalez (jgonzalez@lcc.uma.es)
Tanja Schultz (tanja.schultz@uni-bremen.de)

Speech is a very rich and complex process, the acoustic signal being just one of the biosignals resulting from it. In the last few years, the automatic processing of these speech-related biosignals

has become an active area of research within the speech community. This special session aims to foster research on one emerging area that is growing within the field of silent speech: Direct synthesis. Direct synthesis refers to the generation of speech directly from speech-related biosignals (e.g. ultrasound, EMG, EMA, PMA, lip reading video, BCI) without an intermediate recognition step. This has been made possible by recent developments in supervised machine learning techniques and the availability of high-resolution biosensors. Furthermore, the availability of low-cost computing devices has made something possible that was unthinkable 20 years ago: the generation of audible speech from speech-related biosignals in real time. With this special session, we aim at bringing together researchers working on direct synthesis and related topics to foster work towards direct synthesis toolkits and datasets and to highlight and discuss common challenges and solutions in this emerging research area.

Web-site: http://www.uni-bremen.de/csl/forschung/lautlose-sprachkommunikation/special-session-novel-paradigms-for-direct-synthesis-based-on-speech-related-biosignals.html

## 4. Speech Recognition for Indian Languages

September 4, 2018, 10:00-12:00, MR 1.01-1.02, HICC, Hyderabad

**Organizers:**

Pedro Moreno (pedro@google.com)
Eugene Weinstein (weinstein@google.com)
Sarah Abu Sharkh (sarah@google.com)
Haruko Ishikawa (ishikawa@google.com)
Daan van Esch (dvanesch@google.com)
Rita Singh (rsingh@cs.cmu.edu)
Preethi Jyothi (pjyothi@cse.iitb.ac.in)

Indian languages offer a multitude of challenges that are not observed as widely in languages elsewhere, e.g. code-switching ("Hinglish"). However, research on ASR for Indian languages remains relatively scarce compared to English, European languages, and East-Asian languages. We will invite participants of this Special Session to build their own ASR systems for Indian languages, in the broad sense of the term. You may choose which Indian languages to build systems for, and what data sets to use. In addition to the challenge around building ASR systems, we will also allocate time in the schedule for a show-and-tell: some participants may prefer using an existing ASR system and then building a voice-enabled app. We believe that this Special Session would provide an important boost to academic research on building ASR for Indian languages

Web-site: https://sites.google.com/view/interspeech2018-ss1

## 5. Deep Neural Networks: How Can We Interpret What They Learned?

September 4, 2018, 14:30-16:30, MR 1.01-1.02, HICC, Hyderabad

**Organizers:**

Louis ten Bosch (l.tenbosch@let.ru.nl)
Hugo Van hamme (hugo.vanhamme@esat.kuleuven.be)
Lou Boves (l.boves@ru.nl)

Everybody active in the speech technology area witnesses the advent of deep learning techniques, in particular the use of Deep Neural Nets. DNNs are now being applied in many types of automatic speech recognition systems. The success of DNNs strongly suggests that the parameter set of a trained DNN reflects relevant structure in the training data, but it is not clear how one can reveal this structure. In this session we want to address this issue.

We welcome papers that specifically address one or more of the leading questions listed below, e.g. by aiming at (phonetic or linguistic) interpretations of the representations at the layers of DNNs, or by attempting to use phonetic/linguistic knowledge to guide the design and training of DNNs.

For papers in this special session, leading questions are:
[1] How does a trained DNN encapsulate the structure that exists in a data set ?
[2] How can we visualize this information ?
[3] How can we learn from a DNN, i.e., how can the information in a DNN be used to sharpen our insights ?
[4] What type of knowledge can be encoded in a DNN ?
[5] Can understanding the information encoded in a DNN be used as a guidance in designing and training more powerful networks ?
[6] What architectures and training techniques are most amenable to interpretations?

For more information please contact the organizers. http://cls.ru.nl/interspeech2018

## 6. Low Resource Speech Recognition Challenge For Indian Languages

September 6, 2018, 10:00-12:00, MR 1.01-1.02, HICC, Hyderabad

**Organizers:**
Kalika Bali (kalikab@microsoft.com)
Krishna Doss Mohan (krishna.doss@microsoft.com)
Rupesh Kumar Mehta (rupesh.mehta@microsoft.com)
Niranjan Nayak (niranjan@microsoft.com)
Sunayana Sitaram (t-susita@microsoft.com)
Radhakrishnan Srikanth (rsrikan@microsoft.com)

Most languages in the world lack the amount of text, speech and linguistic resources required to build large Deep Neural Network (DNN) based models. However, there have been many advances in DNN architectures, cross-lingual and multilingual speech processing techniques, and approaches incorporating linguistic knowledge into machine-learning based models, that can help in building systems for low resource languages. In this challenge, we would like to focus on building Automatic Speech Recognition (ASR) systems for Indian languages with constraints on the data available for Acoustic Modeling and Language Modeling. For more details on the challenge and registration please visit the web site.

Web-site: https://www.microsoft.com/en-us/research/event/interspeech-2018-special-session-low-resource-speech-recognition-challenge-indian-languages/

## 7. Integrating Speech Science and Technology for Clinical Applications

September 4, 2018, 14:30-16.30 Hall 4-6: PosterArea 4, HICC, Hyderabad
September 5, 2018, 10:00-12:00, MR G.03-G.04, HICC, Hyderabad

**Organizers:**
Christina Hagedorn (christina.hagedorn@csi.cuny.edu)
Shrikanth Narayanan (shri@ee.usc.edu)
Uttam Sinha (sinha@med.usc.edu)

The broad objectives of this session are to (i) address the current communication and collaboration gaps that exist between the fields of speech science, engineering and technological development, and communication disorders, and (ii) serve as a bridge to unify members of these distinct fields through interactive and dynamic exchanging of experimental findings and ideas for future collaboration. The organizers encourage submissions focused on, though not limited to: characterizing disordered speech using novel imaging techniques and analytical methods, (semi-) automatic detection of speech disorder characteristics, and the efficacy of biofeedback intervention for speech disorders.

Web-site: http://www.speechlab.csi.cuny.edu/ssinterspeech2018.html

## 8. Speech Technologies for Code-Switching in Multilingual Communities

September 5, 2018, 10:00-12:00, MR 1.01-1.02, HICC, Hyderabad

**Organizers:**
Kalika Bali (kalikab@microsoft.com)
Alan Black (awb@cs.cmu.edu)
Mona Diab (mtdiab@gwu.edu)
Julia Hirschberg (julia@cs.columbia.edu)
Sunayana Sitaram (t-susita@microsoft.com)
Thamar Solorio (solorio@cs.uh.edu)

Speech technologies exist for many high resource languages, and attempts are being made to reach the next billion users by building resources and systems for many more languages. Multilingual communities pose many challenges for the design and development of speech processing systems. One of these challenges is code-switching , which is the switching of two or more languages at the conversation, utterance and sometimes even word level.

Code-switching is now found in text in social media, instant messaging and blogs in multilingual communities in addition to conversational speech. Monolingual natural language and speech systems fail when they encounter code-switched speech and text. There is a lack of linguistic data and resources for code-switched speech and text, although one or more of the languages being mixed could be high-resource. Code-switching provides various interesting challenges to the speech community, such as language modeling for mixed languages, acoustic modeling of mixed language speech, pronunciation modeling and language identification from speech.

We conducted the inaugural special session on code-switching at Interspeech 2017, which was organized as a double session spanning four hours. We received several high-quality submissions from research groups all over the world, out of which nine papers were selected

as oral presentations. At the end of the oral presentations, we conducted a panel discussion between researchers in academia and industry about challenges in research, building systems and collecting code-switched data. Our special session was attended by several researchers from academia and industry working on linguistics, NLP and speech technologies.

Web-site: https://www.microsoft.com/en-us/research/event/interspeech-2018-special-session-speech-techologies-code-switching-multilingual-communities/

## 9. Spoken CALL Shared Task, Second Edition

September 5, 2018, 14:30-16:30, MR 1.01-1.02, HICC, Hyderabad

**Organizers:**
Johanna Gerlach (Johanna.Gerlach@unige.ch)
Manny Rayner (Emmanuel.Rayner@unige.ch)
Martin Russell (m.j.russell@bham.ac.uk)
Helmer Strik (w.strik@let.ru.nl)

The Spoken CALL Shared Task is an initiative to create an open challenge dataset for speech-enabled CALL systems, jointly organised by the University of Geneva, the University of Birmingham, Radboud University and Cambridge University. The task is based on data collected from a speech-enabled online tool which has been used to help young Swiss German teens practise skills in English conversation. Items are prompt-response pairs, where the prompt is a piece of German text and the response is a recorded English audio file. The task is to label pairs as "accept" or "reject", accepting responses which are grammatically and linguistically correct to match a set of hidden gold standard answers as closely as possible. Resources are provided so that a scratch system can be constructed with a minimal investment of effort, and in particular without necessarily using a speech recogniser.

The first edition of the task was announced at LREC 2016, with training data released in July 2016 and test data in March 2017, and attracted 20 entries from 9 groups. Results, including seven papers, were presented at the SLaTE workshop in August 2017. Full details, including links to resources, results and papers, can be found on the Shared Task home page.

Following the success of the original task, we are organising a second edition. We have approximately doubled the amount of training data, will provide new test data, and have released improved versions of the accompanying resources. In particular, we have made generally available the open source Kaldi recogniser developed by the University of Birmingham, which achieved the best performance on the original task, together with versions of the training and test data pre-processed through this recogniser.

Web-site: https://regulus.unige.ch/spokencallsharedtask_2ndedition/

# Show & Tell

Session S&T-1: September 3, 2018, 14:00-16:00, MR G.05-G.06

Session S&T-2: September 3, 2018, 16:30-18:30, MR G.05-G.06

Session S&T-3: September 4, 2018, 10:00-12:00, MR G.05-G.06

Session S&T-4: September 4, 2018, 14:30-16:30, MR G.05-G.06

Session S&T-5: September 5, 2018, 10:00-12:00, MR G.05-G.06

Session S&T-6: September 5, 2018, 14:30-16:30, MR G.05-G.06

Session S&T-7: September 6, 2018, 10:00-12:00, MR G.05-G.06

Show & Tell is a special event organized during the conference. Participants are given the opportunity to demonstrate their most recent progress or developments, and interact with the conference attendees in an informal way, such as poster, a mock-up, a demo, or any adapted format of their own choice. These contributions must highlight the innovative side of the concept and may relate to a regular paper. While the emphasis of Show & Tell is on the actual demonstration during the conference, all contributions are allocated two pages in the conference proceedings. Each submission has been peer-reviewed. Reviewers have judged the originality, significance, quality and clarity of the proposed demonstration.

# Special Events

## 1. Workshop for Young Female Researchers in Speech Science & Technology 2018

September 1, 2018
09:00 to 17:00
Venue: Fourth Floor Auditorium, Kohli Research Block,
International Institute of Information Technology (IIIT), Gachibowli, Hyderabad
url: https://www.iiit.ac.in/

**Organizers:**
Amber Afshan (UCLA)
Kay Berkling (Karlsruhe University)
Heidi Christensen (University of Sheffield)
Maxine Eskenazi (CMU)
Milica Gasic (Cambridge University)
Dilek Hakkani-Tür (Google Inc)
Preethi Jyothi (IIT Bombay)
Esther Klabbers (ReadSpeaker)
Lori Lamel (LIMSI CNRS)
Yang Liu (University of Texas Dallas)
Karen Livescu (Toyota Technological Institute at Chicago)
Pratibha Moogi (Samsung Electronics)
Emily Mower Provost (University of Michigan)
Catharine Oertel (EPFL)
Bhuvana Ramabhadran (Google Inc)
Odette Scharenborg (M*Modal)
Elizabeth Shriberg (Ellipsis Health Inc)
Isabel Trancoso (INESC-ID / Instituto Superior Técnico)

A workshop for women undergraduate and masters students who are currently working in speech science and technology is to be held at Interspeech 2018 on September 1, 2018, from 9 am to 5 pm followed by a dinner with invited senior members of the Interspeech community. It will feature panel discussions with senior female researchers in the field, student poster presentations and a mentoring session.

The workshop is the third of its kind, after a successful inaugural event YFRSW 2016, at Interspeech 2016 in San Francisco and YFRSW 2017 at Interspeech 2017 in Stockholm. It is designed to foster interest in research in our field in women at the undergraduate or master level who have not yet committed to getting a PhD in speech science or technology areas, but who have had some research experience in their college and universities via individual or group projects.

Please direct any questions to: yfrs.workshop@gmail.com

## 2. Women in Science (WiS) 2018 – Challenges & Opportunities

September 2, 2018
Time: 17:30-18:30
Venue: MR 1.05-1.06, HICC Hyderabad

**Organizing committee**

- Sharmistha Gray (Nuance)
- Preethi Jyothi (IIT Bombay)
- Pratibha Moogi (Industry)
- Hema Murthy (IIT, Madras)
- Emily Mower Provost (University of Michigan)
- Catharine Oertel (EPFL)
- Bhuvana Ramabhadran (Google Inc)
- Odette Scharenborg (Delft University of Technology)
- Isabel Trancoso (INESC-ID / Instituto Superior Técnico)

A panel discussion by invited members of men and women in speech science and technology with a focus on issues and hurdles women face in academic setup and work places, is to be held at Interspeech 2018 on September 2, 2018, from 5:30 pm to 6:30 pm, accompanied by a light snack.

The panel discussions will touch upon some of these crucial points such as challenges faced by women scientists, how can encouraging work and research environment be built around women who want to pursue their career in research and technology development. Panelist members include experienced researchers, executive members from Industry, will share their rich experiences on current challenges & opportunities that exist currently at work place.

Please direct any questions to: yfrs.workshop@gmail.com

# ISCA Meetings

**1. ISCA Board 1**
(Participation by invitation only)
Chair: John H.L. Hansen
Date: September 2, 2018 (Sunday)
Time: 09:00 - 16:00
Venue: MR 2.06, HICC, Hyderabad

**2. Student Paper Meeting**
(Participation by invitation only)
Chair: Torbjørn Svendsen
Date: September 2, 2018 (Sunday)
Time: 16:00 - 20:00
Venue: MR 2.06, HICC, Hyderabad

**3. IS 2019 Preparation Meeting**
(Graz Team only)
Chair: Gernot Kubin
Date: September 3, 2018 (Monday)
Time: 12:00 - 13:30
Venue: MR 2.05, HICC, Hyderabad

**4. IAC Lunch Meeting (Advisory Committee)**
(Participation by invitation only)
Chair: John H.L. Hansen
Date: September 3, 2018 (Monday)
Time: 12:15 - 13:45
Venue: MR 2.06, HICC, Hyderabad

**5. ISCA Technical Committee Breakfast**
(Participation by invitation only)
Chair: Lori Lamel
Date: September 4, 2018 (Tuesday)
Time: 07:30 - 08:30
Venue: MR 2.06, HICC, Hyderabad

**6. Interspeech Steering Committee Lunch**
(Participation by invitation only)
Chair: Sebastian Möller, Odette Scharenborg and Lori Lamel
Date: September 4, 2018 (Tuesday)
Time: 13:00 - 14:00
Venue: MR 2.05, HICC, Hyderabad

**7. ISCA SIG Topic Lunch**
(Participation by invitation only)
Chair: Gérard Bailly
Date: September 4, 2018 (Tuesday)
Time: 13:15 - 14:15
Venue: MR 2.06, HICC, Hyderabad

**8. Industry Lunch**
(Participation by invitation only)
Chair: Bhuvana Ramabhadran
Date: September 4, 2018 (Tuesday)
Time: 13:30 - 14:30
Venue: MR 1.06, HICC, Hyderabad

**9. Joint Int'l & Language SIG Breakfast**
(Participation by invitation only)
Chair: Satoshi Nakamura
Date: September 5, 2018 (Wednesday)
Time: 07:30 - 08:30
Venue: MR 2.06, HICC, Hyderabad

**10. ISCA - IS 2020, China Preparation Meeting**
(Participation by invitation only)
Chair: Sebastian Möller and Odette Scharenborg
Date: September 5, 2018 (Wednesday)
Time: 07:30 - 08:30
Venue: MR 2.05, HICC, Hyderabad

**11. ISCA - IS 2019, Graz Preparation Meeting**
(Participation by invitation only)
Chair: Sebastian Möller and Odette Scharenborg
Date: September 5, 2018 (Wednesday)
Time: 13:00 - 14:00
Venue: MR 2.06, HICC, Hyderabad

**12. ISCA Fellows Reception**
(Participation by invitation only)
Chair: Phil Green
Date: September 5, 2018 (Wednesday)
Time: 18:00 - 18:45
Venue: MR 2.05, HICC, Hyderabad

**13. SAC Lunch Meeting (students)**
(Participation by invitation only)
Chair: Lori Lamel
Date: September 6, 2018 (Thursday)
Time: 13:30 - 14:30
Venue: MR 2.05, HICC, Hyderabad

**14. CSL Lunch (Elsevier)**
(Participation by invitation only)
Chair: Emmanuelle FOXONET
Date: September 6, 2018 (Thursday)
Time: 13:00 - 14:00
Venue: MR 2.06, HICC, Hyderabad

**15. ISCA Board 2**
(Participation by invitation only)
Chair: John H.L. Hansen
Date: September 7, 2018 (Friday)
Time: 08:30 - 13:00
Venue: MR 2.06, HICC, Hyderabad

# ISCA-SAC Events

## 1. Fourth Doctoral Consortium at INTERSPEECH 2018

Date: September 1, 2018; 10.00 - 17.00
Location: International Institute of Information Technology (IIIT) Hyderabad

The Student Advisory Committee of the International Speech Communication Association (ISCA-SAC) is pleased to call for applications for the 4th Doctoral Consortium. This event extends an opportunity for doctoral candidates to present and discuss their research with a panel of experts. The discussion would include a feedback on the evolution and progress of their research. It also helps them to identify the roadmap and additional studies, which could help refine the shape of their thesis.

The doctoral consortium will be a one-day event (including lunch and coffee breaks) being organised on Saturday, September 1, 2018, at the International Institute of Information Technology (IIIT) Hyderabad, Gachibowli. The applicants are required to submit a two-page extended abstract of their PhD research work. The shortlisted candidates would be invited to the consortium where they are required to present a summary of their research. Each candidate will be given 30 minutes for the presentation, which will be followed by a discussion of 15 minutes, led by a panel of experts.

**Panel of Experts**
- Mathew Magimai Doss (IDIAP Research Institute)
- Isabel Trancoso (INESC ID Lisbon)
- Vikramjit Mitra (SRI International)
- Bhuvana Ramabhadran (Google Inc., NY, USA)
- Richard Stern (CMU, Pittsburgh)
- Prasanta Kumar Ghosh (IISc, Bangalore)
- Chandra Sekhar Seelamantula (IISc, Bangalore)
- Mariapaola D'Imperio (Aix Marseille University, Italy)

**Contact**
Ravi Shankar Prasad; Email: ravi.rythem@gmail.com

## 2. INTERSPEECH 2018: Open Doors

Date: September 3, 2018; 14.00 - 16.00
Location: Genesys and Microsoft India Development Center

International Speech Communication Association (ISCA) - Student Advisory Committee (SAC) is back with the Open Doors event. This year we are hosted by Genesys and Microsoft India Development Center. The Open Doors event is focused to bring industrial research and academia together.

Participants will have the opportunity to visit one of the companies and engage with their speech application team. The industry professionals will demonstrate their products and related technology, followed by an open discussion, and networking opportunities. The event aims to bring students and researchers together discussing state-of-the-art technologies, potential collaboration, and even possible hiring opportunities.

ISCA-SAC will provide transportation from the conference center. Group sizes will be limited.

**Contact**
Berrak Sisman; Email: berraksisman@u.nus.edu

## 3. INTERSPEECH 2018: Students Meet Experts

Date: September 5, 2018; 14.00 - 16.00
Location: Hyderabad International Convention Centre

The Student Advisory Committee of the International Speech Communication Association (ISCA-SAC) is happy to announce this year's edition of the Students meet Experts event. After successful editions in Lyon (2013), Singapore (2014), San Francisco (2016), and Stockholm (2017), the Students meet Experts event is now coming to Interspeech 2018 in Hyderabad. We will have a panel discussion with experts from academia and industry, followed by an informal opportunity to mingle and ask more in depth questions at the SME coffee break.

**Panel of Experts**
- Sharmistha Gray (Nuance Research)
- Dilek Hakkani-Tür (Google Research)
- John H. L. Hansen (The University of Texas at Dallas)
- Haizhou Li (National University of Singapore)
- Satoshi Nakamura (Nara Institute of Science and Technology)
- Shrikanth Narayanan (University of Southern California)
- Vikram Ramanarayanan (ETS Research)
- Odette Scharenborg (Delft University of Technology)

**Contact**
Iona Gessinger; Email: gessinger@coli.uni-saarland.de

# Awards

## ISCA Medal for Scientific Achievement

The ISCA Medal for Scientific Achievement 2018 will be awarded to Professor Bishnu S. Atal by the President of ISCA during the opening ceremony.

## ISCA Best Student Paper Award Finalists

**2572 Automatic Glottis Localization and Segmentation in Stroboscopic Videos Using Deep Neural Network**
Wed-P-3-4-7: Source and Supra-segmentals
September 5, 17:00, Hall 4-6, PosterArea 4
**Achuth Rao M V**, Rahul Krishnamurthy, Pebbili Gopikishore, Veeramani Priyadharshini and Prasanta Kumar Ghosh

**0063 Effects of User Controlled Speech Rate on Intelligibility in Noisy Environments**
Wed-O-1-3-6: Production of Prosody
September 5, 11:40, Hall 2
**John S. Novak**, III, Robert V. Kenyon

**1288 An Interlocutor-Modulated Attentional LSTM for Differentiating between Subgroups of Autism Spectrum Disorder**
Wed-O-2-5-2: Speech and Language Analytics for Mental Health
September 5, 14:50, MR G.03-G.04
**Yun-Shao Lin**, Susan Shur-Fen Ga, Chi-Chun Lee

**1515 An Improved Deep Embedding Learning Method for Short Duration Speaker Verification**
Thu-P-2-1-4: Speaker Verification Using Neural Network Methods II
September 6, 14:30, Hall 4-6, PosterArea 1
**Zhifu Gao**, Yan Song, Ian McLoughlin, Wu Guo, Lirong Dai

**1269 Joint Localization and Classification of Multiple Sound Sources Using a Multi-task Neural Network**
Mon-P-1-3-3: Deep Learning for Source Separation and Pitch Tracking
September 3, 14:00, Hall 4-6, PosterArea 3
**Weipeng He**, Petr Motlicek and Jean-Marc Odobez

**2423 A Priori SNR Estimation Based on a Recurrent Neural Network for Robust Speech Enhancement**
Thu-P-1-1-12: Deep Enhancement
September 6, 10:00, Hall 4-6, PosterArea 1
**Yangyang Xia** and Richard M. Stern

**1830 Multi-target Voice Conversion without Parallel Data by Adversarially Learning Disentangled Audio Representations**
Mon-O-2-4-6: Voice Conversion
September 3, 18:10, MR G.03 - G.04
**Ju-chieh Chou**, **Cheng-chieh Yeh**, Hung-yi Lee and Lin-shan Lee

**2456 Multi-Modal Data Augmentation for End-to-end ASR**
Wed-P-2-1-1: Adjusting to Speaker, Accent, and Domain
September 5, 14:30, Hall 4
**Adithya Renduchintala**, **Shuoyang Ding**, **Matthew Wiesner** and Shinji Watanabe

**1339 A GPU-based WFST Decoder with Exact Lattice Generation**
Wed-O-2-1-2: Recurrent Neural Models for ASR
September 5, 14:50, Hall 3
**Zhehuai Chen**, Justin Luitjens, **Hainan Xu**, **Yiming Wang**, Daniel Povey and Sanjeev Khudanpur

**1352 User-centric Evaluation of Automatic Punctuation in ASR Closed Captioning**
Wed-P-2-4-7: Topics in Speech Recognition
September 5, 14:30, Hall 4-6, Poster Area 4
**Ákos Máté Tündik,** György Szaszák, Gábor Gosztolya and András Bek

**2522 Detecting Depression with Audio/Text Sequence Modeling of Interviews**
Tue-SS-2-2-10: Integrating Speech Science and Technology for Clinical Applications
September 4, 14:30, Hall 4-6, Poster Area 4
**Tuka Alhanai**, **Mohammad Ghassemi** and James Glass

**1346 Joint Learning of Interactive Spoken Content Retrieval and Trainable User Simulator**
Wed-P-1-2-1: Extracting Information from Audio
September 5, 10:00, Hall 4-6, Poster Area 2
**Pei-Hung Chung**, **Kuan Tung**, **Ching-Lun Tai** and Hung-Yi Lee

# Travel Grants

Amber Afshan
Haque Albert
Guozhen An
Rohit Ananthanarayana
Vikram C M
Valliappan CA
Li Chai
Oscar Chen
Gaofeng Cheng
Ju-Chieh Chou
Pei Hung Chung
Nikolaos Flemotomos
Tom Francis
Szu-Wei Fu
Zhifu Gao
Saurabh Garg
Deepanway Ghosal
Debayan Ghosh
Yuan Gong
Jing Han
Nils Holzenberger
Kusha Sridhar Huliyar Sridhara Murthy
Marc Antony Hullebus
Ahmed Imtiaz Humayun
Yusuke Inoue
Abhinav Jain
Jesin James
Sebastian Jilt
Bhavya Karki
Matthew Kelley
Manoj Kumar
Yun-Shao Lin
Karttikeya Mangalam
Ateeq Mohammad
Verkholyak Oxana
Laxmi Pandey
Tanmay Parekh
Yu-Huai Peng
Mengjie Qian
Achuth Rao M V
Rachid Riad
Hardik Sailor
Lena Reed
Latif Siddique
Berrak Sisman
Zhihua Su
Parth Suresh
Gajan Suthokumar
Shuai Wang
Ke Wang
Tifani Warnita
Li Wenjie
Huiyi Wu
Yangyang Xia
Yijia Xu
Özkanca Yasin
Zheng Yibin
Jenny Yu
Yixin Zhang
Haris Bin Zia

In addition, Jinyu Li (Microsoft) and Florian Metze (CMU) set up the "Yajie Miao Memorial Student Travel Fund" to remember and to honor the work of Yajie Miao.

Yajie successfully defended his thesis on "Incorporating Context Information into Deep Neural Network Acoustic Models" at Carnegie Mellon University, and was awarded the PhD degree in August 2016. He had accepted a position at Microsoft in Redmond, and was set to start work there in October 2016. Unfortunately, he died tragically, while visiting his family in China, before he was able to do so.

The "Yajie Miao Memorial Student Travel Fund" supports additional ISCA student travel grants to speech conferences. Depending on the availability of funds, one or more recipients will be selected by ISCA and the organizers. More information about the fund and the background can be found at https://www.youcaring.com/iscainternationalspeechcommunicationassociation815026.

This year, the recipients are:

Kayokwa Nick Chibuye
Trinh Viet Anh
Thomas Zenkel

# Session Index

## September 3, 2018, Monday


Mon-Medal-1 ISCA Medal Talk....104
Mon-O-1-1 End-to-End Speech Recognition....104
Mon-O-1-2 Prosody Modeling and Generation....105
Mon-O-1-3 Speaker Verification I....106
Mon-O-1-4 Spoken Term Detection....107
Mon-SS-1-1 Special Session: The INTERSPEECH 2018 Computational Paralinguistics ChallengE (ComParE): Atypical & Self-Assessed Affect, Crying & Heart Beats 1....109
Mon-S&T-1-1 Show and Tell 1....111
Mon-P-1-1 Speech Segments and Voice Quality....112
Mon-P-1-2 Speaker State and Trait....115
Mon-P-1-3 Deep Learning for Source Separation and Pitch Tracking....117
Mon-P-1-4 Acoustic Analysis-Synthesis of Speech Disorders....119
Mon-O-2-1 ASR Systems and Technologies....120
Mon-O-2-2 Deception, Personality, and Culture Attribute....121
Mon-O-2-3 Automatic Detection and Recognition of Voice and Speech Disorders....122
Mon-O-2-4 Voice Conversion....124
Mon-SS-2-1 Special Session: The INTERSPEECH 2018 Computational Paralinguistics ChallengE (ComParE): Atypical & Self-Assessed Affect, Crying & Heart Beats 2....125
Mon-S&T-2-1 Show and Tell 2....127
Mon-P-2-1 Spoken Dialogue Systems and Conversational Analysis....128
Mon-P-2-2 Spoofing Detection....130
Mon-P-2-3 Speech Analysis and Representation....133
Mon-P-2-4 Sequence Models for ASR....136
Mon-P-2-5 Source Separation and Spatial Analysis....138


## September 4, 2018, Tuesday


Tue-Plenary-1 Plenary Talk-1....140
Tue-O-1-1 Acoustic Model Adaptation....140
Tue-O-1-2 Statistical Parametric Speech Synthesis....141
Tue-O-1-3 Emotion Modeling....142
Tue-O-1-4 Models of Speech Perception....143
Tue-O-1-5 Multimodal Dialogue Systems....145
Tue-SS-1-1 Special Session: Speech Recognition for Indian Languages....146
Tue-S&T-1-1 Show and Tell 3....148
Tue-P-1-1 Speaker Verification II....148
Tue-P-1-2 Novel Approaches to Enhancement....151
Tue-P-1-3 Syllabification, Rhythm, and Voice Activity Detection....153
Tue-P-1-4 Selected Topics in Neural Speech Processing....156
Tue-Perspective-1 Perspective Talk-1....158
Tue-IP-1 Industry Presentation-1....158
Tue-IP-2 Industry Presentation-2....158

Tue-IP-3 Industry Presentation-3....158
Tue-O-2-1 Dereverberation....158
Tue-O-2-2 Audio Events and Acoustic Scenes....159
Tue-O-2-3 Speaker Diarization....160
Tue-O-2-4 Phonation....161
Tue-O-2-5 Cognition and Brain Studies....163
Tue-SS-2-1 Special Session: Deep Neural Networks: How Can We Interpret What They Learned?....164
Tue-S&T-2-1 Show and Tell 4....165
Tue-P-2-1 Speech and Singing Production....166
Tue-P-2-2 Robust Speech Recognition....168
Tue-P-2-3 Applications in Education and Learning....171
Tue-SS-2-2 Special Session: Integrating Speech Science and Technology for Clinical Applications....173
Tue-P-2-5 Speaker Characterization and Analysis....175
Tue-Perspective-2 Perspective Talk-2....177

## September 5, 2018, Wednesday

Wed-Plenary-2 Plenary Talk-2....177
Wed-O-1-1 Novel Neural Network Architectures for Acoustic Modelling....177
Wed-O-1-2 Language Identification....178
Wed-O-1-3 Production of Prosody....180
Wed-O-1-4 Speech Intelligibility and Quality....181
Wed-SS-1-1 Special Session: Integrating Speech Science and Technology for Clinical Applications....182
Wed-SS-1-2 Special Session: Speech Technologies for Code-Switching in Multilingual Communities....183
Wed-S&T-1-1 Show and Tell 5....185
Wed-P-1-1 Voice Conversion and Speech Synthesis....186
Wed-P-1-2 Extracting Information from Audio....188
Wed-P-1-3 Signal Analysis for the Natural, Biological and Social Sciences....191
Wed-P-1-4 Speech Prosody....193
Wed-Perspective-3 Perspective Talk-3....195
Wed-IP-4 Industry Presentation-4....195
Wed-IP-5 Industry Presentation-5....195
Wed-IP-6 Industry Presentation-6....195
Wed-O-2-1 Recurrent Neural Models for ASR....195
Wed-O-2-2 Speaker Verification Using Neural Network Methods I....197
Wed-O-2-3 Speech Perception in Adverse Conditions....198
Wed-O-2-4 Measuring Pitch and Articulation....199
Wed-O-2-5 Speech and Language Analytics for Mental Health....200
Wed-SS-2-1 Special Session: Spoken CALL Shared Task, Second Edition....201
Wed-S&T-2-1 Show and Tell 6....203
Wed-P-2-1 Adjusting to Speaker, Accent, and Domain....204
Wed-P-2-2 Speech Synthesis Paradigms and Methods....206
Wed-P-2-3 Second Language Acquisition and Code-switching....209

Wed-P-2-4 Topics in Speech Recognition....212
Wed-O-3-1 Zero-resource Speech Recognition....214
Wed-O-3-2 Spatial and Phase Cues for Source Separation and Speech Recognition....216
Wed-O-3-3 Dialectal Variation....217
Wed-O-3-4 Spoken Corpora and Annotation....218
Wed-SS-3-1 Special Session: The First DIHARD Speech Diarization Challenge....219
Wed-P-3-1 Text Analysis, Multilingual Issues and Evaluation in Speech Synthesis....221
Wed-P-3-2 Neural Network Training Strategies for ASR....223
Wed-P-3-3 Application of ASR in Medical Practice....225
Wed-P-3-4 Source and Supra-segmentals....226

## September 6, 2018, Thursday

Thu-Plenary-3 Plenary Talk-3....228
Thu-O-1-1 Distant ASR....229
Thu-O-1-2 Expressive Speech Synthesis....230
Thu-O-1-3 Representation Learning for Emotion....231
Thu-O-1-4 Articulatory Information, Modeling and Inversion....232
Thu-SS-1-1 Special Session: Novel Paradigms for Direct Synthesis Based on Speech-RelatedBiosignals....233
Thu-SS-1-2 Special Session: Low Resource Speech Recognition Challenge for Indian Languages....235
Thu-S&T-1-1 Show and Tell 7....236
Thu-P-1-1 Deep Enhancement....237
Thu-P-1-2 Acoustic Scenes and Rare Events....239
Thu-P-1-3 Language Modeling....241
Thu-P-1-4 Speech Pathology, Depression, and Medical Applications....243
Thu-Perspective-4 Perspective Talk-4....246
Thu-IP-7 Industry Presentation-7....246
Thu-IP-8 Industry Presentation-8....246
Thu-IP-9 Industry Presentation-9....246
Thu-O-2-1 Spoken Language Understanding....246
Thu-O-2-2 Source Separation from Monaural Input....247
Thu-O-2-3 Multimodal Systems....249
Thu-O-2-4 Coding....250
Thu-P-2-1 Speaker Verification Using Neural Network Methods II....251
Thu-P-2-2 Emotion Recognition and Analysis....254
Thu-P-2-3 Acoustic Modelling....257
Thu-P-2-4 Speech and Speaker Perception....259

# Abstracts

# ISCA Medal 2018 Talk

Hall 3; 11:00-12:00; Monday, 3 September, 2018
Chair: Bayya Yegnanarayana

## From Vocoders to Code-Excited Linear Prediction: Learning How We Hear What We Hear

*Bishnu S. Atal*
Department of Electrical Engineering, University of Washington, Seattle, WA 98195

**Mon-Medal-1, Time: 11:00-12:00**

It all started almost a century ago, in 1920s. A new undersea transatlantic telegraph cable had been laid. The idea of transmitting speech over the new telegraph cable caught the fancy of Homer Dudley, a young engineer who had just joined Bell Telephone Laboratories. This led to the invention of Vocoder - its close relative Voder was showcased as the first machine to create human speech at the 1939 New York World's Fair. However, the voice quality of vocoders was not good enough for use in commercial telephony. During the time speech scientists were busy with vocoders, several major developments took place outside speech research. Norbert Wiener developed a mathematical theory for calculating the best filters and predictors for detecting signals hidden in noise. Linear Prediction or Linear Predictive Coding became a major tool for speech processing. Claude Shannon established that the highest bit rate in a communication channel in presence of noise is achieved when the transmitted signal resembles random white Gaussian noise. Shannon's theory led to the invention of Code-Excited Linear Prediction (CELP). Nearly all digital cellular standards as well as standards for digital voice communication over the Internet use CELP coders. The success in speech coding came with understanding of what we hear and what we do not. Speech encoding at low bit rates introduce errors and these errors must be hidden under the speech signal to become inaudible. More and more, speech technologies are being used in different acoustic environments raising questions about the robustness of the technology. Human listeners handle situations well when the signal at our ears is not just one signal, but also a superposition of many acoustic signals. We need new research to develop signal-processing methods that can separate the mixed acoustic signal into individual components and provide performance similar or superior to that of human listeners.

# Mon-O-1-1: End-to-End Speech Recognition

Hall 3; 14:00-16:00; Monday, 3 September, 2018
Chairs: Dong Yu and Zoltán Tüske

## Semi-Supervised End-to-End Speech Recognition

*Shigeki Karita[1], Shinji Watanabe[2], Tomoharu Iwata[1], Atsunori Ogawa[1] and Marc Delcroix[1]*
[1]NTT Communication Science Laboratories
[2]Center for Language and Speech Processing, Johns Hopkins University

**Mon-O-1-1-1, Time: 14:00-14:20**

We propose a novel semi-supervised method for end-to-end automatic speech recognition (ASR). It can exploit large unpaired speech and text datasets, which require much less human effort to create paired speech-to-text datasets. Our semi-supervised method targets the extraction of an intermediate representation between speech and text data using a shared encoder network. Autoencoding of text data with this shared encoder improves the feature extraction of text data as well as that of speech data when the intermediate representations of speech and text are similar to each other as an inter-domain feature. In other words, by combining speech-to-text and text-to-text mappings through the shared network, we can improve speech-to-text mapping by learning to reconstruct the unpaired text data in a semi-supervised end-to-end manner. We investigate how to design suitable inter-domain loss, which minimizes the dissimilarity between the encoded speech and text sequences, which originally belong to quite different domains. The experimental results we obtained with our proposed semi-supervised training shows a larger character error rate reduction from 15.8% to 14.4% than a conventional language model integration on the Wall Street Journal dataset.

## Improved Training of End-to-End Attention Models for Speech Recognition

*Albert Zeyer[1,2,3], Kazuki Irie[1], Ralf Schlüter[1] and Hermann Ney[1,2]*
[1]Human Language Technology and Pattern Recognition, Computer Science Department
RWTH Aachen University, 52062 Aachen, Germany
[2]AppTek, USA, http://www.apptek.com/
[3]NNAISENSE, Switzerland, https://nnaisense.com

**Mon-O-1-1-2, Time: 14:20-14:40**

Sequence-to-sequence attention-based models on subword units allow simple open-vocabulary end-to-end speech recognition. In this work, we show that such models can achieve competitive results on the Switchboard 300h and LibriSpeech 1000h tasks. In particular, we report the state-of-the-art word error rates (WER) of 3.54% on the dev-clean and 3.82% on the test-clean evaluation subsets of LibriSpeech.

We introduce a new pretraining scheme by starting with a high time reduction factor and lowering it during training, which is crucial both for convergence and final performance. In some experiments, we also use an auxiliary CTC loss function to help the convergence. In addition, we train long short-term memory (LSTM) language models on subword unit. By shallow fusion, we report up to 27% relative improvements in WER over the attention baseline without a language model.

## End-to-End Speech Recognition Using Lattice-free MMI

*Hossein Hadian[1,2], Hossein Sameti[1], Daniel Povey[2,3] and Sanjeev Khudanpur[2,3]*
[1]Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
[2]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA
[3]Human Language Technology Center of Excellence, Johns Hopkins University, Baltimore, USA

**Mon-O-1-1-3, Time: 14:40-15:00**

We present our work on end-to-end training of acoustic models using the lattice-free maximum mutual information (LF-MMI) objective function in the context of hidden Markov models.

By end-to-end training, we mean flat-start training of a single DNN in one stage without using any previously trained models, forced alignments, or building state-tying decision trees. We use full biphones to enable context-dependent modeling without trees and show that our end-to-end LF-MMI approach can achieve comparable results to regular LF-MMI on well-known large vocabulary tasks.

We also compare with other end-to-end methods such as CTC in character-based and lexicon-free settings and show 5 to 25 percent relative reduction in word error rates on different large vocabulary tasks while using significantly smaller models.

## Multi-channel Attention for End-to-End Speech Recognition

*Stefan Braun, Daniel Neil, Jithendar Anumula, Enea Ceolini and Shih-Chii Liu*

**Notes**

Institute of Neuroinformatics, University of Zurich and ETH Zurich, Switzerland

**Mon-O-1-1-4, Time: 15:00-15:20**

Recent end-to-end models for automatic speech recognition use sensory attention to integrate multiple input channels within a single neural network. However, these attention models are sensitive to the ordering of the channels used during training. This work proposes a sensory attention mechanism that is invariant to the channel ordering and only increases the overall parameter count by 0.09%. We demonstrate that even without re-training, our attention-equipped end-to-end model is able to deal with arbitrary numbers of input channels during inference. In comparison to a recent related model with sensory attention, our model when tested on the real noisy recordings from the multi-channel CHiME-4 dataset, achieves a relative character error rate (CER) improvement of 40.3% to 42.9%. In a two-channel configuration experiment, the attention signal allows the lower signal-to-noise ratio (SNR) sensor to be identified with 97.7% accuracy.

### Quaternion Convolutional Neural Networks for End-to-End Automatic Speech Recognition

*Titouan Parcollet[1,2,4], Ying Zhang[2,5], Mohamed Morchid[1], Chiheb Trabelsi[2], Georges Linarès[1], Renato de Mori[1,3] and Yoshua Bengio[2]*

[1]Université d'Avignon, LIA, France

[2]Université de Montréal, MILA, Canada

[3]McGill University, Montréal, Canada

[4]Orkis, Aix en provence, France

[5]Element AI, Montréal, Canada

**Mon-O-1-1-5, Time: 15:20-15:40**

Recently, the connectionist temporal classification (CTC) model coupled with recurrent (RNN) or convolutional neural networks (CNN), made it easier to train speech recognition systems in an end-to-end fashion. However in real-valued models, time frame components such as mel-filter-bank energies and the cepstral coefficients obtained from them, together with their first and second order derivatives, are processed as individual elements, while a natural alternative is to process such components as composed entities. We propose to group such elements in the form of quaternions and to process these quaternions using the established quaternion algebra. Quaternion numbers and quaternion neural networks have shown their efficiency to process multidimensional inputs as entities, to encode internal dependencies and to solve many tasks with less learning parameters than real-valued models. This paper proposes to integrate multiple feature views in quaternion-valued convolutional neural network (QCNN), to be used for sequence-to-sequence mapping with the CTC model. Promising results are reported using simple QCNNs in phoneme recognition experiments with the TIMIT corpus. More precisely, QCNNs obtain a lower phoneme error rate (PER) with less learning parameters than a competing model based on real-valued CNNs.

### Compression of End-to-End Models

*Ruoming Pang, Tara Sainath, Rohit Prabhavalkar, Suyog Gupta, Yonghui Wu, Shuyuan Zhang and Chung-Cheng Chiu*

Google Inc., U.S.A

**Mon-O-1-1-6, Time: 15:40-16:00**

End-to-end models, which output text directly given speech using a single neural network, have been shown to be competitive with conventional speech recognition models containing separate acoustic, pronunciation and language model components. Such models do not require additional resources for decoding and are typically much smaller than conventional models. This makes them particularly attractive in the context of on-device speech recognition where both small memory footprint and low power consumption are critical. This work explores the problem of compressing end-to-end models with the goal of satisfying device constraints without sacrificing model accuracy. We evaluate matrix factorization, knowledge distillation and parameter sparsity to determine the most effective methods given constraints such as a fixed parameter budget.

**Notes**

## Mon-O-1-2: Prosody Modeling and Generation

Hall 1; 14:00-16:00; Monday, 3 September, 2018

Chairs: Oliver Watts and Esther Klabbers

### Learning Interpretable Control Dimensions for Speech Synthesis by Using External Data

*Zack Hodari, Oliver Watts, Srikanth Ronanki and Simon King*

The Centre for Speech Technology Research, University of Edinburgh, United Kingdom

**Mon-O-1-2-1, Time: 14:00-14:20**

There are many aspects of speech that we might want to control when creating text-to-speech (TTS) systems. We present a general method that enables control of arbitrary aspects of speech, which we demonstrate on the task of emotion control. Current TTS systems use supervised machine learning and are therefore heavily reliant on labelled data. If no labels are available for a desired control dimension, then creating interpretable control becomes challenging. We introduce a method that uses external, labelled data (i.e. not the original data used to train the acoustic model) to enable the control of dimensions that are not labelled in the original data. Adding interpretable control allows the voice to be manually controlled to produce more engaging speech, for applications such as audiobooks. We evaluate our method using a listening test.

### Investigating Accuracy of Pitch-accent Annotations in Neural Network-based Speech Synthesis and Denoising Effects

*Hieu-Thi Luong[1], Xin Wang[1], Junichi Yamagishi[1] and Nobuyuki Nishizawa[2]*

[1]National Institute of Informatics, Tokyo, Japan

[2]KDDI Research Inc., Saitama, Japan

**Mon-O-1-2-2, Time: 14:20-14:40**

We investigated the impact of noisy linguistic features on the performance of a Japanese speech synthesis system based on neural network that uses WaveNet vocoder. We compared an ideal system that uses manually corrected linguistic features including phoneme and prosodic information in training and test sets against a few other systems that use corrupted linguistic features. Both subjective and objective results demonstrate that corrupted linguistic features, especially those in the test set, affected the ideal system's performance significantly in a statistical sense due to a mismatched condition between the training and test sets. Interestingly, while an utterance-level Turing test showed that listeners had a difficult time differentiating synthetic speech from natural speech, it further indicated that adding noise to the linguistic features in the training set can partially reduce the effect of the mismatch, regularize the model and help the system perform better when linguistic features of the test set are noisy.

### An Exploration of Local Speaking Rate Variations in Mandarin Read Speech

*Guan-Ting Liou[1], Chen-Yu Chiang[2], Yih-Ru Wang[1] and Sin-Horng Chen[1]*

[1]Dept. of Electrical and Computer Engineering, National Chiao Tung University, Hsinchu, Taiwan

[2]Dept. of Communication Engineering, National Taipei University, New Taipei City, Taiwan

**Mon-O-1-2-3, Time: 14:40-15:00**

This paper explores speaking rate variation in Mandarin read speech. In contrast to assuming that each utterance is generated in a constant or global speaking rate, this study seeks to estimate local speaking rate for each prosodic unit in an utterance. The exploration is based on the existing speaking rate-dependent hierarchical prosodic model (SR-HPM).

The main idea is to first use the SR-HPM to explore the prosodic structures of utterances and extract the prosodic units. Then, local speaking rate is estimated for each prosodic unit (prosodic phrase in this study). Some major influence factors including tone, base syllable type, prosodic structure and speaking rate of the higher prosodic units (utterance and BG/PG) are compensated in the local SR estimation. A syntactic-local SR model is constructed and use in the prosody generation of Mandarin TTS. Experimental results on a large read speech corpus generated by a professional female announcer showed that the generated prosody with local speaking rate variations is proved to be more vivid than the one with a constant speaking rate.

### BLSTM-CRF Based End-to-End Prosodic Boundary Prediction with Context Sensitive Embeddings in a Text-to-Speech Front-End

*Yibin Zheng[1,2], Jianhua Tao[1,2], Zhengqi Wen[1] and Ya Li[1]*
[1]National Laboratory of Pattern Recognition, Institute of Automation, CAS, China
[2]School of Artificial Intelligence, University of Chinese Academy of Science, China

**Mon-O-1-2-4, Time: 15:00-15:20**

In this paper, we propose a language-independent end-to-end architecture for prosodic boundary prediction based on BLSTM-CRF. The proposed architecture has three components, word embedding layer, BLSTM layer and CRF layer. The word embedding layer is employed to learn the task-specific embeddings for prosodic boundary prediction. The BLSTM layer can efficiently use both past and future input features, while the CRF layer can efficiently use sentence level information. We integrate these three components and learn the whole process end-to-end. In addition, we investigate both character-level embeddings and context sensitive embeddings to this model and employ an attention mechanism for combining alternative word-level embeddings. By using an attention mechanism, the model is able to decide how much information to use from each level of embeddings. Objective evaluation results show the proposed BLSTM-CRF architecture achieves the best results on both Mandarin and English datasets, with an absolute improvement of 3.21% and 3.74% in F1 score, respectively, for intonational phrase prediction, compared to previous state-of-the-art method (BLSTM). The subjective evaluation results further indicate the effectiveness of the proposed methods.

### Wavelet Analysis of Speaker Dependent and Independent Prosody for Voice Conversion

*Berrak Sisman and Haizhou Li*
National University of Singapore, Singapore

**Mon-O-1-2-5, Time: 15:20-15:40**

Thus far, voice conversion studies are mainly focused on the conversion of spectrum. However, speaker identity is also characterized by its prosody features, such as fundamental frequency (F0) and energy contour. We believe that with a better understanding of speaker dependent/independent prosody features, we can devise an analytic approach that addresses voice conversion in a better way. We consider that speaker dependent features reflect speaker's individuality, while speaker independent features reflect the expression of linguistic content. Therefore, the former is to be converted while the latter is to be carried over from source to target during the conversion. To achieve this, we provide an analysis of speaker dependent and speaker independent prosody patterns in different temporal scales by using wavelet transform. The centrepiece of this paper is based on the understanding that a speech utterance can be characterized by speaker dependent and independent features in its prosodic manifestations. Experiments show that the proposed prosody analysis scheme improves the prosody conversion performance consistently under the sparse representation framework.

### Improving Mongolian Phrase Break Prediction by Using Syllable and Morphological Embeddings with BiLSTM Model

*Rui Liu, Feilong Bao, Guanglai Gao, Hui Zhang and Yonghe Wang*
College of Computer Science, Inner Mongolia University, Inner Mongolia Key Laboratory of Mongolian Information Processing Technology, Hohhot 010021, China

**Mon-O-1-2-6, Time: 15:40-16:00**

In the speech synthesis systems, the phrase break (PB) prediction is the first and most important step. Recently, the state-of-the-art PB prediction systems mainly rely on word embeddings. However this method is not fully applicable to Mongolian language, because its word embeddings are inadequate trained, owing to the lack of resources. In this paper, we introduce a bidirectional Long Short Term Memory (BiLSTM) model which combined word embeddings with syllable and morphological embedding representations to provide richer and multi-view information which leverages the agglutinative property. Experimental results show the proposed method outperforms compared systems which only used the word embeddings. In addition, further analysis shows that it is quite robust to the Out-of-Vocabulary (OOV) problem owe to the refined word embedding. The proposed method achieves the state-of-the-art performance in the Mongolian PB prediction.

## Mon-O-1-3: Speaker Verification I

Hall 2; 14:00-16:00; Monday, 3 September, 2018
Chairs: Jean-François Bonastre and Nicholas Evans

### Improved Supervised Locality Preserving Projection for I-vector Based Speaker Verification

*Lanhua You, Wu Guo, Yan Song and Sheng Zhang*
National Engineering Laboratory of Speech and Language Information Processing
University of Science and Technology of China, Hefei, China

**Mon-O-1-3-1, Time: 14:00-14:20**

A Supervised Locality Preserving Projection (SLPP) method is employed for channel compensation in an i-vector based speaker verification system. SLPP preserves more important local information by weighing both the within- and between-speaker nearby data pairs based on the similarity matrices. In this paper, we propose an improved SLPP (P-SLPP) to enhance the channel compensation ability. First, the conventional Euclidean distance in conventional SLPP is replaced with Probabilistic Linear Discriminant Analysis (PLDA) scores. Furthermore, the weight matrices of P-SLPP are generated by using the relative PLDA scores of within- and between-speaker pairs. Experiments are carried out on the five common conditions of NIST 2012 speaker recognition evaluation (SRE) core sets. The results show that SLPP and the proposed P-SLPP outperform all other state-of-the-art channel compensation methods. Among these methods, P-SLPP achieves the best performance.

### Double Joint Bayesian Modeling of DNN Local I-Vector for Text Dependent Speaker Verification with Random Digit Strings

*Ziqiang Shi, Huibin Lin, Liu Liu and Rujie Liu*
Fujitsu Research and Development Center

**Mon-O-1-3-2, Time: 14:20-14:40**

Double joint Bayesian is a recently introduced analysis method that models and explores multiple information explicitly from the samples to improve the verification performance. It was recently applied to voice pass phrase verification, result in better results on text dependent speaker verification task. However little is known about its effectiveness in other challenging situations such as speaker verification for short, text-constrained test utterances, e.g. random digit strings. Contrary to conventional joint Bayesian method that cannot make full use of multi-view information, double joint Bayesian can incorporate both intra-speaker/digit and inter-speaker/digit variation and calculated the likelihood to describe whether

**Notes**

the features having all labels consistent or not. We show that double joint Bayesian outperforms conventional method on modeling DNN local (digit-dependent) i-vectors for speaker verification with random prompted digit strings. Since the strength of both double joint Bayesian and conventional DNN local i-vector appear complementary, the combination significantly outperforms either of its components.

### Fast Variational Bayes for Heavy-tailed PLDA Applied to i-vectors and x-vectors

*Anna Silnova[1], Niko Brümmer[2], Daniel Garcia-Romero[3], David Snyder[3] and Lukáš Burget[1]*

[1]Brno University of Technology, Czech Republic

[2]Nuance Communications, South Africa

[3]Johns Hopkins HLTCOE, USA

**Mon-O-1-3-3, Time: 14:40-15:00**

The standard state-of-the-art backend for text-independent speaker recognizers that use i-vectors or x-vectors is Gaussian PLDA (G-PLDA), assisted by a Gaussianization step involving length normalization. G-PLDA can be trained with both generative or discriminative methods. It has long been known that heavy-tailed PLDA (HT-PLDA), applied without length normalization, gives similar accuracy, but at considerable extra computational cost. We have recently introduced a fast scoring algorithm for a discriminatively trained HT-PLDA backend. This paper extends that work by introducing a fast, variational Bayes, generative training algorithm. We compare old and new backends, with and without length-normalization, with i-vectors and x-vectors, on SRE'10, SRE'16 and SITW.

### Integrated Presentation Attack Detection and Automatic Speaker Verification: Common Features and Gaussian Back-end Fusion

*Massimiliano Todisco[1], Héctor Delgado[1], Kong Aik Lee[2], Md Sahidullah[3], Nicholas Evans[1], Tomi Kinnunen[4] and Junichi Yamagishi[5,6]*

[1]Department of Digital Security, EURECOM, France

[2]Data Science Research Laboratories, NEC Corporation, Japan

[3]MULTISPEECH, Inria, France

[4]School of Computing, University of Eastern Finland, Finland

[5]Digital Content and Media Sciences Research Division, National Institute of Informatics, Japan

[6]Centre of Speech Technology Research, University of Edinburgh, U.K.

**Mon-O-1-3-4, Time: 15:00-15:20**

The vulnerability of automatic speaker verification (ASV) systems to spoofing is widely acknowledged. Recent years have seen an intensification in research efforts to develop spoofing countermeasures, also known as presentation attack detection (PAD) systems. Much of this work has involved the exploration of features that discriminate reliably between bona fide and spoofed speech. While there are grounds to use different front-ends for ASV and PAD systems (they are different tasks) the use of a single front-end has obvious benefits, not least convenience and computational efficiency, especially when ASV and PAD are combined. This paper investigates the performance of a variety of different features used previously for both ASV and PAD and assesses their performance when combined for both tasks. The paper also presents a Gaussian back-end fusion approach to system combination. In contrast to cascaded architectures, it relies upon the modelling of the two-dimensional score distribution stemming from the combination of ASV and PAD in parallel. This approach to combination is shown to generalise particularly well across independent ASVspoof 2017 v2.0 development and evaluation datasets.

### A Generalization of PLDA for Joint Modeling of Speaker Identity and Multiple Nuisance Conditions

*Luciana Ferrer[1] and Mitchell McLaren[2]*

[1]Instituto de Investigación en Ciencias de la Computación, CONICET-Universidad de Buenos Aires, Buenos Aires, Argentina

[2]Speech Technology and Research Lab, SRI International, Menlo Park, USA

**Mon-O-1-3-5, Time: 15:20-15:40**

Probabilistic linear discriminant analysis (PLDA) is the leading method for computing scores in speaker recognition systems. The method models the vectors representing each audio sample as a sum of three terms: one that depends on the speaker identity, one that models the within-speaker variability and one that models any remaining variability. The last two terms are assumed to be independent across samples. We recently proposed an extension of the PLDA method, which we termed Joint PLDA (JPLDA), where the second term is considered dependent on the type of nuisance condition present in the data (e.g., the language or channel). The proposed method led to significant gains for multilanguage speaker recognition when taking language as the nuisance condition. In this paper, we present a generalization of this approach that allows for multiple nuisance terms. We show results using language and several nuisance conditions describing the acoustic characteristics of the sample and demonstrate that jointly including all these factors in the model leads to better results than including only language or acoustic condition factors. Overall, we obtain relative improvements in detection cost function between 5% and 47% for various systems and test conditions with respect to standard PLDA approaches.

### An Investigation of Non-linear i-vectors for Speaker Verification

*Nanxin Chen, Jesús Villalba and Najim Dehak*

Center for Language and Speech Processing

Johns Hopkins University, Baltimore, MD

**Mon-O-1-3-6, Time: 15:40-16:00**

Speaker verification becomes increasingly important due to the popularity of speech assistants and smart home. i-vectors are used broadly for this topic, which use factor analysis to model the shift of average parameter in Gaussian Mixture Models. Recently by the progress of deep learning, high-level non-linearity improves results in many areas. In this paper we proposed a new framework of i-vectors which uses stochastic gradient descent to solve the problem of i-vectors. From our preliminary results stochastic gradient descent can get same performance as expectation-maximization algorithm. However, by backpropagation the assumption can be more flexible, so both linear and non-linear assumption is possible in our framework. From our result, both maximum a posteriori estimation and maximum likelihood lead to slightly better result than conventional i-vectors and both linear and non-linear system has similar performance.

## Mon-O-1-4: Spoken Term Detection

MR G.01-G.02; 14:00-16:00; Monday, 3 September, 2018

Chairs: Xiaodong Cui and Thiago Fraga da Silva

### CNN Based Query by Example Spoken Term Detection

*Dhananjay Ram, Lesly Miculicich and Hervé Bourlard*

Idiap Research Institute, Martigny, Switzerland

École polytechnique fédérale de Lausanne, Switzerland

**Mon-O-1-4-1, Time: 14:00-14:20**

In this work, we address the problem of query by example spoken term detection (QbE-STD) in zero-resource scenario. State of the art solutions usually rely on dynamic time warping (DTW) based template matching. In contrast, we propose here to tackle the problem as binary classification of images. Similar to the DTW approach, we rely on deep neural network (DNN) based posterior probabilities as feature vectors. The posteriors from a spoken query and a test utterance are used to compute frame-level similarities in a matrix form. This matrix contains somewhere a quasi-

**Notes**

diagonal pattern if the query occurs in the test utterance. We propose to use this matrix as an image and train a convolutional neural network (CNN) for identifying the pattern and make a decision about the occurrence of the query. This language independent system is evaluated on SWS 2013 and is shown to give 10% relative improvement over a highly competitive baseline system based on DTW. Experiments on QUESST 2014 database gives similar improvements showing that the approach generalizes to other database as well.

## Learning Acoustic Word Embeddings with Temporal Context for Query-by-Example Speech Search

*Yougen Yuan[1], Cheung-Chi Leung[2], Lei Xie[1], Hongjie Chen[1], Bin Ma[2] and Haizhou Li[3]*

[1]School of Computer Science, Northwestern Polytechnical University, Xi'an, China

[2]Alibaba Inc., Singapore

[3]Department of Electrical and Computer Engineering, National University of Singapore, Singapore

**Mon-O-1-4-2, Time: 14:20-14:40**

We propose to learn acoustic word embeddings with temporal context for query-by-example (QbE) speech search. The temporal context includes the leading and trailing word sequences of a word. We assume that there exist spoken word pairs in the training database. We pad the word pairs with their original temporal context to form fixed-length speech segment pairs. We obtain the acoustic word embeddings through a deep convolutional neural network (CNN) which is trained on the speech segment pairs with a triplet loss. By shifting a fixed-length analysis window through the search content, we obtain a running sequence of embeddings. In this way, searching for the spoken query is equivalent to the matching of acoustic word embeddings. The experiments show that our proposed acoustic word embeddings learned with temporal context are effective in QbE speech search. They outperform the state-of-the-art frame-level feature representations and reduce run-time computation since no dynamic time warping is required in QbE speech search. We also find that it is important to have sufficient speech segment pairs to train the deep CNN for effective acoustic word embeddings.

## Siamese Recurrent Auto-Encoder Representation for Query-by-Example Spoken Term Detection

*Ziwei Zhu1, Zhiyong Wu[1,2], Runnan Li[1], Helen Meng[1,2] and Lianhong Cai[1]*

[1]Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University, China

[2]Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, China

**Mon-O-1-4-3, Time: 14:40-15:00**

With the explosive development of human-computer speech interaction, spoken term detection is widely required and has attracted increasing interest. In this paper, we propose a weak supervised approach using Siamese recurrent auto-encoder (RAE) to represent speech segments for query-by-example spoken term detection (QbyE-STD). The proposed approach exploits word pairs that contain different instances of the same/different word content as input to train the Siamese RAE. The encoder last hidden state vector of Siamese RAE is used as the feature for QbyE-STD, which is a fixed dimensional embedding feature containing mostly semantic content related information. The advantages of the proposed approach are: 1) extracting more compact feature with fixed dimension while keeping the semantic information for STD; 2) the extracted feature can describe the sequential phonetic structure of similar sounds to degree, which can be applied for zero-resource QbyE-STD. Evaluations on real scene Chinese speech interaction data and TIMIT confirm the effectiveness and efficiency of the proposed approach compared to the conventional ones.

## Fast Derivation of Cross-lingual Document Vectors from Self-attentive Neural Machine Translation Model

*Wei Li and Brian Mak*

Department of Computer Science and Engineering, The Hong Kong University of Science and Technology

**Mon-O-1-4-4, Time: 15:00-15:20**

A universal cross-lingual representation of documents, which can capture the underlying semantics is very useful in many natural language processing tasks. In this paper, we develop a new document vectorization method which effectively selects the most salient sequential patterns from the inputs to create document vectors via a self-attention mechanism using a neural machine translation (NMT) model. The model used by our method can be trained with parallel corpora that are unrelated to the task at hand. During testing, our method will take a monolingual document and convert it into a "Neural machine Translation framework based cross-lingual Document Vector" (NTDV). NTDV has two comparative advantages. Firstly, the NTDV can be produced by the forward pass of the encoder in the NMT and the process is very fast and does not require any training/optimization. Secondly, our model can be conveniently adapted from a pair of existing attention based NMT models and the training requirement on parallel corpus can be reduced significantly. In a cross-lingual document classification task, our NTDV embeddings surpass the previous state-of-the-art performance in the English-to-German classification test and, to our best knowledge, it also achieves the best performance among the fast decoding methods in the German-to-English classification test.

## LSTM Based Attentive Fusion of Spectral and Prosodic Information for Keyword Spotting in Hindi Language

*Laxmi Pandey[1] and Karan Nathwani[2]*

[1]Indian Institute of Technology, Kanpur, India

[2]Indian Institute of Technology, Jammu, India

**Mon-O-1-4-5, Time: 15:20-15:40**

In this paper, a DNN based keyword spotting framework, that utilizes both spectral as well as prosodic information present in the speech signal, is proposed. A DNN is first trained to learn a set of hierarchical non-linear transformation parameters that project the original spectral and prosodic feature vectors onto a feature space where the distance between similar syllable pairs is small and between dissimilar syllable pairs is large. These transformed features are then fused using an attention-based long short-term memory (LSTM) network. As a side result, a deep denoising autoencoder based fine-tuning technique is used to improve the performance of sequence predictions. A sequence matching method called the sliding syllable protocol is also developed for keyword spotting. Syllable recognition and keyword spotting (KWS) experiments are conducted specifically for the Hindi language which is one of the widely spoken languages across the globe but is not addressed significantly by the speech processing community. The proposed framework indicates reasonable improvements when compared to baseline methods available in the literature.

## Spoken Keyword Detection Using Joint DTW-CNN

*Ravi Shankar[1], Vikram C M[2] and S R Mahadeva Prasanna[2,3]*

[1]Johns Hopkins University, Baltimore

[2]Indian Institute of Technology Guwahati

[3]Indian Institute of Technology Dharwad

**Mon-O-1-4-6, Time: 15:40-16:00**

A method to detect spoken keywords in a given speech utterance is proposed, called as joint Dynamic Time Warping (DTW)- Convolution Neural Network (CNN). It is a combination of DTW approach with a strong classifier like CNN. Both these methods have independently shown significant results in solving problems related to optimal sequence alignment and object recognition, respectively. The proposed method modifies the original DTW formulation and converts the warping matrix into a gray scale image. A CNN is trained on these images to classify the presence

**Notes**

or absence of keyword by identifying the texture of warping matrix. The TIMIT corpus has been used for conducting experiments and our method shows significant improvement over other existing techniques.

# Mon-SS-1-1: Special Session: The INTERSPEECH 2018 Computational Paralinguistics ChallengE (ComParE): Atypical & Self-Assessed Affect, Crying & Heart Beats 1

MR 1.01-1.02; 14:00-16:00; Monday, 3 September, 2018
Chairs: Björn Schuller, Stefan Steidl, Anton Batliner, Peter B. Marschik, Harald Baumeister and Fengquan Dong

## The INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical & Self-Assessed Affect, Crying & Heart Beats

*Björn Schuller[1,2,3], Stefan Steidl[4], Anton Batliner[2,4], Peter B. Marschik[5,6,7], Harald Baumeister[8], Fengquan Dong[9], Simone Hantke[2,10], Florian B. Pokorny[5,10], Eva-Maria Rathner[8], Katrin D. Bartl-Pokorny[5], Christa Einspieler[5], Dajie Zhang[5,6], Alice Baird[2], Shahin Amiriparian[2,10], Kun Qian[2,10], Zhao Ren[2], Maximilian Schmitt[2], Panagiotis Tzirakis[1] and Stefanos Zafeiriou[1,11]*
[1]GLAM – Group on Language, Audio & Music, Imperial College London, UK
[2]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[3]audEERING GmbH, Gilching, Germany
[4]Pattern Recognition Lab, FAU Erlangen-Nuremberg, Germany
[5]iDN – interdisciplinary Developmental Neuroscience, Medical University of Graz, Austria
[6]University Medical Center Göttingen, Germany
[7]Department of Women's and Children's Health, Karolinska Institutet, Stockholm, Sweden
[8]Department of Clinical Psychology and Psychotherapy, University of Ulm, Germany
[9]Shenzhen University General Hospital, Shenzhen, P.R. China
[10]Machine Intelligence & Signal Processing Group, Technische Universität München, Germany
[11]University of Oulu, Finland

**Mon-SS-1-1-1, Time: 14:00-14:10**

The INTERSPEECH 2018 Computational Paralinguistics Challenge addresses four different problems for the first time in a research competition under well-defined conditions: In the Atypical Affect Sub-Challenge, four basic emotions annotated in the speech of handicapped subjects have to be classified; in the Self-Assessed Affect Sub-Challenge, valence scores given by the speakers themselves are used for a three-class classification problem; in the Crying Sub-Challenge, three types of infant vocalisations have to be told apart; and in the Heart Beats Sub-Challenge, three different types of heart beats have to be determined. We describe the Sub-Challenges, their conditions and baseline feature extraction and classifiers, which include data-learnt (supervised) feature representations by end-to-end learning, the 'usual' ComParE and BoAW features and deep unsupervised representation learning using the AUDEEP toolkit for the first time in the challenge series.

## Heart Beat Sub-Challenge

*Björn Schuller[1,2,3], Stefan Steidl[4], Anton Batliner[2,4], Peter B. Marschik[5,6,7], Harald Baumeister[8], Fengquan Dong[9], Simone Hantke[2,10], Florian B. Pokorny[5,10], Eva-Maria Rathner[8], Katrin D. Bartl-Pokorny[5], Christa Einspieler[5], Dajie Zhang[5,6], Alice Baird[2], Shahin Amiriparian[2,10], Kun Qian[2,10], Zhao Ren[2], Maximilian Schmitt[2], Panagiotis Tzirakis[1] and Stefanos Zafeiriou[1,11]*
[1]GLAM – Group on Language, Audio & Music, Imperial College London, UK
[2]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[3]audEERING GmbH, Gilching, Germany
[4]Pattern Recognition Lab, FAU Erlangen-Nuremberg, Germany
[5]iDN – interdisciplinary Developmental Neuroscience, Medical University of Graz, Austria
[6]University Medical Center Göttingen, Germany
[7]Department of Women's and Children's Health, Karolinska Institutet, Stockholm, Sweden
[8]Department of Clinical Psychology and Psychotherapy, University of Ulm, Germany
[9]Shenzhen University General Hospital, Shenzhen, P.R. China
[10]Machine Intelligence & Signal Processing Group, Technische Universität München, Germany
[11]University of Oulu, Finland

**Mon-SS-1-1-2, Time: 14:10-14:20**

## An Ensemble of Transfer, Semi-supervised and Supervised Learning Methods for Pathological Heart Sound Classification

*Ahmed Imtiaz Humayun[1], Md. Tauhiduzzaman Khan[1], Shabnam Ghaffarzadegan[2], Zhe Feng[2] and Taufiq Hasan[1]*
[1]mHealth Lab, Dept. of Biomedical Engineering, Bangladesh University of Engineering and Technology (BUET), Bangladesh.
[2]Human Machine Interaction Group-2, Robert Bosch Research and Technology Center (RTC), Sunnyvale, CA

**Mon-SS-1-1-3, Time: 14:20-14:30**

In this work, we propose an ensemble of classifiers to distinguish between various degrees of abnormalities of the heart using Phonocardiogram (PCG) signals acquired using digital stethoscopes in a clinical setting, for the INTERSPEECH 2018 Computational Paralinguistics (ComParE) Heart Beats Sub-Challenge. Our primary classification framework constitutes a convolutional neural network with 1D-CNN time-convolution (tConv) layers, which uses features transferred from a model trained on the 2016 Physionet Heart Sound Database. We also employ a Representation Learning (RL) approach to generate features in an unsupervised manner using Deep Recurrent Autoencoders and use Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA) classifiers. Finally, we utilize an SVM classifier on a high-dimensional segment-level feature extracted using various functionals on short-term acoustic features, i.e., Low-Level Descriptors (LLD). An ensemble of the three different approaches provides a relative improvement of 11.13% compared to our best single sub-system in terms of the Unweighted Average Recall (UAR) performance metric on the evaluation dataset.

## Crying Sub-Challenge

*Björn Schuller[1,2,3], Stefan Steidl[4], Anton Batliner[2,4], Peter B. Marschik[5,6,7], Harald Baumeister[8], Fengquan Dong[9], Simone Hantke[2,10], Florian B. Pokorny[5,10], Eva-Maria Rathner[8], Katrin D. Bartl-Pokorny[5], Christa Einspieler[5], Dajie Zhang[5,6], Alice Baird[2], Shahin Amiriparian[2,10], Kun Qian[2,10], Zhao Ren[2], Maximilian*

**Notes**

*Schmitt[2], Panagiotis Tzirakis[1] and Stefanos Zafeiriou[1,11]*
[1]GLAM – Group on Language, Audio & Music, Imperial College London, UK
[2]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[3]audEERING GmbH, Gilching, Germany
[4]Pattern Recognition Lab, FAU Erlangen-Nuremberg, Germany
[5]iDN – interdisciplinary Developmental Neuroscience, Medical University of Graz, Austria
[6]University Medical Center Göttingen, Germany
[7]Department of Women's and Children's Health, Karolinska Institutet, Stockholm, Sweden
[8]Department of Clinical Psychology and Psychotherapy, University of Ulm, Germany
[9]Shenzhen University General Hospital, Shenzhen, P.R. China
[10]Machine Intelligence & Signal Processing Group, Technische Universität München, Germany
[11]University of Oulu, Finland

**Mon-SS-1-1-4, Time: 14:30-14:40**

## Monitoring Infant's Emotional Cry in Domestic Environments Using the Capsule Network Architecture

*Mehmet Ali Tuğtekin Turan and Engin Erzin*
Multimedia, Vision and Graphics Laboratory,
College of Engineering, Koç University, Istanbul, Turkey

**Mon-SS-1-1-5, Time: 14:40-14:50**

Automated recognition of an infant's cry from audio can be considered as a preliminary step for the applications like remote baby monitoring. In this paper, we implemented a recently introduced deep learning topology called capsule network (CapsNet) for the cry recognition problem. A capsule in the CapsNet, which is defined as a new representation, is a group of neurons whose activity vector represents the probability that the entity exists. Active capsules at one level make predictions, via transformation matrices, for the parameters of higher-level capsules. When multiple predictions agree, a higher level capsule becomes active. We employed spectrogram representations from the short segments of an audio signal as an input of the CapsNet. For experimental evaluations, we apply the proposed method on INTERSPEECH 2018 computational paralinguistics challenge (ComParE), crying sub-challenge, which is a three-class classification task using an annotated database (CRIED). Provided audio samples contains recordings from 20 healthy infants and categorized into the three classes namely neutral, fussing and crying. We show that multi-layer CapsNet outperforms baseline performance on CRIED corpus and is considerably better than a conventional convolutional net.

## Neural Network Architecture That Combines Temporal and Summative Features for Infant Cry Classification in the Interspeech 2018 Computational Paralinguistics Challenge

*Mark Huckvale*
Speech, Hearing and Phonetic Sciences, University College London, U.K.

**Mon-SS-1-1-6, Time: 14:50-15:00**

This paper describes the application of a novel deep neural network architecture to the classification of infant vocalisations as part of the Interspeech 2018 Computational Paralinguistics Challenge. Previous approaches to infant cry classification have either applied a statistical classifier to summative features of the whole cry, or applied a syntactic pattern recognition technique to a temporal sequence of features. In this work we explore a deep neural network architecture that exploits both temporal and summative features to make a joint classification. The temporal input comprises centi-second frames of low-level signal features which are input to LSTM nodes, while the summative vector comprises a large set of statistical functionals of the same frames that are input to MLP nodes. The combined network is jointly optimized and evaluated using leave-one-speaker-out cross-validation on the challenge training set. Results are compared to independently-trained temporal and summative networks and to a baseline SVM classifier. The combined model outperforms the other models and the challenge baseline on the training set. While problems remain in finding the best configuration and training protocol for such networks, the approach seems promising for future signal classification tasks.

## Evolving Learning for Analysing Mood-Related Infant Vocalisation

*Zixing Zhang[1], Jing Han[2], Kun Qian [2,3] and Björn Schuller [1,2]*
[1]GLAM – Group on Language, Audio & Music, Imperial College London, UK
[2]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[3]Machine Intelligence & Signal Processing Group, MMK, Technical University of Munich, Germany

**Mon-SS-1-1-7, Time: 15:00-15:10**

Infant vocalisation analysis plays an important role in the study of the development of pre-speech capability of infants, while machine-based approaches nowadays emerge with an aim to advance such an analysis. However, conventional machine learning techniques require heavy feature-engineering and refined architecture designing. In this paper, we present an evolving learning framework to automate the design of neural network structures for infant vocalisation analysis. In contrast to manually searching by trial and error, we aim to automate the search process in a given space with less interference. This framework consists of a controller and its child networks, where the child networks are built according to the controller's estimation. When applying the framework to the Interspeech 2018 Computational Paralinguistics (ComParE) Crying Sub-challenge, we discover several deep recurrent neural network structures, which are able to deliver competitive results to the best ComParE baseline method.

## Atypical Affect Sub-Challenge

*Björn Schuller[1,2,3], Stefan Steidl[4], Anton Batliner[2,4], Peter B. Marschik[5,6,7], Harald Baumeister[8], Fengquan Dong[9], Simone Hantke[2,10], Florian B. Pokorny[5,10], Eva-Maria Rathner[8], Katrin D. Bartl-Pokorny[5], Christa Einspieler[5], Dajie Zhang[5,6], Alice Baird[2], Shahin Amiriparian[2,10], Kun Qian[2,10], Zhao Ren[2], Maximilian Schmitt[2], Panagiotis Tzirakis[1] and Stefanos Zafeiriou[1,11]*
[1]GLAM – Group on Language, Audio & Music, Imperial College London, UK
[2]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[3]audEERING GmbH, Gilching, Germany
[4]Pattern Recognition Lab, FAU Erlangen-Nuremberg, Germany
[5]iDN – interdisciplinary Developmental Neuroscience, Medical University of Graz, Austria
[6]University Medical Center Göttingen, Germany
[7]Department of Women's and Children's Health, Karolinska Institutet, Stockholm, Sweden
[8]Department of Clinical Psychology and Psychotherapy, University of Ulm, Germany
[9]Shenzhen University General Hospital, Shenzhen, P.R. China
[10]Machine Intelligence & Signal Processing Group, Technische Universität München, Germany
[11]University of Oulu, Finland

**Mon-SS-1-1-8, Time: 15:10-15:20**

**Notes**

### Deep Learning in Paralinguistic Recognition Tasks: Are Hand-crafted Features Still Relevant?

*Johannes Wagner, Dominik Schiller, Andreas Seiderer and Elisabeth André*

Human-Centered Multimedia, Augsburg University, Germany

**Mon-SS-1-1-9, Time: 15:20-15:30**

In the past, the performance of machine learning algorithms depended heavily on the representation of the data. Well-designed features therefore played a key role in speech and paralinguistic recognition tasks. Consequently, engineers have put a great deal of work into manually designing large and complex acoustic feature sets. With the emergence of Deep Neural Networks (DNNs), however, it is now possible to automatically infer higher abstractions from simple spectral representations or even learn directly from raw waveforms. This raises the question if (complex) hand-crafted features will still be needed in the future. We take this year's INTERSPEECH Computational Paralinguistic Challenge as an opportunity to approach this issue by means of two corpora - Atypical Affect and Crying. At first, we train a Recurrent Neural Network (RNN) to evaluate the performance of several hand-crafted feature sets of varying complexity. Afterwards, we make the network do the feature engineering all on its own by prefixing a stack of convolutional layers. Our results show that there is no clear winner (yet). This creates room to discuss chances and limits of either approach.

### Investigation on Joint Representation Learning for Robust Feature Extraction in Speech Emotion Recognition

*Danqing Luo[1], Yuexian Zou[1] and Dongyan Huang[2]*

[1]ADSPLAB, School of ECE, Peking University, Shenzhen, China

[2]Human Language Technology, Institute for Infocomm Research/ A*STAR, Singapore

**Mon-SS-1-1-10, Time: 15:30-15:40**

Speech emotion recognition (SER) is a challenging task due to its difficulty in finding proper representations for emotion embedding in speech. Recently, Convolutional Recurrent Neural Network (CRNN), which is combined by convolution neural network and recurrent neural network, is popular in this field and achieves state-of-art on related corpus. However, most of work on CRNN only utilizes simple spectral information, which is not capable to capture enough emotion characteristics for the SER task. In this work, we investigate two joint representation learning structures based on CRNN aiming at capturing richer emotional information from speech. Cooperating the handcrafted high-level statistic features with CRNN, a two-channel SER system (HSF-CRNN) is developed to jointly learn the emotion-related features with better discriminative property. Furthermore, considering that the time duration of speech segment significantly affects the accuracy of emotion recognition, another two-channel SER system is proposed where CRNN features extracted from different time scale of spectrogram segment are used for joint representation learning. The systems are evaluated over Atypical Affect Challenge of ComParE2018 and IEMOCAP corpus. Experimental results show that our proposed systems outperform the plain CRNN.

### Using Voice Quality Supervectors for Affect Identification

*Soo Jin Park, Amber Afshan, Zhi Ming Chua and Abeer Alwan*

Electrical and Computer Engineering Department, University of California Los Angeles, USA

**Mon-SS-1-1-11, Time: 15:40-15:50**

The voice quality of speech sounds often conveys perceivable information about the speaker's affect. This study proposes perceptually important voice quality features to recognize affect represented in speech excerpts from individuals with mental, neurological and/or physical disabilities. The voice quality feature set consists of F0, harmonic amplitude differences between the first, second, fourth harmonics and the harmonic near 2 kHz, the center frequency and amplitudes of the first 3 formants and cepstral peak prominence. The feature distribution of each utterance was represented with a supervector and the Gaussian mixture model and support vector machine classifiers were used for affect classification. Similar classification systems using the MFCCs and ComParE16 feature set were implemented. The systems were fused by taking the confidence mean of the classifiers. Applying the fused system to the Interspeech 2018 Atypical Affect subchallenge task resulted in unweighted average recalls of 43.9% and 41.0% on the development and test dataset, respectively. Additionally, we investigated clusters obtained by unsupervised learning to address gender-related differences.

### An End-to-End Deep Learning Framework for Speech Emotion Recognition of Atypical Individuals

*Dengke Tang[2], Junlin Zeng[2] and Ming Li[1]*

[1]Data Science Research Center, Duke Kunshan University, Kunshan, China

[2]School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China

**Mon-SS-1-1-12, Time: 15:50-16:00**

The goal of the ongoing ComParE 2018 Atypical Affect sub-challenge is to recognize the emotional states of atypical individuals. In this work, we present three modeling methods under the end-to-end learning framework, namely CNN combined with extended features, CNN+RNN and ResNet, respectively. Furthermore, we investigate multiple data augmentation, balancing and sampling methods to further enhance the system performance. The experimental results show that data balancing and augmentation increase the unweighted accuracy (UAR) by 10% absolutely. After score level fusion, our proposed system achieves 48.8% UAR on the develop dataset.

## Mon-S&T-1-1: Show and Tell 1

MR G.05-G.06; 14:00-16:00; Monday, 3 September, 2018

Chairs: R Padmanabhan and Ramakrishnan A G

### DialogOS: Simple and Extensible Dialogue Modeling

*Alexander Koller[1], Timo Baumann[2] and Arne Köhn[3]*

[1]Saarland University

[2]Carnegie Mellon University

[3]Universität Hamburg

**Mon-S&T-1-1-1, Time: 14:00-16:00**

We present the open-source extensible dialog manager DialogOS. DialogOS features simple finite-state based dialog management (which can be expanded to more complex DM strategies via a full-fledged scripting language) in combination with integrated speech recognition and synthesis in multiple languages. DialogOS runs on all major platforms, provides a simple-to-use graphical interface and can easily be extended via well-defined plugin and client interfaces, or can be integrated server-side into larger existing software infrastructures. We hope that DialogOS will help foster research and teaching given that it lowers the bar of entry into building and testing spoken dialog systems and provides paths to extend one's system as development progresses.

### A Framework for Speech Recognition Benchmarking

*Franck Dernoncourt, Trung Bui and Walter Chang*

Adobe Research, USA

**Mon-S&T-1-1-2, Time: 14:00-16:00**

Over the past few years, the number of APIs for automated speech recognition (ASR) has significantly increased. It is often time-consuming to evaluate how the performance of these ASR systems compare with each other and against newly proposed algorithms. In this paper, we present a lightweight, open source framework that allows users to easily benchmark ASR APIs on the corpora of their choice. The framework currently supports 7 ASR APIs and is easily extendable to more APIs.

**Notes**

### Flexible Tongue Housed in a Static Model of the Vocal Tract With Jaws, Lips and Teeth

*Takayuki Arai*

Department of Information and Communication Sciences, Japan

**Mon-S&T-1-1-3, Time: 14:00-16:00**

Physical models of the human vocal tract with a moveable tongue have been reported in past literature. In this study, we developed a new model with a flexible tongue. As with previous models by the author, the flexible tongue is made of gel material. The shape of this model's tongue is still an abstraction, although it is more realistic than previous models. Apart from the tongue, the model is static and solid; the gel tongue is the main part that can be manipulated. The static portion of the model is an extension of our recent static model with lips, teeth and tongue. The entire model looks like a sagittal splice taken from an artificial human head. Because the thin, acrylic plates on the outside are transparent, the interior of the oral and pharyngeal cavities are visible. When we feed a glottal sound through a hole in the laryngeal region on the bottom of the model, different vowels are produced, dependent upon the shape of the tongue. This model is the most useful and realistic looking of the models we've made for speech science education so far.

### Voice Analysis Using Acoustic and Throat Microphones for Speech Therapy

*Lani Mathew[1] and Gopakumar K.[2]*

[1]Mar Baselios College of Engineering, Thiruvananthapuram, Kerala, India

[2]TKM College of Engineering, Kollam, Kerala, India

**Mon-S&T-1-1-4, Time: 14:00-16:00**

Diagnosis of voice disorders by a speech therapist involves the process of voice recording with the patient, followed by software-aided analysis. In this paper, we propose a novel voice diagnosis system which gives voice report information based on Praat software, using voice samples from a throat microphone and an acoustic microphone, making the diagnosis near real-time, as well as robust to background noise. Results show that throat microphones give reliable Jitter and Shimmer values in ambient noise levels of 47~50 dB, while acoustic microphones show high variance in these parameters.

### A Robust Context-Dependent Speech-to-Speech Phraselator Toolkit for Alexa

*Manny Rayner[1], Nikos Tsourakis[1] and Jan Stanek[2]*

[1]University of Geneva, FTI/TIM, Switzerland

[2]University of South Australia, Adelaide, Australia

**Mon-S&T-1-1-5, Time: 14:00-16:00**

We present an open source toolkit for creating robust speech-to-speech phraselators, suitable for medical and other safety-critical domains, that can be hosted on the Amazon Alexa platform. Supported functionality includes context-dependent translation of incomplete utterances. We describe a preliminary evaluation on an English medical examination grammar.

## Mon-P-1-1: Speech Segments and Voice Quality

Hall 4-6: Poster-1; 14:00-16:00; Monday, 3 September, 2018

Chair: Narenda N P

### Discriminating Nasals and Approximants in English Language Using Zero Time Windowing

*RaviShankar Prasad, Sudarsana Reddy Kadiri, Suryakanth V Gangashetty and Bayya Yegnanarayana*

Speech Processing Laboratory, International Institute of Information Technology, Hyderabad, India

**Mon-P-1-1-1, Time: 14:00-16:00**

Nasals and approximants consonants are often confused with each other. Despite the distinction in the production mechanism, these two sound classes exhibit a similar low frequency behavior and lack significant high frequency content. The present study uses a spectral representation obtained using the zero time windowing (ZTW) analysis of speech, for the task of distinction between these two. The instantaneous spectral representation has good resolution at resonances, which helps to highlight the difference in the acoustic vocal tract system response for these sounds. The ZTW spectra around the regions of glottal closure instants are averaged to derive parameters for their classification in continuous speech. A set of parameters based on the dominant resonances, center of gravity, band energy ratio and cumulative spectral sum in low frequencies, is derived from the average spectrum. The paper proposes classification using a knowledge-based approach and training a support vector machine. These classifiers are tested on utterances from different English speakers in the TIMIT dataset. The proposed methods result in an average classification accuracy of 90% between the two classes in continuous speech.

### Gestural Lenition of Rhotics Captures Variation in Brazilian Portuguese

*Phil Howson and Alexei Kochetov*

University of Toronto

**Mon-P-1-1-2, Time: 14:00-16:00**

The goal of this study is to examine the rhotics in Brazilian Portuguese (BP), /ɾ,ʁ/ and the 'archetypal' coda /R/, to determine if: (1) they can be characterized as a coordination of the tongue dorsum and tongue body or tip and (2) manipulation of the gestural settings accounts for rhotic allophony in BP.

Six native speakers of BP participated in an ultrasound experiment and produced target phonemes in #CV, VCV and VC# environments with the vowels /i, e, a, o/. Tongue contours for the rhotics were compared using Smoothing Spline ANOVAs. /ɾ, ʁ/ were produced with a tongue body and dorsum gesture, while /ɾ/ also had an apical gesture. Archetypal /R/ was realized variably, as any of [ɾ, ɻ, ɹ, χ].

BP rhotics can be described as the coordination of a tongue dorsum and a tongue body or tip gesture. 'Archetypal' /R/ is posited to be /ɾ/. Allophony between /ɾ/ and [ɻ, ɹ, χ] is due to tongue tip lenition. Allophony between /ʁ/ and [h] is due to weakening of the tongue dorsum and body gestures. This analysis suggests synchronic and diachronic changes of rhotics result from lenition. It also captures the rarity of diachronic changes from uvulars to alveolars.

### Identification and Classification of Fricatives in Speech Using Zero Time Windowing Method

*RaviShankar Prasad and Bayya Yegnanarayana*

Speech Processing Laboratory, International Institute of Information Technology, Hyderabad, India

**Mon-P-1-1-3, Time: 14:00-16:00**

Fricatives are produced by creating a turbulence in the air-flow by passing it through a stricture in the vocal tract cavity. Fricatives are characterized by their noise-like behavior, which makes it difficult to analyze. Difference in the place of articulation leads to different classes of fricatives. Identification of fricative segment boundaries in speech helps in improving the performance of several applications. The present study attempts towards the identification and classification of fricative segments in continuous speech, based on the statistical behavior of instantaneous spectral characteristics. The proposed method uses parameters such as the dominant resonance frequencies, the center of gravity along with the statistical moments of the spectrum obtained using the zero time windowing (ZTW) method. The ZTW spectra exhibits a high temporal resolution and therefore gives accurate segment boundaries in speech. The proposed algorithm is tested on the TIMIT dataset for English language. A high identification rate of 97.5% is achieved for segment boundaries of the sibilant fricative class. Voiced nonsibilants

**Notes**

show a lower identification rate than their voiceless counterparts due to their vowel-like spectral characteristics. A high classification rate of 93.2% is achieved between sibilants and nonsibilants.

## GlobalTIMIT: Acoustic-Phonetic Datasets for the World's Languages

*Nattanun Chanchaochai[1], Christopher Cieri[1], Japhet Debrah[1], Hongwei Ding[2], Yue Jiang[3], Sishi Liao[2], Mark Liberman[1], Jonathan Wright[1], Jiahong Yuan[1], Juhong Zhan[3] and Yuqing Zhan[2]*

[1]Linguistic Data Consortium, University of Pennsylvania, U.S.A.

[2]School of Foreign Languages, Shanghai Jiao Tong University, P.R.C.

[3]School of Foreign Studies, Xi'an Jiaotong University, P.R.C.

**Mon-P-1-1-4, Time: 14:00-16:00**

Although the TIMIT acoustic-phonetic dataset ([1], [2]) was created three decades ago, it remains in wide use, with more than 20000 Google Scholar references and more than 1000 since 2017. Despite TIMIT's antiquity and relatively small size, inspection of these references shows that it is still used in many research areas: speech recognition, speaker recognition, speech synthesis, speech coding, speech enhancement, voice activity detection, speech perception, overlap detection and source separation, diagnosis of speech and language disorders and linguistic phonetics, among others.

Nevertheless, comparable datasets are not available even for other widely-studied languages, much less for under-documented languages and varieties. Therefore, we have developed a method for creating TIMIT-like datasets in new languages with modest effort and cost and we have applied this method in standard Thai, standard Mandarin Chinese, English from Chinese L2 learners, the Guanzhong dialect of Mandarin Chinese and the Ga language of West Africa. Other collections are planned or underway.

The resulting datasets will be published through the LDC, along with instructions and open-source tools for replicating this method in other languages, covering the steps of sentence selection and assignment to speakers, speaker recruiting and recording, proof-listening and forced alignment.

## Structural Effects on Properties of Consonantal Gestures in Tashlhiyt

*Anne Hermes[1], Doris Mücke[1], Bastian Auris[1] and Rachid Ridouane[2]*

[1]IfL – Phonetics, University of Cologne, Germany

[2]LPP, UMR7018, CNRS-Sorbonne Nouvelle, Paris, France

**Mon-P-1-1-5, Time: 14:00-16:00**

Tashlhiyt Berber is a language in which every consonant can take up the nucleus position in a syllable. The present study investigates how gestural properties are modified when the consonants occur in different syllable positions (onset, nucleus, coda). Furthermore, the effect of higher structural components such as morphology on the respective gestural organization patterns are examined.

Therefore, we collected articulographic data for different consonantal roots, such as /bdg/ and /gzm/ with varying affixes, entailing different syllabification patterns in Tashlhiyt. Consonantal properties in different syllable positions are investigated with respect to their intragestural properties and intergestural properties, i.e. bonding strength. Furthermore, gestural coherence with respect to prefixation were examined.

Results reveal that consonantal gestures were not modified on the intragestural level in terms of duration, velocity, stiffness or displacement, when the morphological structure was kept constant. However, on the intergestural level syllable relation was encoded, revealing a tighter bonding for onset-nucleus relations than for heterosyllabic sequences. Furthermore, when changing the morphological marker, modifications of intragestural parameters occur, inducing temporal changes of consonantal gestures. We conclude that higher structural components should be taken into account when investigating syllable internal timing patterns.

**Notes**

## The Retroflex-dental Contrast in Punjabi Stops and Nasals: A Principal Component Analysis of Ultrasound Images

*Alexei Kochetov[1], Matthew Faytak[2] and Kiranpreet Nara[1]*

[1]University of Toronto

[2]University of California, Berkeley

**Mon-P-1-1-6, Time: 14:00-16:00**

Many languages of South Asia show a phonemic contrast between retroflexes and dentals across different manners of articulation. This contrast, however, tends to be less phonetically distinct and more variable in nasals.

The goal of this paper is to examine the overall similarity of the retroflex-dental contrasts in Punjabi stops and nasals. Ultrasound tongue imaging recordings were obtained from 14 Punjabi speakers producing /ʈ,ɳ,t,n/ in the /ba_ab/ nonsense word context. Selected video frames were fed to a principal component analysis (PCA); the output was used for (1) training a linear discriminant model on one manner that discriminates place and (2) testing it on the other manner.

The results showed 100% correct classification of the contrast (retroflex or dental) in stops and 92% correct classification in nasals in the training data. The classification was much poorer across different manners: on average 67% of stops and 57% of nasals were classified correctly based on training sets with nasals or stops, respectively. In both cases, retroflex responses were more common.

These results suggest that the tongue configurations for Punjabi retroflex and dental consonants differ by manner of articulation. The contrast is also overall less robust in nasals than in stops, confirming previous reports.

## Vowels and Diphthongs in Hangzhou Wu Chinese Dialect

*Yang Yue[1] and Fang Hu[2]*

[1]Department of Linguistics, Graduate School, Chinese Academy of Social Sciences

[2]Institute of Linguistics, Chinese Academy of Social Sciences

**Mon-P-1-1-7, Time: 14:00-16:00**

This paper gives an acoustic phonetic description of the vowels and diphthongs in Hangzhou Wu Chinese dialect. Data from 12 speakers, 6 male and 6 female, were measured and analyzed. Monophthongs were investigated in CV, CVN and CVC syllables; diphthongs were examined in terms of temporal organization, spectral properties and dynamic aspects. Results suggest that falling diphthongs tend to have a single dynamic target, while rising diphthongs have two static spectral targets.

## Resyllabification in Indian Languages and Its Implications in Text-to-speech Systems

*Mahesh M, Jeena J Prakash and Hema Murthy*

Department of Computer Science and Engineering

Indian Institute of Technology Madras

**Mon-P-1-1-8, Time: 14:00-16:00**

Resyllabification is a phonological process in continuous speech in which the coda of a syllable is converted into the onset of the following syllable, either in the same word or in the subsequent word. This paper presents an analysis of resyllabification across words in different Indian languages and its implications in Indian language text-to-speech (TTS) synthesis systems. The evidence for resyllabification is evaluated based on the acoustic analysis of a read speech corpus of the corresponding language. This study shows that the resyllabification obeys the maximum onset principle and introduces the notion of prominence resyllabification in Indian languages. This paper finds acoustic evidence for total resyllabification.

The resyllabification rules obtained are applied to TTS systems. The correctness of the rules is evaluated quantitatively by comparing the acoustic log-likelihood scores of the speech utterances with the original and resyllabified texts and by performing a pair comparison (PC) listening

test on the synthesized speech output. An improvement in the log-likelihood score with the resyllabified text is observed and the synthesized speech with the resyllabified text is preferred 3 times over those without resyllabification.

## Voice Source Contribution to Prominence Perception: Rd Implementation

*Andy Murphy, Irena Yanushevskaya, Ailbhe Ní Chasaide and Christer Gobl*

Trinity College Dublin, Ireland

**Mon-P-1-1-9, Time: 14:00-16:00**

This paper explores the contribution of voice source modulation to the perception of prominence, following on previous analyses of accentuation, focus and deaccentuation. A listening test was carried out on a sentence of Irish with three accented, prominent syllables (P1, P2, P3). Using inverse filtering and resynthesis, a 'flattened' version was generated, with only slight declination of f0 and other voice source parameters. The global waveshape parameter Rd was modulated to provide (i) source boosting (tenser phonation) on either P1 or P2 and/or (ii) source attenuation (laxer phonation) following (Post-attenuation) or preceding (Pre-attenuation) P1 or P2. Rd variation was achieved in two different ways to generate two series of stimuli. f0 was not varied in either series. Twenty-nine listeners rated the prominence level of all syllables in the utterance. Results show that the phrasal position (P1 vs. P2) makes a large difference to prominence judgements. P1 emerged as overall more prominent and more readily 'enhanced' by the source modifications. Post-attenuation was particularly important for P1, with effects equal to or greater than local P-boosting. In the case of P2, Pre-attenuation was much more important than Post-attenuation.

## On the Relationship between Glottal Pulse Shape and Its Spectrum: Correlations of Open Quotient, Pulse Skew and Peak Flow with Source Harmonic Amplitudes

*Christer Gobl, Andy Murphy, Irena Yanushevskaya and Ailbhe Ní Chasaide*

Phonetics and Speech Laboratory, School of Linguistic, Speech and Communication Sciences, Trinity College Dublin, Ireland

**Mon-P-1-1-10, Time: 14:00-16:00**

This paper explores the relationship between the glottal pulse amplitude (Up) and the amplitude of the first harmonic (H1), as well as the combined effects of Up, the open quotient (Oq) and degree of pulse asymmetry/skew (Rk) on the low end of the source spectrum. This serves to elucidate their relationship to the H1-H2 estimate, widely used to make inferences on changes in Oq and voice quality. It has been suggested that H1 is mainly determined by Up and that the pulse shape has a relatively small impact. To investigate this, a series of glottal pulses were generated using the LF model, where Up was kept constant, while Oq and Rk were systematically varied. The resulting harmonic amplitudes of these pulses show that Up is not the sole determinant of H1. Rather, H1 is highly dependent on Oq and to a certain degree also on Rk. Although the effects of these parameters on the lowest harmonics is rather complex, we find that the H1-H2 measure is broadly correlated with Oq. However, there is also a strong effect of differences in glottal skew, particularly at high Oq values, which could invalidate inferences on Oq and voice quality from estimates of H1-H2.

## The Individual and the System: Assessing the Stability of the Output of a Semi-automatic Forensic Voice Comparison System

*Vincent Hughes[1], Philip Harrison[1,2], Paul Foulkes[1], Peter French[1,2], Colleen Kavanagh[1] and Eugenia San Segundo Fernández[1]*

[1]Department of Language and Linguistic Science, University of York, UK

[2]J P French Associates, York, UK

**Mon-P-1-1-11, Time: 14:00-16:00**

Semi-automatic systems based on traditional linguistic-phonetic features are increasingly being used for forensic voice comparison (FVC) casework. In this paper, we examine the stability of the output of a semi-automatic system, based on the long-term formant distributions (LTFDs) of F1, F2 and F3, as the channel quality of the input recordings decreases. Cross-validated, calibrated GMM-UBM log likelihood-ratios (LLRs) were computed for 97 Standard Southern British English speakers under four conditions. In each condition the same speech material was used, but the technical properties of the recordings changed (high quality studio recording, landline telephone recording, high bit-rate GSM mobile telephone recording and low bit-rate GSM mobile telephone recording). Equal error rate (EER) and the log LR cost function (Cllr) were compared across conditions. System validity was found to decrease with poorer technical quality, with the largest differences in EER (21.66%) and Cllr (0.46) found between the studio and the low bit-rate GSM conditions. However, importantly, performance for individual speakers was affected differently by channel quality. Speakers that produced stronger evidence overall were found to be more variable. Mean F3 was also found to be a predictor of LLR variability, however no effects were found based on speakers' voice quality profiles. of H1-H2.

## Breathy to Tense Voice Discrimination using Zero-Time Windowing Cepstral Coefficients (ZTWCCs)

*Sudarsana Reddy Kadiri and Bayya Yegnanarayana*

Speech Processing Laboratory,

International Institute of Information Technology, Hyderabad, India

**Mon-P-1-1-12, Time: 14:00-16:00**

In this paper, we consider breathy to tense voices, which are often considered to be opposite ends of a voice quality continuum. Along with these, other aspects of a speaker's voice play an important role to convey the information to the listener such as mood, attitude and emotional state. The glottal pulse characteristics in different phonation types vary due to the tension of laryngeal muscles together with the respiratory effort. In the present study, we are deriving the features that can capture effects of excitation on the vocal tract system through a signal processing method, called as zero-time windowing (ZTW) method. The ZTW method gives the instantaneous spectrum which captures the changes in the speech production mechanism, providing higher spectral resolution. The cepstral coefficients derived from ZTW method are used for the classification of phonation types. Along with zero-time windowing cepstral coefficients (ZTWCCs), we use the excitation source features derived from zero frequency filtering (ZFF) method. The excitation features used are: strength of excitation, energy of excitation, loudness measure and ZFF signal energy. Classification experiments using ZTWCC and excitation features reveal a significant improvement in the detection of phonation type compared to the existing voice quality features and MFCC features.

## Analysis of Breathiness in Contextual Vowel of Voiceless Nasals in Mizo

*Pamir Gogoi [1], Sishir Kalita [2], Parismita Gogoi [2], Ratree Wayland [1], Priyankoo Sarmah [2] and S R Mahadeva Prasanna [2,3]*

[1]University of Florida, Gainesville, USA

[2]Indian Institute of Technology Guwahati, Guwahati, India

[3]Indian Institute of Technology Dharwad, Dharwad, India

**Mon-P-1-1-13, Time: 14:00-16:00**

This study analyses the source characteristics of voiced and voiceless nasals in Mizo, a Tibeto-Burman language spoken in North-East India. Mizo is one of the few languages that has voiced and voiceless nasals in its phoneme inventory. This analysis is motivated by the interaction between breathiness and nasality reported in a number of speech perception studies using synthetic stimuli. However, there are no studies examining this interaction in vowels after voiced and voiceless nasals. Existing research has also documented the interaction between breathy phonation and vowel height. The current study is an acoustic analysis of breathiness in high and low vowels following voiced and voiceless nasals in Mizo. The acoustic parameter measures are: H1H2 ratio, spectral balance (SB), strength of excitation (SoE) and waveform peak factor (WPF). The values obtained for all the four acoustic measures suggest that vowels following voiceless nasals exhibit stronger acoustic characteristics associated with breathy phonation than vowels following voiced nasals. In addition, the degree of acoustic breathiness is affected by vowel height.

**Notes**

## Mon-P-1-2: Speaker State and Trait

Hall 4-6: Poster-2; 14:00-16:00; Monday, 3 September, 2018
Chair: Vidhyasaharan Sethu

### Infant Emotional Outbursts Detection in Infant-parent Spoken Interactions

*Yijia Xu[1,2], Mark Hasegawa-Johnson[1,2] and Nancy McElwain[2,3]*
[1]Department of Electrical and Computer Engineering
[2]Beckman Institute for Advanced Science and Technology
[3]Department of Human Development and Family Studies,
University of Illinois at Urbana Champaign, IL, USA

**Mon-P-1-2-1, Time: 14:00-16:00**

Detection of infant emotional outbursts, such as crying, in large corpora of recorded infant speech, is essential to the study of dyadic social process, by which infants learn to identify and regulate their own emotions. Such large corpora now exist with the advent of LENA speech monitoring systems, but are not labeled for emotional outbursts. This paper reports on our efforts to manually code child utterances as being of type "laugh", "cry", "fuss", "babble" and "hiccup" and to develop algorithms capable of performing the same task automatically. Human labelers achieve much higher rates of inter-coder agreement for some of these categories than for others. Linear discriminant analysis (LDA) achieves better accuracy on tokens that have been coded by two human labelers than on tokens that have been coded by only one labeler, but the difference is not as much as we expected, suggesting that the acoustic and contextual features being used by human labelers are not yet available to the LDA. Convolutional neural network and hidden markov model achieve better accuracy than LDA, but worse F-score, because they over-weight the prior. Discounting the transition probability does not solve the problem.

### Deep Neural Networks for Emotion Recognition Combining Audio and Transcripts

*Jaejin Cho[1], Raghavendra Pappagari[1], Purva Kulkarni[2], Jesús Villalba[1], Yishay Carmiel[2] and Najim Dehak[1]*
[1]Center for Language Speech Processing, Johns Hopkins University, Baltimore, MD, USA
[2]IntelligentWire, Seattle, WA, USA

**Mon-P-1-2-2, Time: 14:00-16:00**

In this paper, we propose to improve emotion recognition by combining acoustic information and conversation transcripts. On the one hand, an LSTM network was used to detect emotion from acoustic features like f0, shimmer, jitter, MFCC, etc. On the other hand, a multi-resolution CNN was used to detect emotion from word sequences. This CNN consists of several parallel convolutions with different kernel sizes to exploit contextual information at different levels. A temporal pooling layer aggregates the hidden representations of different words into a unique sequence level embedding, from which we compute the emotion posteriors. We optimized a weighted sum of classification and verification losses. The verification loss tries to bring embeddings from same emotions closer while it separates embeddings for different emotions. We also compared our CNN with state-of-the-art text-based hand-crafted features (e-vector). We evaluated our approach on the USC-IEMOCAP dataset as well as the dataset consisting of US English telephone speech. In the former, we used human transcripts while in the latter, we used ASR transcripts. The results showed fusing audio and transcript information improved unweighted accuracy by relative 24% for IEMOCAP and relative 3.4% for the telephone data compared to a single acoustic system.

### Preference-Learning with Qualitative Agreement for Sentence Level Emotional Annotations

*Srinivas Parthasarathy and Carlos Busso*
Multimodal Signal Processing(MSP) lab, Department of Electrical and Computer Engineering, The University of Texas at Dallas, Richardson TX 75080, USA

**Mon-P-1-2-3, Time: 14:00-16:00**

The perceptual evaluation of emotional attributes is noisy due to inconsistencies between annotators. The low inter-evaluator agreement arises due to the complex nature of emotions. Conventional approaches average scores provided by multiple annotators. While this approach reduces the influence of dissident annotations, previous studies have showed the value of considering individual evaluations to better capture the underlying ground-truth. One of these approaches is the qualitative agreement (QA) method, which provides an alternative framework that captures the inherent trends amongst the annotators. While previous studies have focused on using the QA method for time-continuous annotations from a fixed number of annotators, most emotional databases are annotated with attributes at the sentence-level (e.g., one global score per sentence). This study proposes a novel formulation based on the QA framework to estimate reliable sentence-level annotations for preference-learning. The proposed relative labels between pairs of sentences capture consistent trends across evaluators. The experimental evaluation shows that preference-learning methods to rank-order emotional attributes trained with the proposed QA-based labels achieve significantly better performance than the same algorithms trained with relative scores obtained by averaging absolute scores across annotators. These results show the benefits of QA-based labels for preference-learning using sentence-level annotations.

### Transfer Learning for Improving Speech Emotion Classification Accuracy

*Siddique Latif[1,3], Rajib Rana[2], Shahzad Younis[3], Junaid Qadir[1] and Julien Epps[4]*
[1]Information Technology University (ITU)-Punjab, Pakistan
[2]University of Southern Queensland, Australia
[3]National University of Sciences and Technology (NUST), Pakistan
[4]The University of New South Wales, Sydney, Australia

**Mon-P-1-2-4, Time: 14:00-16:00**

The majority of existing speech emotion recognition research focuses on automatic emotion detection using training and testing data from same corpus collected under the same conditions. The performance of such systems has been shown to drop significantly in cross-corpus and cross-language scenarios. To address the problem, this paper exploits a transfer learning technique to improve the performance of speech emotion recognition systems that is novel in cross-language and cross-corpus scenarios. Evaluations on five different corpora in three different languages show that Deep Belief Networks (DBNs) offer better accuracy than previous approaches on cross-corpus emotion recognition, relative to a Sparse Autoencoder and Support Vector Machine (SVM) baseline system. Results also suggest that using a large number of languages for training and using a small fraction of the target data in training can significantly boost accuracy compared with baseline also for the corpus with limited training examples.

### What Do Classifiers Actually Learn? A Case Study on Emotion Recognition Datasets

*Patrick Meyer, Eric Buschermöhle and Tim Fingscheidt*
Institute for Communications Technology
Technische Universität Braunschweig
38106 Braunschweig, Germany

**Mon-P-1-2-5, Time: 14:00-16:00**

In supervised learning, a typical method to ensure that a classifier has desirable generalization properties, is to split the available data into training, validation and test subsets. Given a proper data split, we typically then trust our results on the test data. But what do classifiers actually learn? In this case study we show how important it is to analyze precisely the available data, its inherent dependencies w.r.t. class labels and present an example of a popular database for speech emotion recognition, where a minor change of the data split results in an accuracy decrease of about 55% absolute, leading to the conclusion that linguistic content has been learned instead of the desired speech emotions.

**Notes**

## State of Mind: Classification through Self-reported Affect and Word Use in Speech

*Eva-Maria Rathner[1], Yannik Terhorst[1], Nicholas Cummins[2], Björn Schuller[2,3] and Harald Baumeister[1]*

[1]Clinical Psychology and Psychotherapy, University of Ulm, Germany

[2]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany

[3]GLAM - Group on Language, Audio & Music, Imperial College London, UK

**Mon-P-1-2-6, Time: 14:00-16:00**

Human state–of-mind (SOM; e.g.: perception, cognition, attention) constantly shifts due to internal and external demands. Mental health is influenced by the habitual use of either adaptive or maladaptive SOM. Therefore, the training of conscious regulation of SOM could be promising in self-help (e- and m-health), blended care and psychotherapy. The presented study indicates that SOM can be influenced by telling personal narratives. Furthermore, SOM and narrative sentiment (positive vs. negative) can be predicted through word use. Such results lay the groundwork for the development of applications that analyse text and speech for: i) the early detection of mental health; ii) the early detection of maladaptive changes in emotion dynamics; (iii) the use of personal narratives to improve emotion regulation skills; iv) the distribution of tailored interventions; and finally, v) evaluation of therapy outcome.

## Exploring Spatio-Temporal Representations by Integrating Attention-based Bidirectional-LSTM-RNNs and FCNs for Speech Emotion Recognition

*Ziping Zhao[1], Yu Zheng[1], Zixing Zhang[2], Haishuai Wang[3], Yiqin Zhao[1], Chao Li[1]*

[1]College of Computer and Information Engineering, Tianjin Normal University

[2]Group on Language, Audio & Music, Imperial College London, UK

[3]Department of Biomedical Informatics, Harvard University

**Mon-P-1-2-7, Time: 14:00-16:00**

Automatic emotion recognition from speech, which is an important and challenging task in the field of affective computing, heavily relies on the effectiveness of the speech features for classification. Previous approaches to emotion recognition have mostly focused on the extraction of carefully hand-crafted features. How to model spatio-temporal dynamics for speech emotion recognition effectively is still under active investigation. In this paper, we propose a method to tackle the problem of emotional relevant feature extraction from speech by leveraging Attention-based Bidirectional Long Short-Term Memory Recurrent Neural Networks with fully convolutional networks in order to automatically learn the best spatio-temporal representations of speech signals. The learned high-level features are then fed into a deep neural network (DNN) to predict the final emotion. The experimental results on the Chinese Natural Audio-Visual Emotion Database (CHEAVD) and the Interactive Emotional Dyadic Motion Capture (IEMOCAP) corpora show that our method provides more accurate predictions compared with other existing emotion recognition algorithms.

## End-to-End Deep Neural Network Age Estimation

*Pegah Ghahremani[1], Phani Sankar Nidadavolu[1], Nanxin Chen[1], Jesús Villalba[1], Daniel Povey[1,2], Sanjeev Khudanpur[1,2] and Najim Dehak[1]*

[1]Center for Language and Speech Processing

[2]Human Language Technology Center Of Excellence, Johns Hopkins University, Baltimore, MD

**Mon-P-1-2-8, Time: 14:00-16:00**

In this paper, we apply the recently proposed x-vector neural network architecture for the task of age estimation. This architecture maps a variable length utterance into a fixed dimensional embedding which retains the relevant sequence level information. This is achieved by a temporal pooling layer. From the embedding, a series of layers is applied to make predictions. The full network is trained end-to-end in a discriminative fashion. This kind of network is starting to outperform the state-of-the-art i-vector embeddings in tasks like speaker and language recognition. Motivated by this, we investigated the optimum way to train x-vectors for the age estimation task. Despite that a regression objective is typical for this task, we found that optimizing a mixture of classification and regression losses provides better results. We trained our models on the NIST SRE08 dataset and evaluated on SRE10. The proposed approach improved mean absolute error (MAE) by 12% w.r.t the i-vector baseline.

## Improving Gender Identification in Movie Audio Using Cross-Domain Data

*Rajat Hebbar, Krishna Somandepalli and Shrikanth Narayanan*

Signal Analysis and Interpretation Laboratory, Department of Electrical Engineering, University of Southern California, Los Angeles

**Mon-P-1-2-9, Time: 14:00-16:00**

Gender identification from audio is an important task for quantitative gender analysis in multimedia and to improve tasks like speech recognition. Robust gender identification requires speech segmentation that relies on accurate voice activity detection (VAD). These tasks are challenging in movie audio due to diverse and often noisy acoustic conditions. In this work, we acquire VAD labels for movie audio by aligning it with subtitle text and train a recurrent neural network model for VAD. Subsequently, we apply transfer learning to predict gender using feature embeddings obtained from a model pre-trained for large-scale audio classification. In order to account for the diverse acoustic conditions in movie audio, we use audio clips from YouTube labeled for gender. We compare the performance of our proposed method with baseline experiments that were setup to assess the importance of feature embeddings and training data used for gender identification task. For systematic evaluation, we extend an existing benchmark dataset for movie VAD, to include precise gender labels. The VAD system shows comparable results to state-of-the-art in movie domain. The proposed gender identification system outperforms existing baselines, achieving an accuracy of 85% for movie audio. We have made the data and related code publicly available.

## On Learning to Identify Genders from Raw Speech Signal Using CNNs

*Selen Hande Kabil[1,2], Hannah Muckenhirn[1,2] and Mathew Magimai Doss[1]*

[1]Idiap Research Institute, Martigny, CH

[2]École Polytechnique Fédérale de Lausanne, CH

**Mon-P-1-2-10, Time: 14:00-16:00**

Automatic Gender Recognition (AGR) is the task of identifying the gender of a speaker given a speech signal. Standard approaches extract features like fundamental frequency and cepstral features from the speech signal and train a binary classifier. Inspired from recent works in the area of automatic speech recognition (ASR), speaker recognition and presentation attack detection, we present a novel approach where relevant features and classifier are jointly learned from the raw speech signal in end-to-end manner. We propose a convolutional neural networks (CNN) based gender classifier that consists of: (1) convolution layers, which can be interpreted as a feature learning stage and (2) a multilayer perceptron (MLP), which can be interpreted as a classification stage. The system takes raw speech signal as input and outputs gender posterior probabilities. Experimental studies conducted on two datasets, namely AVspoof and ASVspoof 2015, with different architectures show that with simple architectures the proposed approach yields better system than standard acoustic features based approach. Further analysis of the CNNs show that the CNNs learn formant and fundamental frequency information for gender identification.

## Denoising and Raw-waveform Networks for Weakly-Supervised Gender Identification on Noisy Speech

*Jilt Sebastian[1,3,4], Manoj Kumar[2], Pavan Kumar D. S.[3,4], Mathew Magimai Doss[3], Hema Murthy[1] and Shrikanth Narayanan[2]*

**Notes**

[1]Indian Institute of Technology Madras, India
[2]University of Southern California, Los Angeles, USA
[3]Idiap Research Institute, Martigny, CH
[4]École Polytechnique Fédérale de Lausanne, CH

**Mon-P-1-2-11, Time: 14:00-16:00**

This paper presents a raw-waveform neural network and uses it along with a denoising network for clustering in weakly-supervised learning scenarios under extreme noise conditions. Specifically, we consider language independent gender identification on a set of varied noise conditions and signal to noise ratios (SNRs). We formulate the denoising problem as a source separation task and train the system using a discriminative criterion in order to enhance output SNRs. A denoising recurrent neural network (RNN) is first trained on a small subset (roughly one-fifth) of the data for learning a speech-specific mask. The denoised speech signal is then directly fed as input to a raw-waveform convolutional neural network (CNN) trained with denoised speech. We evaluate the standalone performance of denoiser in terms of various signal-to-noise measures and discuss its contribution towards robust gender identification. An absolute improvement of 11.06% and 13.33% is achieved by the combined pipeline over the i-vector SVM baseline system for 0 dB and -5 dB SNR conditions, respectively. We further analyse the information captured by the first CNN layer in both noisy and denoised speech.

### The Effect of Exposure to High Altitude and Heat on Speech Articulatory Coordination

*James Williamson[1], Thomas Quatieri[1], Adam Lammert[1], Katherine Mitchell[2], Katherine Finkelstein[2], Nicole Ekon[2], Caitlin Dillon[2], Robert Kenefick[2] and Kristin Heaton[2]*
[1]MIT Lincoln Laboratory
[2]USARIEM

**Mon-P-1-2-12, Time: 14:00-16:00**

The effects of altitude and heat on speech articulatory coordination following exercise and approximately three hours of exposure are explored. Recordings of read speech and free response speech before and after exercise in moderate altitude, moderate heat and both moderate altitude and heat are analyzed using features that characterize articulatory coordination. It is found that 1) moderate altitude causes small changes and moderate heat negligible changes to articulatory coordination features after brief exposure prior to exercise; 2) moderate altitude and heat produce similar large feature changes following exercise and longer exposure; 3) moderate altitude and heat produce larger feature changes in combination than individually immediately following exercise. Finally, using cross-validation training of a statistical classifiers, the features are sufficient to classify the four experimental conditions with an overall accuracy of 0.50 and to detect the presence of any one of the experimental conditions with an accuracy of 0.90.

## Mon-P-1-3: Deep Learning for Source Separation and Pitch Tracking

Hall 4-6: Poster-3; 14:00-16:00; Monday, 3 September, 2018
Chair: Mathew Magimai Doss

### Permutation Invariant Training of Generative Adversarial Network for Monaural Speech Separation

*Lianwu Chen[1], Meng Yu[1], Yanmin Qian[2], Dan Su[1] and Dong Yu[3]*
[1]Tencent AI Lab, Shenzhen, China
[2]Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
[3]Tencent AI Lab, Bellevue, WA, USA

**Mon-P-1-3-1, Time: 14:00-16:00**

We explore generative adversarial networks (GANs) for speech separation, particularly with permutation invariant training (SSGAN-PIT). Prior work demonstrates that GANs can be implemented for suppressing additive noise in noisy speech waveform and improving perceptual speech quality. In this work, we train GANs for speech separation which enhances multiple speech sources simultaneously with the permutation issue addressed by the utterance level PIT in the training of the generator network. We propose operating GANs on the power spectrum domain instead of waveforms to reduce computation. To better explore time dependencies, recurrent neural networks (RNNs) with long short-term memory (LSTM) are adopted for both generator and discriminator in this study. We evaluated SSGAN-PIT on the WSJ0 two-talker mixed speech separation task and found that SSGAN-PIT outperforms SSGAN without PIT and the neural networks based speech separation with or without PIT. The evaluation confirms the feasibility of the proposed model and training approach for efficient speech separation. The convergence behavior of permutation invariant training and adversarial training are analyzed.

### Deep Extractor Network for Target Speaker Recovery from Single Channel Speech Mixtures

*Jun Wang[1], Jie Chen[1], Dan Su[1], Lianwu Chen[1], Meng Yu[1], Yanmin Qian[3] and Dong Yu[2]*
[1]Tencent AI Lab, Shenzhen, China
[2]Tencent AI Lab, Bellevue WA, USA
[3]Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China

**Mon-P-1-3-2, Time: 14:00-16:00**

Speaker-aware source separation methods are promising workarounds for major difficulties such as arbitrary source permutation and unknown number of sources. However, it remains challenging to achieve satisfying performance provided a very short available target speaker utterance (anchor). Here we present a novel "deep extractor network" which creates an extractor point for the target speaker in a canonical high dimensional embedding space and pulls together the time-frequency bins corresponding to the target speaker. The proposed model is different from prior works that the carnonical embedding space encodes knowledges of both the anchor and the mixture during training phase: first, embeddings for the anchor and mixture speech are separately constructed in a primary embedding space and then combined as an input to feed-forward layers to transform to a carnonical embedding space which we discover more stable than the primary one. Experimental results show that given a very short utterance, the proposed model can efficiently recover high quality target speech from a mixture, which outperforms various baseline models, with 5.2% and 6.6% relative improvements in SDR and PESQ respectively compared with a baseline oracle deep attracor model. Meanwhile, we show it can be generalized well to more than one interfering speaker.

### Joint Localization and Classification of Multiple Sound Sources Using a Multi-task Neural Network

*Weipeng He[1,2], Petr Motlicek[1] and Jean-Marc Odobez[1,2]*
[1]Idiap Research Institute, Switzerland
[2]École Polytechnique Fédérale de Lausanne(EPFL), Switzerland

**Mon-P-1-3-3, Time: 14:00-16:00**

We propose a novel multi-task neural network-based approach for joint sound source localization and speech/non-speech classification in noisy environments. The network takes raw short time Fourier transform as input and outputs the likelihood values for the two tasks, which are used for the simultaneous detection, localization and classification of an unknown number of overlapping sound sources, Tested with real recorded data, our method achieves significantly better performance in terms of speech/non-speech classification and localization of speech sources, compared to method that performs localization and classification separately. In addition, we demonstrate that incorporating the temporal context can further improve the performance.

**Notes**

## Detection of Glottal Closure Instants from Speech Signals: A Convolutional Neural Network Based Method

*Shuai Yang[1], Zhiyong Wu[1,3], Binbin Shen[2] and Helen Meng[1,3]*

[1]Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University

[2]Pachira Information Technology (Beijing) Co., Ltd.

[3]Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong

**Mon-P-1-3-4, Time: 14:00-16:00**

Most conventional methods to detect glottal closure instants (GCI) are based on signal processing technologies and different GCI candidate selection methods. This paper proposes a classification method to detect glottal closure instants from speech waveforms using convolutional neural network (CNN). The procedure is divided into two successive steps. Firstly, a low-pass filtered signal is computed, whose negative peaks are taken as candidates for GCI placement. Secondly, a CNN-based classification model determines for each peak whether it corresponds to a GCI or not. The method is compared with three existing GCI detection algorithms on two publicly available databases. For the proposed method, the detection accuracy in terms of F1-score is 98.23%. Additional experiment indicates that the model can perform better after trained with the speech data from the speakers who are the same as those in the test set.

## Robust TDOA Estimation Based on Time-Frequency Masking and Deep Neural Networks

*Zhong-Qiu Wang[1], Xueliang Zhang[3], DeLiang Wang[1,2]*

[1]Department of Computer Science and Engineering, The Ohio State University, USA

[2]Center for Cognitive and Brain Sciences, The Ohio State University, USA

[3]Department of Computer Science, Inner Mongolia University, China

**Mon-P-1-3-5, Time: 14:00-16:00**

Deep learning based time-frequency (T-F) masking has dramatically advanced monaural speech separation and enhancement. This study investigates its potential for robust time difference of arrival (TDOA) estimation in noisy and reverberant environments. Three novel algorithms are proposed to improve the robustness of conventional cross-correlation-, beamforming- and subspace-based algorithms for speaker localization. The key idea is to leverage the power of deep neural networks (DNN) to accurately identify T-F units that are relatively clean for TDOA estimation. All of the proposed algorithms exhibit strong robustness for TDOA estimation in environments with low input SNR, high reverberation and low direction-to-reverberant energy ratio.

## Waveform to Single Sinusoid Regression to Estimate the F0 Contour from Noisy Speech Using Recurrent Deep Neural Networks

*Akihiro Kato and Tomi Kinnunen*

University of Eastern Finland

**Mon-P-1-3-6, Time: 14:00-16:00**

The fundamental frequency (F0) represents pitch in speech that determines prosodic characteristics of speech and is needed in various tasks for speech analysis and synthesis. Despite decades of research on this topic, F0 estimation at low signal-to-noise ratios (SNRs) in unexpected noise conditions remains difficult. This work proposes a new approach to noise robust F0 estimation using a recurrent neural network (RNN) trained in a supervised manner. Recent studies employ deep neural networks (DNNs) for F0 tracking as a frame-by-frame classification task into quantised frequency states but we propose waveform-to-sinusoid regression instead to achieve both noise robustness and accurate estimation with increased frequency resolution. Experimental results with PTDB-TUG corpus contaminated by additive noise (NOISEX-92) demonstrate that the proposed method improves gross pitch error (GPE) rate and fine pitch error (FPE) by more than 35% at SNRs between -10 dB and +10 dB compared with well-known noise robust F0 tracker, PEFAC. Furthermore, the proposed method also outperforms state-of-the-art DNN-based approaches by more than 15% in terms of both FPE and GPE rate over the preceding SNR range.

## Reducing Interference with Phase Recovery in DNN-based Monaural Singing Voice Separation

*Paul Magron[1], Konstantinos Drossos[1], Stylianos Ioannis Mimilakis[2] and Tuomas Virtanen[1]*

[1]Laboratory of Signal Processing, Tampere University of Technology, Finland

[2]Fraunhofer IDMT, Ilmenau, Germany

**Mon-P-1-3-7, Time: 14:00-16:00**

State-of-the-art methods for monaural singing voice separation consist in estimating the magnitude spectrum of the voice in the short-time Fourier transform (STFT) domain by means of deep neural networks (DNNs). The resulting magnitude estimate is then combined with the mixture's phase to retrieve the complex-valued STFT of the voice, which is further synthesized into a time-domain signal. However, when the sources overlap in time and frequency, the STFT phase of the voice differs from the mixture's phase, which results in interference and artifacts in the estimated signals. In this paper, we investigate on recent phase recovery algorithms that tackle this issue and can further enhance the separation quality. These algorithms exploit phase constraints that originate from a sinusoidal model or from consistency, a property that is a direct consequence of the STFT redundancy. Experiments conducted on real music songs show that those algorithms are efficient for reducing interference in the estimated voice compared to the baseline approach.

## Nebula: F0 Estimation and Voicing Detection by Modeling the Statistical Properties of Feature Extractors

*Kanru Hua*

University of Illinois, U.S.A.

**Mon-P-1-3-8, Time: 14:00-16:00**

A F0 and voicing status estimation algorithm for high quality speech analysis/synthesis is proposed. This problem is approached from a different perspective that models the behavior of feature extractors under noise, instead of directly modeling speech signals. Under time-frequency locality assumptions, the joint distribution of extracted features and target F0 can be characterized by training a bank of Gaussian mixture models (GMM) on artificial data generated from Monte-Carlo simulations. The trained GMMs can then be used to generate a set of conditional distributions on the predicted F0, which are then combined and post-processed by Viterbi algorithm to give a final F0 trajectory. Evaluation on CSTR and CMU Arctic speech databases shows that the proposed method, trained on fully synthetic data, achieves lower gross error rates than state-of-the-art methods.

## Real-time Single-channel Dereverberation and Separation with Time-domain Audio Separation Network

*Yi Luo and Nima Mesgarani*

Department of Electrical Engineering, Columbia University, New York, NY

**Mon-P-1-3-9, Time: 14:00-16:00**

We investigate the recently proposed Time-domain Audio Separation Network (TasNet) in the task of real-time single-channel speech dereverberation. Unlike systems that take time-frequency representation of the audio as input, TasNet learns an adaptive front-end in replacement of the time-frequency representation by a time-domain convolutional non-negative autoencoder. We show that by formulating the dereverberation problem as a denoising problem where the direct path is separated from

**Notes**

the reverberations, a TasNet denoising autoencoder can outperform a deep LSTM baseline on log-power magnitude spectrogram input in both causal and non-causal settings. We further show that adjusting the stride size in the convolutional autoencoder helps both the dereverberation and separation performance.

### Music Source Activity Detection and Separation Using Deep Attractor Network

*Rajath Kumar, Yi Luo and Nima Mesgarani*
Department of Electrical Engineering, Columbia University, New York, NY

**Mon-P-1-3-10, Time: 14:00-16:00**

In music signal processing, singing voice detection and music source separation are widely researched topics. Recent progress in deep neural network based source separation has advanced the state of the performance in the problem of vocal and instrument separation, while the problem of joint source activity detection and separation remains unexplored. In this paper, we propose an approach to perform source activity detection using the high-dimensional embedding generated by Deep Attractor Network (DANet) when trained for music source separation. By defining both tasks together, DANet is able to dynamically estimate the number of outputs depending on the active sources. We propose an Expectation-Maximization (EM) training paradigm for DANet which further improves the separation performance of the original DANet. Experiments show that our network achieves higher source separation and comparable source activity detection against a baseline system.

### Improving Mandarin Tone Recognition Using Convolutional Bidirectional Long Short-Term Memory with Attention

*Longfei Yang, Yanlu Xie and Jinsong Zhang*
Beijing Advanced Innovation Center for Language Resources, Beijing Language and Culture University, Beijing 100083, China

**Mon-P-1-3-11, Time: 14:00-16:00**

Automatic tone recognition is useful for Mandarin spoken language processing. However, the complex F0 variations from the tone co-articulations and the interplay effects among tonality make it rather difficult to perform tone recognition of Chinese continuous speech. This paper explored the application of Bidirectional Long Short-Term Memory (BLSTM), which had the capability of modeling time series, to Mandarin tone recognition to handle the tone variations in continuous speech. In addition, we introduced attention mechanism to guide the model to select the suitable context information. The experimental results showed that the performance of proposed CNN-BLSTM with attention mechanism was the best and it achieved the tone error rate (TER) of 9.30% with a 17.6% relative error reduction from the DNN baseline system with TER of 11.28%. It demonstrated that our proposed model was more effective to handle the complex F0 variations than other models.

## Mon-P-1-4: Acoustic Analysis-Synthesis of Speech Disorders

Hall 4-6: Poster-4; 14:00-16:00; Monday, 3 September, 2018
Chair: Vikramjit Mitra

### Vowel Space as a Tool to Evaluate Articulation Problems

*Rob van Son[1,2], Catherine Middag[3] and Kris Demuynck[3]*
[1]NKI-AVL, Amsterdam;
[2]ACLC, University of Amsterdam, The Netherlands
[3]IDLab, University of Ghent, Belgium

**Mon-P-1-4-1, Time: 14:00-16:00**

Treatment for oral tumors can lead to long term changes in the anatomy and physiology of the vocal tract and result in problems with articulation. There are currently no readily available automatic methods to evaluate changes in articulation. We developed a Praat script which plots and measures vowel space coverage. The script reproduces speaker specific vowel space use and speaking-style dependent vowel reduction in normal speech from a Dutch corpus. Speaker identity and speaking style explain more than 60% of the variance in the measured area of the vowel triangle. In recordings of patients treated for oral tumors, vowel space use before and after treatment is still significantly correlated. Articulation before and after treatment is evaluated in a listening experiment and from a maximal articulation speed task. Linear models can explain 50-75% of variance in perceptual ratings and relative articulation rate from values at previous recordings and vowel space measures.

### Towards a Better Characterization of Parkinsonian Speech: A Multidimensional Acoustic Study

*Veronique Delvaux[1,2], Kathy Huet[2], Myriam Piccaluga[2], Sophie van Malderen[2] and Bernard Harmegnies[2]*
[1]FNRS, Belgium
[2]IRSTL, UMONS, Belgium

**Mon-P-1-4-2, Time: 14:00-16:00**

This paper reports on a first attempt at adopting a new perspective in characterizing speech disorders in Parkinson's Disease (PD) based on individual patient profiles. Acoustic data were collected on 13 Belgian French speakers with PD, 6 male, 7 female, aged 45-81 and 50 healthy controls (HC) using the "MonPaGe" protocol (Fougeron et al., LREC18). In this protocol, various kinds of linguistic material are recorded in different speech conditions, in order to assess multiple speech dimensions for each speaker. First, we compared a variety of voice and speech parameters across groups (HC vs. PD patients). Second, we examined individual profiles of PD patients. Results showed that as a group PD participants most systematically differed from HC in terms of speech tempo and rythm. Moreover, the analysis of individual profiles revealed that other parameters, related to pneumophonatory control and linguistic prosody, were valuable to describe the speech specificities of several PD patients.

### Self-similarity Matrix Based Intelligibility Assessment of Cleft Lip and Palate Speech

*Sishir Kalita[1], S R Mahadeva Prasanna[1,2] and Samarendra Dandapat[1]*
[1]Indian Institute of Technology Guwahati, Guwahati-781039, India
[2]Indian Institute of Technology Dharwad, Dharwad-580011, India

**Mon-P-1-4-3, Time: 14:00-16:00**

This work presents a comparison based framework by exploiting the self-similarity matrices matching technique to estimate the speech intelligibility of cleft lip and palate (CLP) children. Self-similarity matrix (SSM) of a feature sequence is a square matrix, which encodes the acoustic-phonetic composition of the underlying speech signal. Deviations in the acoustic characteristics of underlying sound units due to the degradation of intelligibility will deviate the CLP speech's SSM structure from that of normal. This degree of deviations in CLP speech's SSM from the corresponding normal speech's SSM may provide information about the severity profile of speech intelligibility. The degree of deviations is quantified using the structural similarity (SSIM) index, which is considered as the representative of objective intelligibility score. The proposed method is evaluated using two parameterizations of speech signals: Mel-frequency cepstral coefficients and Gaussian posteriorgrams and compared with dynamic time warping (DTW) based intelligibility assessment method. The proposed SSM based method shows the better correlation with the perceptual ratings of intelligibility when compared to the DTW based method.

### Pitch-Adaptive Front-end Feature for Hypernasality Detection

*Akhilesh Kumar Dubey[1], S R Mahadeva Prasanna[1,2] and S Dandapat[1]*

**Notes**

[1]Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati, India
[2]Department of Electrical Engineering, Indian Institute of Technology Dharwad, Dharwad, India

**Mon-P-1-4-4, Time: 14:00-16:00**

Hypernasality in cleft palate (CP) children is due to the velopharyngeal insufficiency. The vowels get nasalized in hypernasal speech and the nasality evidence are mainly present in low-frequency region around the first formant (F1) of vowels. The detection of hypernasality using Mel-frequency cepstral coefficient (MFCC) feature may get affected because the feature might not be able to capture the nasality evidence present in the low-frequency region. This is due to the fact that the MFCC feature extracted from high pitched children speech contains the pitch harmonics effect of magnitude spectrum. The pitch harmonics effect results in high variance for the higher dimensions of MFCC coefficients. This problem may increase due to high perturbation in pitch of CP speech. So in this work, a pitch-adaptive MFCC feature is used for hypernasality detection. The feature is derived from the cepstral smooth spectrum instead of magnitude spectrum. A pitch-adaptive low time liftering is done to smooth out the pitch harmonics. This feature when used for the detection of hypernasality using support vector machine (SVM) gives an accuracy of 83.45%, 88.04% and 85.58% for /a/, /i/ and /u/ vowels respectively, which is better than the accuracy of MFCC feature.

### Detection of Amyotrophic Lateral Sclerosis (ALS) via Acoustic Analysis

*Raquel Norel[1], Mary Pietrowicz[1], Carla Agurto[1], Shay Rishoni[2] and Guillermo Cecchi[1]*
[1]Computational Biology Center, T.J. Watson IBM Research Laboratory, New York, USA.
[2]Prize4Life Israel, Ramat HaSharon, Israel

**Mon-P-1-4-5, Time: 14:00-16:00**

ALS is a fatal neurodegenerative disease with no cure. Experts typically measure disease progression via the ALSFRS-R score, which includes measurements of various abilities known to decline. We propose instead the use of speech analysis as a proxy for ALS progression. This technique enables 1) frequent non-invasive, inexpensive, longitudinal analysis, 2) analysis of data recorded in the wild and 3) creation of an extensive ALS databank for future analysis. Patients and trained medical professionals need not be co-located, enabling more frequent monitoring of more patients from the convenience of their own homes. The goals of this study are the identification of acoustic speech features in naturalistic contexts which characterize disease progression and development of machine models which can recognize the presence and severity of the disease. We evaluated subjects from the Prize4Life Israel dataset, using a variety of frequency, spectral and voice quality features. The dataset was generated using the ALS Mobile Analyzer, a cell-phone app that collects data regarding disease progress using a self-reported ALSFRS-R questionnaire and several active tasks that measure speech and motor skills. Classification via leave-five-subjects-out cross-validation resulted in an accuracy rate of 79% (61% chance) for males and 83% (52% chance) for females.

### Detection of Glottal Activity Errors in Production of Stop Consonants in Children with Cleft Lip and Palate

*Vikram C M[1], S R Mahadeva Prasanna[1,2], Ajish K Abraham[3], Pushpavathi M[3] and Girish K S[3]*
[1]Indian Institute of Technology Guwahati, Guwahati, India
[2]Indian Institute of Technology Dharwad, Dharwad, India
[3]All India Institute of Speech and Hearing, Mysuru, India

**Mon-P-1-4-6, Time: 14:00-16:00**

Individuals with cleft lip and palate (CLP) alter the glottal activity characteristics during the production of stop consonants. The presence/absence of glottal vibrations during the production of unvoiced/voiced stops is referred as glottal activity error (GAE). In this work, acoustic-phonetic and production based knowledge of stop consonants are exploited to propose an algorithm for the automatic detection of GAE. The algorithm uses zero frequency filtered and band-pass (500-4000 Hz) filtered speech signals to identify the syllable nuclei positions, followed by the detection of glottal activity characteristics of consonant present within the syllable. Based on the identified glottal activity characteristics of consonant and a priori voicing information of target stop consonant, the presence or absence of GAE is detected. The algorithm is evaluated over the database containing the responses of normal children and children with repaired CLP for the target consonant-vowel-consonant-vowel words with stop consonants.

## Mon-O-2-1: ASR Systems and Technologies

Hall 1; 16:30-18:30; Monday, 3 September, 2018
Chairs: Mittul Singh and Ariya Rastrow

### Cold Fusion: Training Seq2Seq Models Together with Language Models

*Anuroop Sriram, Heewoo Jun, Sanjeev Satheesh and Adam Coates*
Baidu Research, Sunnyvale, USA

**Mon-O-2-1-1, Time: 16:30-16:50**

Sequence-to-sequence (Seq2Seq) models with attention have excelled at tasks which involve generating natural language sentences such as machine translation, image captioning and speech recognition. Performance has further been improved by leveraging unlabeled data, often in the form of a language model. In this work, we present the Cold Fusion method, which leverages a pre-trained language model during training and show its effectiveness on the speech recognition task. We show that Seq2Seq models with Cold Fusion are able to better utilize language information enjoying i) faster convergence and better generalization and ii) almost complete transfer to a new domain while using less than 10% of the labeled training data.

### Investigation on Estimation of Sentence Probability by Combining Forward, Backward and Bi-directional LSTM-RNNs

*Kazuki Irie, Zhihong Lei, Liuhui Deng, Ralf Schlüter and Hermann Ney*
Human Language Technology and Pattern Recognition Group, Computer Science Department, RWTH Aachen University, D-52056 Aachen, Germany

**Mon-O-2-1-2, Time: 16:50-17:10**

A combination of forward and backward long short-term memory (LSTM) recurrent neural network (RNN) language models is a popular model combination approach to improve the estimation of the sequence probability in the second pass N-best list rescoring in automatic speech recognition (ASR). In this work, we further push such an idea by proposing a combination of three models: a forward LSTM language model, a backward LSTM language model and a bi-directional LSTM based gap completion model. We derive such a combination method from a forward backward decomposition of the sequence probability. We carry out experiments on the Switchboard speech recognition task. While we empirically find that such a combination gives slight improvements in perplexity over the combination of forward and backward models, we finally show that a combination of the same number of forward models gives the best perplexity and word error rate (WER) overall.

### Subword and Crossword Units for CTC Acoustic Models

*Thomas Zenkel[1], Ramon Sanabria[2], Florian Metze[2] and Alex Waibel[1,2]*
[1]Karlsruhe Institute of Technology, Karlsruhe, Germany

**Notes**

[2]Carnegie Mellon University, Pittsburgh, PA, U.S.A.

**Mon-O-2-1-3, Time: 17:10-17:30**

This paper proposes a novel approach to create a unit set for CTC-based speech recognition systems. By using Byte-Pair Encoding we learn a unit set of an arbitrary size on a given training text. In contrast to using characters or words as units this allows us to find a good trade-off between the size of our unit set and the available training data. We investigate both Crossword units, that may span multiple word and Subword units. By evaluating these unit sets with decodings methods using a separate language model we are able to show improvements over a purely character-based unit set.

### Neural Error Corrective Language Models for Automatic Speech Recognition

*Tomohiro Tanaka, Ryo Masumura, Hirokazu Masataki and Yushi Aono*

NTT Media Intelligence Laboratories, NTT Corporation

**Mon-O-2-1-4, Time: 17:30-17:50**

We present novel neural network based language models that can correct automatic speech recognition (ASR) errors by using speech recognizer output as a context. These models, called neural error corrective language models (NECLMs), utilizes ASR hypotheses of a target utterance as a context for estimating the generative probability of words. NECLMs are expressed as conditional generative models composed of an encoder network and a decoder network. In the models, the encoder network constructs context vectors from N-best lists and ASR confidence scores generated in a speech recognizer. The decoder network rescores recognition hypotheses by computing a generative probability of words using the context vectors so as to correct ASR errors. We evaluate the proposed models in Japanese lecture ASR tasks. Experimental results show that NECLM achieve better ASR performance than a state-of-the-art ASR system that incorporate a convolutional neural network acoustic model and a long short-term memory recurrent neural network language model.

### Entity-Aware Language Model as an Unsupervised Reranker

*Mohammad Sadegh Rasooli[1] and Sarangarajan Parthasarathy[2]*

[1]Department of Computer Science, Columbia University, New York, NY, 10027

[2]Artificial Intelligence and Research, Microsoft Corporation, USA

**Mon-O-2-1-5, Time: 17:50-18:10**

In language modeling, it is difficult to incorporate entity relationships from a knowledge-base. One solution is to use a reranker trained with global features, in which global features are derived from n-best lists. However, training such a reranker requires manually annotated n-best lists, which is expensive to obtain. We propose a method based on the contrastive estimation method that alleviates the need for such data. Experiments in the music domain demonstrate that global features, as well as features extracted from an external knowledge-base, can be incorporated into our reranker. Our final model, a simple ensemble of a language model and reranker, achieves a 0.44% absolute word error rate improvement over an LSTM language model on the blind test data.

### Character-level Language Modeling with Gated Hierarchical Recurrent Neural Networks

*Iksoo Choi, Jinhwan Park and Wonyong Sung*

Department of Electrical and Computer Engineering, Seoul National University, Gwanak-ro, Gwanak-gu, Seoul, 08826 Korea

**Mon-O-2-1-6, Time: 18:10-18:30**

Recurrent neural network (RNN)-based language models are widely used for speech recognition and translation applications. We propose a gated hierarchical recurrent neural network (GHRNN) and apply it to the character-level language modeling. GHRNN consists of multiple RNN units that operate with different time scales and the frequency of operation at each unit is controlled by the learned gates from training data. In our model, GHRNN learns the hierarchical structure of character, sub-word and word. Timing gates are included in the hierarchical connections to control the operating frequency of these units. The performance was measured for Penn Treebank and Wikitext-2 datasets. Experimental results showed lower bit per character (BPC) when compared to simply layered or skip-connected RNN models. Also, when a continuous cache model is added, the BPC of 1.192 is recorded, which is comparable to the state of the art result.

## Mon-O-2-2: Deception, Personality and Culture Attribute

Hall 2; 16:30-18:30; Monday, 3 September, 2018

Chairs: Ailbhe Ni Chasaide and Zixing Zhang

### Acoustic-Prosodic Indicators of Deception and Trust in Interview Dialogues

*Sarah Ita Levitan, Angel Maredia and Julia Hirschberg*

Department of Computer Science, Columbia University, USA

**Mon-O-2-2-1, Time: 16:30-16:50**

We analyze a set of acoustic-prosodic features in both truthful and deceptive responses to interview questions, identifying differences between truthful and deceptive speech. We also study the perception of deception, identifying acoustic-prosodic characteristics of speech that is perceived as truthful or deceptive by interviewers. In addition to studying differences across all speakers, we identify individual variations in deception production and perception across gender and native language. We conduct machine learning classification experiments aimed at distinguishing between truthful and deceptive speech, using acoustic-prosodic features. We also explore methods of leveraging individual traits for deception classification. Our results show that acoustic-prosodic features are highly effective at classifying deceptive speech. Our best classifier achieved an F1-score of 72.77, well above both the random baseline and above human performance at this task. This work advances our understanding of deception production and perception and has implications for automatic deception detection and the development of synthesized speech that is trustworthy.

### Deep Personality Recognition for Deception Detection

*Guozhen An[1,2], Sarah Ita Levitan[4], Julia Hirschberg[4] and Rivka Levitan[1,3]*

[1]Department of Computer Science, CUNY Graduate Center, USA

[2]Department of Mathematics and Computer Science, York College (CUNY), USA

[3]Department of Computer and Information Science, Brooklyn College (CUNY), USA

[4]Department of Computer Science, Columbia University, USA

**Mon-O-2-2-2, Time: 16:50-17:10**

Researchers in both psychology and computer science have suggested that modeling individual differences may improve the performance of automatic deception detection systems. In this study, we fuse a personality classifier with a deception classifier and evaluate various ways to combine the two tasks, either as a single network with shared layers, or by feeding the results of the personality classifier into the deception classifier. We show that including personality recognition improves the performance of deception detection on the Columbia X-Cultural Deception (CXD) corpus by more than 6% relative, achieving new state-of-the-art results on classification of phrase-like units in this corpus.

### Cross-cultural (A)symmetries in Audio-visual Attitude Perception

*Hansjörg Mixdorff[1], Albert Rilliard[2], Tan Lee[3], Matthew K. H. Ma[3] and Angelika Hönemann[1,4]*

[1]Department of Computer Science and Media, Beuth University

**Notes**

Berlin, Germany
[2]LIMSI, CNRS, Université Paris-Saclay, France & Federal University of Rio de Janeiro, Brazil
[3]The Chinese University of Hong Kong, Hong Kong
[4]Faculty of Linguistics & Literary Studies, University of Bielefeld, Germany

**Mon-O-2-2-3, Time: 17:10-17:30**

This paper evaluates results from a cross-cultural and cross-language experiment series employing short audio-visual utterances produced with varying attitudinal expressions. German and Cantonese-speaking participants freely labeled such utterances in the two languages and assigned to each stimulus a verbal label. Based on the results of the four experiments we were able to establish to what degree the attitudinal frames of reference of the two groups overlap and how they differ. Verbal labels were assessed regarding their emotional content in terms of valence, activation and dominance and for the linguistic opposition between assertive and interrogative speech act and hence permit to abstract from the language of the rater and ultimately even abstract from the attitudinal categories used when eliciting the stimuli. Instead we regard each utterance as a data-point in the emotional space. We found that the judgments of the two rater groups agree well with respect to the valence of attitudinal expressions and diverge most as to the perceived activation of the stimulus presenter. Cantonese speaking participants seem to mirror Germans' ratings of German stimuli better than vice versa, which suggests an interesting asymmetry of attitudinal perception. As for the modality of presentation, the audio channel primarily transmits linguistically relevant information regarding the opposition of assertion and interrogation while the visual information signals the emotional content.

### An Active Feature Transformation Method for Attitude Recognition of Video Bloggers

*Fasih Haider[1], Fahim A. Salim[2], Owen Conlan[2] and Saturnino Luz[1]*
[1]IPHSI, University of Edinburgh, UK
[2]ADAPT Centre, Trinity College Dublin, Ireland

**Mon-O-2-2-4, Time: 17:30-17:50**

Video blogging is a form of unidirectional communication where a video blogger expresses his/her opinion about different issues. The success of a video blog is measured using metrics like the number of views and comments by online viewers. Researchers have highlighted the importance of non-verbal behaviours (e.g. attitudes) in the context of video blogging and showed that it correlates with the level of attention (number of views) gained by a video blog. Therefore, an automatic attitude recognition system can help potential video bloggers to train their attitudes. It can also be useful in developing video blogs summarization and search tools. This study proposes a novel Active Feature Transformation (AFT) method for automatic recognition of attitudes (a form of non-verbal behaviour) in video blogs. The proposed method transforms the Mel-frequency Cepstral Coefficient (MFCC) features for the classification task. The Principal Component Analysis (PCA) transformation is also used for comparison. Our results show that AFT outperforms PCA in terms of accuracy and dimensionality reduction for attitude recognition using linear discrimination analysis, 1-nearest neighbour and decision tree classifiers.

### Automatic Assessment of Individual Culture Attribute of Power Distance Using a Social Context-Enhanced Prosodic Network Representation

*Fu-Sheng Tsai[1,3], Hao-Chun Yang[1,3], Wei-Wen Chang[2] and Chi-Chun Lee[1,3]*
[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan
[2]International Human Resource Development, National Taiwan Normal University, Taiwan
[3]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan

**Mon-O-2-2-5, Time: 17:50-18:10**

Culture is a collective social norm of human societies that often influences a person's values, thoughts and social behaviors during interactions at an individual level. In this work, we present a computational analysis toward automatic assessing an individual's culture attribute of power distance, i.e., a measure of his/her belief about status, authority and power in organizations, by modeling their expressive prosodic structures during social encounters with people of different power status. Specifically, we propose a center-loss embedded network architecture to jointly consider the effect of social interaction contexts on individuals' prosodic manifestations in order to learn an enhanced representation for power distance recognition. Our proposed prosodic network achieves an overall accuracy of 78.6% in binary classification task of recognizing high versus low power distance. Our experiment demonstrates an improved discriminability (17.6% absolute improvement) over prosodic neural network without social context enhancement. Further visualization reveals that the diversity in the prosodic manifestation for individuals with low power distance seems to be higher than those of high power distance.

### Analysis and Detection of Phonation Modes in Singing Voice using Excitation Source Features and Single Frequency Filtering Cepstral Coefficients (SFFCC)

*Sudarsana Reddy Kadiri and Bayya Yegnanarayana*
Speech Processing Laboratory,
International Institute of Information Technology, Hyderabad, India

**Mon-O-2-2-6, Time: 18:10-18:30**

In this study, classification of the phonation modes in singing voice is carried out. Phonation modes in singing voice can be described using four categories: breathy, neutral, flow and pressed phonations. Previous studies on the classification of phonation modes use voice quality features derived from inverse filtering which lack in accuracy. This is due to difficulty in deriving the excitation source features using inverse filtering from singing voice. We propose to use the excitation source features that are derived directly from the signal. It is known that, the characteristics of the excitation source vary in different phonation types due to the vibration of the vocal folds together with the respiratory effort (lungs effort). In the present study, we are exploring excitation source features derived from the modified zero frequency filtering (ZFF) method. Apart from excitation source features, we also explore cepstral coefficients derived from single frequency filtering (SFF) method for the analysis and classification of phonation types in singing voice.

## Mon-O-2-3: Automatic Detection and Recognition of Voice and Speech Disorders

MR G.01-G.02; 16:30-18:30; Monday, 3 September, 2018
Chairs: Rob van Son and Ajish Abraham

### A Deep Learning Method for Pathological Voice Detection Using Convolutional Deep Belief Networks

*Huiyi Wu[1], John Soraghan[1], Anja Lowit[2] and Gaetano Di-Caterina[1]*
[1]Centre for Signal and Image Processing, Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, Scotland
[2]Speech and Language Therapy, School of Psychological Sciences and Health, University of Strathclyde, Glasgow, Scotland

**Mon-O-2-3-1, Time: 16:30-16:50**

Automatically detecting pathological voice disorders such as vocal cord paralysis or Reinke's edema is an important medical classification problem. While deep learning techniques have achieved significant progress in the speech recognition field, there has been less research work in the area of pathological voice disorders detection. A novel system for pathological voice detection using Convolutional Neural Network (CNN) as the basic

**Notes**

architecture is presented in this work. The novel system uses spectrograms of normal and pathological speech recordings as the input to the network. Initially Convolutional Deep Belief Network (CDBN) are used to pre-train the weights of CNN system. This acts as a generative model to explore the structure of the input data using statistical methods. Then a CNN is trained using supervised back-propagation learning algorithm to fine tune the weights. Results show that a small amount of data can be used to achieve good results in classification with this deep learning approach. A performance analysis of the novel method is provided using real data from the Saarbrucken Voice database.

## Dysarthric Speech Recognition Using Time-delay Neural Network Based Denoising Autoencoder

*Chitralekha Bhat, Biswajit Das, Bhavik Vachhani and Sunil Kumar Kopparapu*

TCS Research and Innovation, Mumbai, India

**Mon-O-2-3-2, Time: 16:50-17:10**

Dysarthria is a manisfestation of the disruption in the neuro-muscular physiology resulting in uneven, slow, slurred, harsh or quiet speech. Dysarthric speech poses serious challenges to automatic speech recognition, considering this speech is difficult to decipher for both humans and machines. The objective of this work is to enhance dysarthric speech features to match that of healthy control speech. We use a Time-Delay Neural Network based Denoising Autoencoder (TDNN-DAE) to enhance the dysarthric speech features. The dysarthric speech thus enhanced is recognized using a DNN-HMM based Automatic Speech Recognition (ASR) engine. This methodology was evaluated for speaker-independent (SI) and speaker-adapted (SA) systems. Absolute improvements of 13% and 3% was observed in the ASR performance for SI and SA systems respectively as compared with unenhanced dysarthric speech recognition.

## A Multitask Learning Approach to Assess the Dysarthria Severity in Patients with Parkinson's Disease

*Juan Camilo Vásquez Correa[1,2], Tomas Arias[1,2,3], Juan Rafael Orozco-Arroyave[1,2] and Elmar Nöth[2]*

[1]Faculty of Engineering, University of Antioquia UdeA, Calle 70 No. 52-21, Medelín, Colombia

[2]Pattern Recognition Lab, Friedrich-Alexander-University Erlangen-Nüremberg, Germany

[3]Ludwig-Maximilians-University, Munich, Germany

**Mon-O-2-3-3, Time: 17:10-17:30**

Parkinson's disease is a neurodegenerative disorder characterized by a variety of motor and non-motor symptoms. Particularly, several speech impairments appear in the initial stages of the disease, which affect aspects related to respiration and the movement of muscles and limbs in the vocal tract. Most of the studies in the literature aim to assess only one specific task from the patients, such as the classification of patients vs. healthy speakers, or the assessment of the neurological state of the patients. This study proposes a multitask learning approach based on convolutional neural networks to assess at the same time several speech deficits of the patients. A total of eleven speech aspects are considered, including difficulties of the patients to move articulators such as lips, palate, tongue and larynx. According to the results, the proposed approach improves the generalization of the convolutional network, producing more representative feature maps to assess the different speech symptoms of the patients. The multitask learning scheme improves in of up to 4% the average accuracy relative to single networks trained to assess each individual speech aspect.

## The Use of Machine Learning and Phonetic Endophenotypes to Discover Genetic Variants Associated with Speech Sound Disorder

*Jason Lilley, Erin Crowgey and H Timothy Bunnell*

Nemours Biomedical Research, Wilmington, DE, USA

**Mon-O-2-3-4, Time: 17:30-17:50**

Thirty-four (34) children with reported speech sound disorders (SSD) were recruited for a prior study, as well as 31 of their siblings, many of whom also showed SSD. Using data-clustering techniques, we assigned each child to one or more endophenotypes defined by the number and type of speech errors made on the GFTA-2. The genetic samples of 53 of the participants underwent whole exome sequencing. Variant alleles were detected, filtered and annotated from the sequences and the data were filtered using quality checks, annotations and phenotypes. We then used Random Forest classification to search for associations between variants and endophenotypes. In this preliminary report, we highlight one promising association with a common variant of COMT, a dopamine metabolizer in the brain.

## Whistle-blowing ASRs: Evaluating the Need for More Inclusive Speech Recognition Systems

*Meredith Moore, Hemanth Venkateswara and Sethuraman Panchanathan*

Arizona State University's Center for Cognitive Ubiquitous Computing

**Mon-O-2-3-5, Time: 17:50-18:10**

Speech is a complex process that can break in many different ways and lead to a variety of voice disorders. Dysarthria is a voice disorder where individuals are unable to control one or more of the aspects of speech—the articulation, breathing, voicing, or prosody—leading to less intelligible speech. In this paper, we evaluate the accuracy of state-of-the-art automatic speech recognition systems (ASRs) on two dysarthric speech datasets and compare the results to ASR performance on control speech. The limits of ASR performance using different voices have not been explored since the field has shifted from generative models of speech recognition to deep neural network architectures. To test how far the field has come in recognizing disordered speech, we test two different ASR systems: (1) Carnegie Mellon University's Sphinx Open Source Recognition and (2) Google®Speech Recognition. While (1) uses generative models of speech recognition, (2) uses deep neural networks. As expected, while (2) achieved lower word error rates (WER) on dysarthric speech than (1), control speech had a WER 59% lower than dysarthric speech. Future studies should be focused not only on making ASRs robust to environmental noise, but also more robust to different voices.

## Data Augmentation Using Healthy Speech for Dysarthric Speech Recognition

*Bhavik Vachhani, Chitralekha Bhat and Sunil Kumar Kopparapu*

TCS Research and Innovation, Mumbai, India

**Mon-O-2-3-6, Time: 18:10-18:30**

Dysarthria refers to a speech disorder caused by trauma to the brain areas concerned with motor aspects of speech giving rise to effortful, slow, slurred or prosodically abnormal speech. Traditional Automatic Speech Recognizers (ASR) perform poorly on dysarthric speech recognition tasks, owing mostly to insufficient dysarthric speech data. Speaker related challenges complicates data collection process for dysarthric speech. In this paper, we explore data augmentation using temporal and speed modifications of healthy speech to simulate dysarthric speech. DNN-HMM based Automatic Speech Recognition (ASR) and Random Forest based classification were used for evaluation of the proposed method. Dysarthric speech generated synthetically is classified for severity using a Random Forest classifier that is trained on actual dysarthric speech. ASR trained on healthy speech augmented with simulated dysarthric speech is evaluated for dysarthric speech recognition. All evaluations were carried out using Universal Access dysarthric speech corpus. An absolute improvement of 4.24% and 2% was achieved using tempo based and speed based data augmentation respectively as compared to ASR performance using healthy speech alone for training.

**Notes**

# Mon-O-2-4: Voice Conversion

MR G.03-G.04; 16:30-18:30; Monday, 3 September, 2018

Chairs: Tomoki Toda and Yannis Stylianou

## Improving Sparse Representations in Exemplar-Based Voice Conversion with a Phoneme-Selective Objective Function

*Shaojin Ding, Guanlong Zhao, Christopher Liberatore and Ricardo Gutierrez-Osuna*

Department of Computer Science and Engineering, Texas A&M University, USA

**Mon-O-2-4-1, Time: 16:30-16:50**

The acoustic quality of exemplar-based voice conversion (VC) degrades whenever the phoneme labels of the selected exemplars do not match the phonetic content of the frame being represented. To address this issue, we propose a Phoneme-Selective Objective Function (PSOF) that promotes a sparse representation of each speech frame with exemplars from a few phoneme classes. Namely, PSOF enforces group sparsity on the representation, where each group corresponds to a phoneme class. The sparse representation for exemplars within a phoneme class tends to activate or suppress simultaneously using the proposed objective function. We conducted two sets of experiments on the ARCTIC corpus to evaluate the proposed method. First, we evaluated the ability of PSOF to reduce phoneme mismatches. Then, we assessed its performance on a VC task and compared it against three baseline methods from previous studies. Results from objective measurements and subjective listening tests show that the proposed method effectively reduces phoneme mismatches and significantly improves VC acoustic quality while retaining the voice identity of the target speaker.

## Learning Structured Dictionaries for Exemplar-based Voice Conversion

*Shaojin Ding, Christopher Liberatore and Ricardo Gutierrez-Osuna*

Department of Computer Science and Engineering

Texas A&M University, United States

**Mon-O-2-4-2, Time: 16:50-17:10**

Incorporating phonetic information has been shown to improve the performance of exemplar-based voice conversion. A standard approach is to build a phonetically structured dictionary, where exemplars are categorized into sub-dictionaries according to their phoneme labels. However, acquiring phoneme labels can be expensive and the phoneme labels can have inaccuracies. The latter problem becomes more salient when the speakers are non-native speakers. This paper presents an iterative dictionary-learning algorithm that avoids the need for phoneme labels and instead learns the structured dictionaries in an unsupervised fashion. At each iteration, two steps are alternatively performed: cluster update and dictionary update. In the cluster update step, each training frame is assigned to a cluster whose sub-dictionary represents it with the lowest residual. In the dictionary update step, the sub-dictionary for a cluster is updated using all the speech frames in the cluster. We evaluate the proposed algorithm through objective and subjective experiments on a new corpus of non-native English speech. Compared to previous studies, the proposed algorithm improves the acoustic quality of voice-converted speech while retaining the target speaker's identity.

## Exemplar-Based Spectral Detail Compensation for Voice Conversion

*Yu-Huai Peng[1], Hsin-Te Hwang[1], Yichiao Wu[2], Yu Tsao[3] and Hsin-Min Wang[1]*

[1]Institute of Information Science, Academia Sinica, Taipei, Taiwan

[2]Graduate School of Informatics, Nagoya University, Japan

[3]Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan

**Mon-O-2-4-3, Time: 17:10-17:30**

Most voice conversion (VC) systems are established under the vocoder-based VC framework. When performing spectral conversion (SC) under this framework, the low-dimensional spectral features, such as mel-ceptral coefficients (MCCs), are often adopted to represent the high-dimensional spectral envelopes. The joint density Gaussian mixture model (GMM)-based SC method with the STRAIGHT vocoder is a well-known representative. Although it is reasonably effective, the loss of spectral details in the converted spectral envelopes inevitably deteriorates speech quality and similarity. To overcome this problem, we propose a novel exemplar-based spectral detail compensation method for VC. In the offline stage, the paired dictionaries of source spectral envelopes and target spectral details are constructed. In the online stage, the locally linear embedding (LLE) algorithm is applied to predict the target spectral details from the source spectral envelopes and then, the predicted spectral details are used to compensate the converted spectral envelopes obtained by a baseline GMM-based SC method with the STRAIGHT vocoder. Experimental results show that the proposed method can notably improve the baseline system in terms of objective and subjective tests.

## Whispered Speech to Neutral Speech Conversion Using Bidirectional LSTMs

*G. Nisha Meenakshi and Prasanta Kumar Ghosh*

Electrical Engineering, Indian Institute Science

Bangalore-560012, Karnataka, India

**Mon-O-2-4-4, Time: 17:30-17:50**

We propose a bidirectional long short-term memory (BLSTM) based whispered speech to neutral speech conversion system that employs the STRAIGHT speech synthesizer. We use a BLSTM to map the spectral features of whispered speech to those of neutral speech. Three other BLSTMs are employed to predict the pitch, periodicity levels and the voiced/unvoiced phoneme decisions from the spectral features of whispered speech. We use objective measures to quantify the quality of the predicted spectral features and excitation parameters, using data recorded from six subjects, in a four fold setup. We find that the temporal smoothness of the spectral features predicted using the proposed BLSTM based system is statistically more compared to that predicted using deep neural network based baseline schemes. We also observe that while the performance of the proposed system is comparable to the baseline scheme for pitch prediction, it is superior in terms of classifying voicing decisions and predicting periodicity levels. From subjective evaluation via listening test, we find that the proposed method is chosen as the best performing scheme 26.61% (absolute) more often than the best baseline scheme. This reveals that the proposed method yields a more natural sounding neutral speech from whispered speech.

## Voice Conversion Across Arbitrary Speakers Based on a Single Target-Speaker Utterance

*Songxiang Liu[1], Jinghua Zhong[1], Lifa Sun[1,2], Xixin Wu[1], Xunying Liu[1] and Helen Meng[1]*

[1]Human-Computer Communications Laboratory, Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China

[2]SpeechX Limited, Shenzhen, China

**Mon-O-2-4-5, Time: 17:50-18:10**

Developing a voice conversion (VC) system for a particular speaker typically requires considerable data from both the source and target speakers. This paper aims to effectuate VC across arbitrary speakers, which we call any-to-any VC, with only a single target-speaker utterance. Two systems are studied: (1) the i-vector-based VC (IVC) system and (2) the speaker-encoder-based VC (SEVC) system. Phonetic PosteriorGrams are adopted as speaker-independent linguistic features extracted from speech samples. Both systems train a multi-speaker deep bidirectional long-short term memory (DBLSTM) VC model, taking in additional inputs that encode speaker identities, in order to generate the outputs. In the IVC system, the speaker identity of a new target speaker is represented by i-vectors. In the SEVC system, the speaker identity is represented by speaker embedding predicted from a separately trained model. Experiments verify

**Notes**

the effectiveness of both systems in achieving VC based only on a single target-speaker utterance. Furthermore, the IVC approach is superior to SEVC, in terms of the quality of the converted speech and its similarity to the utterance produced by the genuine target speaker.

### Multi-target Voice Conversion without Parallel Data by Adversarially Learning Disentangled Audio Representations

*Ju-chieh Chou, Cheng-chieh Yeh, Hung-yi Lee and Lin-shan Lee*
College of Electrical Engineering and Computor Science, National Taiwan University

**Mon-O-2-4-6, Time: 18:10-18:30**

Recently, cycle-consistent adversarial network (Cycle-GAN) has been successfully applied to voice conversion to a different speaker without parallel data, although in those approaches an individual model is needed for each target speaker. In this paper, we propose an adversarial learning framework for voice conversion, with which a single model can be trained to convert the voice to many different speakers, all without parallel data, by separating the speaker characteristics from the linguistic content in speech signals. An autoencoder is first trained to extract speaker-independent latent representations and speaker embedding separately using another auxiliary speaker classifier to regularize the latent representation. The decoder then takes the speaker-independent latent representation and the target speaker embedding as the input to generate the voice of the target speaker with the linguistic content of the source utterance. The quality of decoder output is further improved by patching with the residual signal produced by another pair of generator and discriminator. A target speaker set size of 20 was tested in the preliminary experiments and very good voice quality was obtained. Conventional voice conversion metrics are reported. We also show that the speaker information has been properly reduced from the latent representations.

## Mon-SS-2-1: Special Session: The INTERSPEECH 2018 Computational Paralinguistics ChallengE (ComParE): Atypical & Self-Assessed Affect, Crying & Heart Beats 2

MR 1.01-1.02; 16:30-18:30; Monday, 3 September, 2018
Chairs: Björn Schuller, Stefan Steidl, Anton Batliner, Peter B. Marschik, Harald Baumeister and Fengquan Dong

### Self-assessed Affect Sub-Challenge

*Björn Schuller[1,2,3], Stefan Steidl[4], Anton Batliner[2,4], Peter B. Marschik[5,6,7], Harald Baumeister[8], Fengquan Dong[9], Simone Hantke[2,10], Florian B. Pokorny[5,10], Eva-Maria Rathner[8], Katrin D. Bartl-Pokorny[5], Christa Einspieler[5], Dajie Zhang[5,6], Alice Baird[2], Shahin Amiriparian[2,10], Kun Qian[2,10], Zhao Ren[2], Maximilian Schmitt[2], Panagiotis Tzirakis[1] and Stefanos Zafeiriou[1,11]*
[1]GLAM – Group on Language, Audio & Music, Imperial College London, UK
[2]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[3]audEERING GmbH, Gilching, Germany
[4]Pattern Recognition Lab, FAU Erlangen-Nuremberg, Germany
[5]iDN – interdisciplinary Developmental Neuroscience, Medical University of Graz, Austria
[6]University Medical Center Göttingen, Germany
[7]Department of Women's and Children's Health, Karolinska Institutet, Stockholm, Sweden
[8]Department of Clinical Psychology and Psychotherapy, University of Ulm, Germany
[9]Shenzhen University General Hospital, Shenzhen, P.R. China
[10]Machine Intelligence & Signal Processing Group, Technische Universität München, Germany
[11]University of Oulu, Finland

**Mon-SS-2-1-1, Time: 16:30-16:42**

### Attention-based Sequence Classification for Affect Detection

*Cristina Gorrostieta, Richard Brutti, Kye Taylor, Avi Shapiro, Joseph Moran, Ali Azarbayejani and John Kane*
Cogito Corporation

**Mon-SS-2-1-2, Time: 16:42-16:54**

This paper presents the Cogito submission to the Interspeech Computational Paralinguistics Challenge (ComParE), for the second sub-challenge. The aim of this second sub-challenge is to recognize self-assessed affect from short clips of speech-containing audio data. We adopt a sequence classification-based approach where we use a long-short term memory (LSTM) network for modeling the evolution of low-level spectral coefficients, with added attention mechanism to emphasize salient regions of the audio clip. Additionally to deal with the underrepresentation of the negative valence class we use a combination of mitigation strategies including oversampling and loss function weighting. Our experiments demonstrate improvements in detection accuracy when including the attention mechanism and class balancing strategies in combination, with the best models outperforming the best single challenge baseline model.

### Computational Paralinguistics: Automatic Assessment of Emotions, Mood and Behavioural State from Acoustics of Speech

*Zafi Sherhan Syed, Julien Schroeter, Kirill Sidorov and David Marshall*
Cardiff University, UK

**Mon-SS-2-1-3, Time: 16:54-17:06**

Paralinguistic analysis of speech remains a challenging task due to the many confounding factors which affect speech production. In this paper, we address the Interspeech 2018 Computational Paralinguistics Challenge (ComParE) which aims to push the boundaries of sensitivity to non-textual information that is conveyed in the acoustics of speech. We attack the problem on several fronts. We posit that a substantial amount of paralinguistic information is contained in spectral features alone. To this end, we use a large ensemble of Extreme Learning Machines for classification of spectral features. We further investigate the applicability of (an ensemble of) CNN-GRUs networks to model temporal variations therein. We report on the details of the experiments and the results for three ComParE sub-challenges: Atypical Affect, Self-Assessed Affect and Crying. Our results compare favourably and in some cases exceed the published state-of-the-art performance.

### Investigating Utterance Level Representations for Detecting Intent from Acoustics

*SaiKrishna Rallabandi[1], Bhavya Karki[1], Carla Viegas[1,2], Eric Nyberg[1] and Alan W Black[1]*
[1]Language Technologies Institute, Carnegie Mellon University, PA, USA
[2]NOVA Laboratory for Computer Science and Informatics, FCT NOVA, Campus Caparica, Almada, Portugal

**Mon-SS-2-1-4, Time: 17:06-17:18**

Recognizing paralinguistic cues from speech has applications in varied domains of speech processing. In this paper we present approaches to identify the expressed intent from acoustics in the context of INTERSPEECH 2018 ComParE challenge. We have made submissions in three sub-

**Notes**

challenges: prediction of 1) self-assessed affect and 2) atypical affect 3) Crying Sub challenge. Since emotion and intent are perceived at suprasegmental levels, we explore a variety of utterance level embeddings. The work includes experiments with both automatically derived as well as knowledge-inspired features that capture spoken intent at various acoustic levels. Incorporation of utterance level embeddings at the text level using an off the shelf phone decoder has also been investigated. The experiments impose constraints and manipulate the training procedure using heuristics from the data distribution. We conclude by presenting the preliminary results on the development and blind test sets.

## LSTM Based Cross-corpus and Cross-task Acoustic Emotion Recognition

*Heysem Kaya[1], Dmitrii Fedotov[2,3], Ali Yeşilkanat[4], Oxana Verkholyak[2], Yang Zhang[5] and Alexey Karpov[2]*

[1]Department of Computer Engineering, Namık Kemal University, Çorlu, Tekirdağ, Turkey

[2]St. Petersburg Institute for Informatics and Automation of Russian Academy of Sciences, Russia

[3]Institute of Communications Engineering, Ulm University, Ulm, Germany

[4]Department of Computer Engineering, Boğaziçi University, İstanbul, Turkey

[5]Noah's Ark Lab, Huawei Technologies, Shenzhen, China

**Mon-SS-2-1-5, Time: 17:18-17:30**

Acoustic emotion recognition is a popular and central research direction in paralinguistic analysis, due its relation to a wide range of affective states/traits and manifold applications. Developing highly generalizable models still remains as a challenge for researchers and engineers, because of multitude of nuisance factors. To assert generalization, deployed models need to handle spontaneous speech recorded under different acoustic conditions compared to the training set. This requires that the models are tested for cross-corpus robustness. In this work, we first investigate the suitability of Long-Short-Term-Memory (LSTM) models trained with time- and space-continuously annotated affective primitives for cross-corpus acoustic emotion recognition. We next employ an effective approach to use the frame level valence and arousal predictions of LSTM models for utterance level affect classification and apply this approach on the ComParE 2018 challenge corpora. The proposed method alone gives motivating results both on development and test set of the Self-Assessed Affect Sub-Challenge. On the development set, the cross-corpus prediction based method gives a boost to performance when fused with top components of the baseline system. Results indicate the suitability of the proposed method for both time-continuous and utterance level cross-corpus acoustic emotion recognition tasks.

## Implementing Fusion Techniques for the Classification of Paralinguistic Information

*Bogdan Vlasenko[1], Jilt Sebastian[1,2,3], Pavan Kumar D. S.[1,3] and Mathew Magimai Doss[1]*

[1]Idiap Research Institute, CH-1920 Martigny, Switzerland

[2]Indian Institute of Technology Madras, India

[3]École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland

**Mon-SS-2-1-6, Time: 17:30-17:42**

This work tests several classification techniques and acoustic features and further combines them using late fusion to classify paralinguistic information for the ComParE 2018 challenge. We use Multiple Linear Regression (MLR) with Ordinary Least Squares (OLS) analysis to select the most informative features for Self-Assessed Affect (SSA) sub-Challenge. We also propose to use raw-waveform convolutional neural networks (CNN) in the context of three paralinguistic sub-challenges. By using combined evaluation split for estimating codebook, we obtain better representation for Bag-of-Audio-Words approach. We preprocess the speech to vocalized segments to improve classification performance. For fusion of our leading classification techniques, we use weighted late fusion approach applied for confidence scores. We use two mismatched evaluation phases by exchanging the training and development sets and this estimates the optimal fusion weight. Weighted late fusion provides better performance on development sets in comparison with baseline techniques. Raw-waveform techniques perform comparable to the baseline.

## General Utterance-Level Feature Extraction for Classifying Crying Sounds, Atypical & Self-Assessed Affect and Heart Beats

*Gábor Gosztolya[1], Tamás Grósz[1,2] and László Tóth[2]*

[1]MTA-SZTE Research Group on Artificial Intelligence, Szeged, Hungary

[2]Institute of Informatics, University of Szeged, Hungary

**Mon-SS-2-1-7, Time: 17:42-17:54**

In the area of computational paralinguistics, there is a growing need for general techniques that can be applied in a variety of tasks and which can be easily realized using standard and publicly available tools. In our contribution to the 2018 Interspeech Computational Paralinguistic Challenge (ComParE), we test four general ways of extracting features. Besides the standard ComParE feature set consisting of 6373 diverse attributes, we experiment with two variations of Bag-of-Audio-Words representations, and define a simple feature set inspired by Gaussian Mixture Models. Our results indicate that the UAR scores obtained via the different approaches vary among the tasks. In our view, this is mainly because most feature sets tested were local by nature and they could not properly represent the utterances of the Atypical Affect and Self-Assessed Affect Sub- Challenges. On the Crying Sub-Challenge, however, a simple combination of all four feature sets proved to be effective.

## Self-Assessed Affect Recognition Using Fusion of Attentional BLSTM and Static Acoustic Features

*Bo-Hao Su[1,2], Sung-Lin Yeh[1,2], Ming-Ya Ko[1,2], Huan-Yu Chen[1,2], Shun-Chang Zhong[1,2], Jeng-Lin Li[1,2] and Chi-Chun Lee[1,2]*

[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan

[2]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan

**Mon-SS-2-1-8, Time: 17:54-18:06**

In this study, we present a computational framework to participate in the Self-Assessed Affect Sub-Challenge in the INTERSPEECH 2018 Computation Paralinguistics Challenge. The goal of this sub-challenge is to classify the valence scores given by the speaker themselves into three different levels, i.e., low, medium and high. We explore fusion of Bi-directional LSTM with baseline SVM models to improve the recognition accuracy. In specifics, we extract frame-level acoustic LLDs as input to the BLSTM with a modified attention mechanism and separate SVMs are trained using the standard ComParE_16 baseline feature sets with minority class upsampling. These diverse prediction results are then further fused using a decision-level score fusion scheme to integrate all of the developed models. Our proposed approach achieves a 62.94% and 67.04% unweighted average recall (UAR), which is an 6.24% and 1.04% absolute improvement over the best baseline provided by the challenge organizer. We further provide a detailed comparison analysis between different models.

## Vocalic, Lexical and Prosodic Cues for the INTERSPEECH 2018 Self-Assessed Affect Challenge

*Claude Montacié[1] and Marie-José Caraty[2]*

[1]STIH Laboratory, Paris Sorbonne University, 28 rue Serpente, 75006, Paris, France

[2]STIH Laboratory, Paris Descartes University, 45 rue des Saints-Pères, 75006, Paris, France

**Mon-SS-2-1-9, Time: 18:06-18:18**

The INTERSPEECH 2018 Self-Assessed Affect Challenge consists in the prediction of the affective state of mind from speech. Experiments were

**Notes**

conducted on the Ulm State-of-Mind in Speech database (USoMS) where subjects self-report their affective state. Dimensional representation of emotion (valence) is used for labeling. We have investigated cues related to the perception of the emotional valence according to three main relevant linguistic levels: phonetics, lexical and prosodic. For this purpose we studied: the degree-of-articulation, the voice quality, an affect lexicon and the expressive prosodic contours. For the phonetics level, a set of gender-dependent audio-features was computed on vowel analysis (voice quality and speech articulation measurements). At the lexical level, an affect lexicon was extracted from the automatic transcription of the USoMS database. This lexicon has been assessed for the Challenge task comparatively to a reference polarity lexicon. In order to detect expressive prosody, N-gram models of the prosodic contours were computed from an intonation labeling system. At last, an emotional valence classifier was designed combining ComParE and eGeMAPS feature sets with other phonetic, prosodic and lexical features. Experiments have shown an improvement of 2.4% on the Test set, compared to the baseline performance of the Challenge.

### The INTERSPEECH 2018 Computational Paralinguistics Challenge: Summary of results

*Björn Schuller[1,2,3], Stefan Steidl[4], Anton Batliner[2,4], Peter B. Marschik[5,6,7], Harald Baumeister[8], Fengquan Dong[9], Simone Hantke[2,10], Florian B. Pokorny[5,10], Eva-Maria Rathner[8], Katrin D. Bartl-Pokorny[5], Christa Einspieler[5], Dajie Zhang[5,6], Alice Baird[2], Shahin Amiriparian[2,10], Kun Qian[2,10], Zhao Ren[2], Maximilian Schmitt[2], Panagiotis Tzirakis[1] and Stefanos Zafeiriou[1,11]*

[1]GLAM – Group on Language, Audio & Music, Imperial College London, UK

[2]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany

[3]audEERING GmbH, Gilching, Germany

[4]Pattern Recognition Lab, FAU Erlangen-Nuremberg, Germany

[5]iDN – interdisciplinary Developmental Neuroscience, Medical University of Graz, Austria

[6]University Medical Center Göttingen, Germany

[7]Department of Women's and Children's Health, Karolinska Institutet, Stockholm, Sweden

[8]Department of Clinical Psychology and Psychotherapy, University of Ulm, Germany

[9]Shenzhen University General Hospital, Shenzhen, P.R. China

[10]Machine Intelligence & Signal Processing Group, Technische Universität München, Germany

[11]University of Oulu, Finland

**Mon-SS-2-1-10, Time: 18:18-18:30**

.....

## Mon-S&T-2-1: Show and Tell 2

MR G.05-G.06; 16:30-18:30; Monday, 3 September, 2018

Chairs: Veena T and C. Krishna Mohan

.....

### Intonation tutor by SPIRE (In-SPIRE): An Online Tool for an Automatic Feedback to the Second Language Learners in Learning Intonation

*Anand P A[1], Chiranjeevi Yarra[2], Kausthubha N K[2], Prasanta Kumar Ghosh[2]*

[1]National Institute of Technology Karnataka (NITK), Surathkal-575025, India

[2]Electrical Engineering, Indian Institute of Science (IISc), Bangalore-560012, India

**Mon-S&T-2-1-1, Time: 16:30-18:30**

In spoken communication, intonation often conveys meaning of an utterance. Thus, incorrect intonation, typically made by second language (L2) learners, could result in miscommunication. We demonstrate In-SPIRE tool that helps the L2 learners to learn intonation in a self-learning manner. For this, we design an interactive self-explanatory front end, which is also used to send learner`s audio and hand-shake signals to the back-end. At the back-end, we implement a system that takes the learner`s audio against a specific stimuli and computes pitch patterns representing the intonation. For this, we apply pitch stylization on each syllable segment in the audio. Further, we compute a quality score using the learner`s patterns and the respective ground-truth patterns. Finally, the score, the patterns of the learners and the ground-truth are sent to the front-end for display as a feedback to the learners. Thus, the learner could correct any mismatch in his/her intonation with respect to the ground-truth. The proposed tool benefits the learners who do not have access to effective spoken language training.

### Game-based Spoken Dialog Language Learning Applications for Young Students

*Keelan Evanini[1], Veronika Timpe-Laughlin[1], Eugene Tsuprun[1], Ian Blood[1], Jeremy Lee[1], James Bruno[1], Vikram Ramanarayanan[2], Patrick Lange[2] and David Suendermann-Oeft[2]*

[1]Educational Testing Service R&D, 660 Rosedale Rd., Princeton, NJ, USA

[2]Educational Testing Service R&D, 90 New Montgomery Street, Suite 1500, San Francisco, CA, USA

**Mon-S&T-2-1-2, Time: 16:30-18:30**

This demo presents three different spoken dialog applications that were developed to provide young learners of English an opportunity to practice speaking and to receive feedback on particular aspects of their English speaking ability. The speaking tasks were designed as game-based interactions in order to engage young students and they provide feedback about grammar yes/no question formation and simple past tense verb formation) and vocabulary. A pilot study with 27 primary-level English as a foreign language (EFL) learners investigated the usefulness of these applications.

### The IBM Virtual Voice Creator

*Alexander Sorin[1], Slava Shechtman[1], Zvi Kons[1], Ron Hoory[1], Shay Ben-David[1], Joe Pavitt[2], Shai Rozenberg[1], Carmel Rabinovitz[1] and Tal Drory[1]*

[1]IBM Research – Haifa, Israel

[2]IBM Research – Hursley, UK

**Mon-S&T-2-1-3, Time: 16:30-18:30**

The IBM Virtual Voice Creator (IVVC) is an end-to-end cloud-based solution for TTS voice customization and voiceover generation in games and animated movies. The solution is based on the IBM expressive TTS technology with built-in online voice transformation capabilities. It is endowed with an interactive web GUI studio.

IVVC lets the users create unique voice personas according to their needs and imagination and control the vocal performance of the virtual speakers. IVVC provides a powerful set of controls over the voice characteristics, including the vocal tract, glottal pulse, breathiness, pitch, rate and special voice effects. IVVC also allows the user to control emotional style and emphasis in the synthesized speech.

The virtual voice design and performance controls are interactive, intuitive, fast and do not require any special skills.

**Notes**

### Mobile Application for Learning Languages for the Unlettered

*Gayathri G, Mohana N, Radhika Pal and Hema Murthy*
Indian Institute of Technology, Madras, India

**Mon-S&T-2-1-4, Time: 16:30-18:30**

Mobile based technologies have become ubiquitous and various applications from games to readers are mobile oriented. In this paper, we propose development of speech based language learning app. Conventionally language learning tools start with words, followed by sentences. The fundamental assumption is that the person is literate. In the Indian context, the literacy levels are as low as 65%. In addition, each Indian language has its own scripts. The objective of this work is to develop a mobile app that starts from the script to teach the language to a person who can speak the language but is unlettered. Since the focus is on the unlettered, writing should be easy. A script centric approach is used to learn a language. The application starts with teaching a simple letter of the alphabet, followed by another letter that can be obtained by simple modification to the previously learned letter, followed by words using the letters that are learned, followed by sentences using the learned words. At every step, a text-to-speech system is used which articulates the letters and words. The learning app is based on a book called Tamil Karpom (P Nannan). The ideas from the book are adapted for learning Hindi.

### Mandarin-English Code-switching Speech Recognition

*Haihua Xu[1], Van Tung Pham[1,2], Zin Tun Kyaw[2], Zhi Hao Lim[1], Eng Siong Chng[1,2] and Haizhou Li[3]*
[1]Temasek Laboratories, Nanyang Technological University, Singapore
[2]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[3]Department of Electrical and Computer Engineering, National University of Singapore, Singapore

**Mon-S&T-2-1-5, Time: 16:30-18:30**

This work presents the development of a Mandarin-English code-switching speech recognition system. We demonstrate three key novelties in our system. First, we increase our lexicon coverage to 360K words, where phone sets of different languages are maintained separately. Secondly, we used over 1000 hours of training data combining both mono-lingual and code-switch corpus to develop the acoustic model. Finally, for language modelling, we applied context-aware text normalization and word-class language model. When testing on our internal code-switch close talk microphone recording, the system achieves recognition performance that can support real applications.

## Mon-P-2-1: Spoken Dialogue Systems and Conversational Analysis

Hall 4-6: Poster-1; 16:30-18:30; Monday, 3 September, 2018
Chair: Helena Moniz

### Joint Learning of Domain Classification and Out-of-Domain Detection with Dynamic Class Weighting for Satisficing False Acceptance Rates

*Joo-Kyung Kim and Young-Bum Kim*
Amazon Alexa

**Mon-P-2-1-1, Time: 16:30-18:30**

In domain classification for spoken dialog systems, correct detection of out-of-domain (OOD) utterances is crucial because it reduces confusion and unnecessary interaction costs between users and the systems. Previous work usually utilizes OOD detectors that are trained separately from in-domain (IND) classifiers and confidence thresholding for OOD detection given target evaluation scores. In this paper, we introduce a neural joint learning model for domain classification and OOD detection, where dynamic class weighting is used during the model training to satisfice a given OOD false acceptance rate (FAR) while maximizing the domain classification accuracy. Evaluating on two domain classification tasks for the utterances from a large spoken dialogue system, we show that our approach significantly improves the domain classification performance with satisficing given target FARs.

### Analyzing Vocal Tract Movements During Speech Accommodation

*Sankar Mukherjee[1], Thierry Legou[2], Leonardo Lancia[2], Pauline Hilt[1], Alice Tomassini[1], Luciano Fadiga[1,3], Alessandro D'Ausilio[1,3], Leonardo Badino[1] and Noël Nguyen[2]*
[1]Center for Translational Neurophysiology of Speech and Communication, Istituto Italiano di Tecnologia, Ferrara, Italy
[2]Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France
[3]Section of Human Physiology, University of Ferrara, Italy

**Mon-P-2-1-2, Time: 16:30-18:30**

When two people engage in verbal interaction, they tend to accommodate on a variety of linguistic levels. Although recent attention has focused on to the acoustic characteristics of convergence in speech, the underlying articulatory mechanisms remain to be explored. Using 3D electromagnetic articulography (EMA), we simultaneously recorded articulatory movements in two speakers engaged in an interactive verbal game, the domino task. In this task, the two speakers take turn in chaining bi-syllabic words according to a rhyming rule. By using a robust speaker identification strategy, we identified for which specific words speakers converged or diverged. Then, we explored the different vocal tract features characterizing speech accommodation. Our results suggest that tongue movements tend to slow down during convergence whereas maximal jaw opening during convergence and divergence differs depending on syllable position.

### Cross-Lingual Multi-Task Neural Architecture for Spoken Language Understanding

*Yujiang Li[1,2], Xuemin Zhao[1,2], Weiqun Xu[1,2] and Yonghong Yan[1,2,3]*
[1]Key Laboratory of Speech Acoustic and Content Understanding, Institute of Acoustics, China
[2]University of Chinese Academy of Sciences, China
[3]Xinjiang Laboratory of Minority Speech and Language Information Processing, China

**Mon-P-2-1-3, Time: 16:30-18:30**

Cross-lingual spoken language understanding (SLU) systems traditionally require machine translation services for language portability and liberation from human supervision. However, restriction exists in parallel corpora and model architectures. Assuming reliable data are provided with human-supervision, which encourages non-parallel corpora and alleviate translation errors, this paper aims to explore cross-lingual knowledge transfer from multiple levels by taking advantage of neural architectures. We first investigate a joint model of slot filling and intent determination for SLU, which alleviates the out-of-vocabulary problem and explicitly models dependencies between output labels by combining character and word representations, bidirectional Long Short-Term Memory and conditional random fields together, while attention-based classifier is introduced for intent determination. Knowledge transfer is further operated on character-level and sequence-level, aiming to share morphological and phonological information between languages with similar alphabets by sharing character representations and characterize the sequence with language-general and language-specific knowledge adaptively acquired by separate encoders. Experimental results on the MIT-Restaurant-Corpus and the ATIS corpora in different languages demonstrate the effectiveness of the proposed methods.

**Notes**

## Statistical Model Compression for Small-Footprint Natural Language Understanding

*Grant P. Strimel, Kanthashree Mysore Sathyendra and Stanislav Peshterliev*

Alexa Machine Learning, Amazon.com

**Mon-P-2-1-4, Time:16:30-18:30**

In this paper we investigate statistical model compression applied to natural language understanding (NLU) models. Small-footprint NLU models are important for enabling offline systems on hardware restricted devices and for decreasing on-demand model loading latency in cloud-based systems. To compress NLU models, we present two main techniques, parameter quantization and perfect feature hashing. These techniques are complementary to existing model pruning strategies such as L1 regularization. We performed experiments on a large scale NLU system. The results show that our approach achieves 14-fold reduction in memory usage compared to the original models with minimal predictive performance impact.

## Comparison of an End-to-End Trainable Dialogue System with a Modular Statistical Dialogue System

*Norbert Braunschweiler and Alexandros Papangelis*

Toshiba Research Europe Limited, Cambridge Research Laboratory, Cambridge, UK

**Mon-P-2-1-5, Time: 16:30-18:30**

This paper presents a comparison of two dialogue systems: one is end-to-end trainable and the other uses a more traditional, modular architecture. End-to-end trainable dialogue systems recently attracted a lot of attention because they offer several advantages over traditional systems. One of them is the avoidance to train each system module independently, by creating a single network architecture which maps an input to the corresponding output without the need for intermediate representations. While the end-to-end system investigated here had been tested in a text-in/out scenario it remained an open question how the system would perform in a speech-in/out scenario, with noisy input from a speech recognizer and output speech generated by a speech synthesizer.

To evaluate this, both dialogue systems were trained on the same corpus, including human-human dialogues in the Cambridge restaurant domain, and then compared in both scenarios by human evaluation. The results show, that in both interfaces the end-to-end system receives significantly higher ratings on all metrics than the traditional modular system, an indication that it enables users to reach their goals faster and experience both a more natural system response and a better comprehension by the dialogue system.

## A Discriminative Acoustic-Prosodic Approach for Measuring Local Entrainment

*Megan Willi[1], Stephanie A. Borrie[2], Tyson S. Barrett[2], Ming Tu[1] and Visar Berisha[1]*

[1]Arizona State University, Tempe, AZ, USA

[2]Utah State University, Logan, UT, USA

**Mon-P-2-1-6, Time: 16:30-18:30**

Acoustic-prosodic entrainment describes the tendency of humans to align or adapt their speech acoustics to each other in conversation. This alignment of spoken behavior has important implications for conversational success. However, modeling the subtle nature of entrainment in spoken dialogue continues to pose a challenge. In this paper, we propose a straightforward definition for local entrainment in the speech domain and operationalize an algorithm based on this: acoustic-prosodic features that capture entrainment should be maximally different between real conversations involving two partners and sham conversations generated by randomly mixing the speaking turns from the original two conversational partners. We propose an approach for measuring local entrainment that quantifies alignment of behavior on a turn-by-turn basis, projecting the differences between interlocutors' acoustic-prosodic features for a given turn onto a discriminative feature subspace that maximizes the difference between real and sham conversations. We evaluate the method using the derived features to drive a classifier aiming to predict an objective measure of conversational success (i.e., low versus high), on a corpus of task-oriented conversations. The proposed entrainment approach achieves 72% classification accuracy using a Naive Bayes classifier, outperforming three previously established approaches evaluated on the same conversational corpus.

## Investigating Speech Features for Continuous Turn-Taking Prediction Using LSTMs

*Matthew Roddy[1], Gabriel Skantze[2] and Naomi Harte[1]*

[1]ADAPT Centre, School of Engineering, Trinity College Dublin, Ireland

[2]Department of Speech Music and Hearing, KTH, Stockholm, Sweden

**Mon-P-2-1-7, Time: 16:30-18:30**

For spoken dialog systems to conduct fluid conversational interactions with users, the systems must be sensitive to turn-taking cues produced by a user. Models should be designed so that effective decisions can be made as to when it is appropriate, or not, for the system to speak. Traditional end-of-turn models, where decisions are made at utterance end-points, are limited in their ability to model fast turn-switches and overlap. A more flexible approach is to model turn-taking in a continuous manner using RNNs, where the system predicts speech probability scores for discrete frames within a future window. The continuous predictions represent generalized turn-taking behaviors observed in the training data and can be applied to make decisions that are not just limited to end-of-turn detection. In this paper, we investigate optimal speech-related feature sets for making predictions at pauses and overlaps in conversation. We find that while traditional acoustic features perform well, part-of-speech features generally perform worse than word features. We show that our current models outperform previously reported baselines.

## Classification of Correction Turns in Multilingual Dialogue Corpus

*Ivan Kraljevski and Diane Hirschfeld*

voice INTER connect GmbH, Dresden, Germany

**Mon-P-2-1-8, Time: 16:30-18:30**

This paper presents a multiclass classification of correction dialog turns using machine learning. The classes are determined by the type of the introduced recognition errors while performing WOz trials and creating the multilingual corpus. Three datasets were obtained using different sets of acoustic-prosodic features on the multilingual dialogue corpus.

The classification experiments were done using different machine learning paradigms: Decision Trees, Support Vector Machines and Deep Learning. After careful experiments setup and optimization on the hyper-parameter space, the obtained classification results were analyzed and compared in the terms of accuracy, precision, recall and F1 score. The achieved results are comparable with those obtained in similar experiments on different tasks and speech databases.

## Contextual Slot Carryover for Disparate Schemas

*Chetan Naik, Arpit Gupta, Hancheng Ge, Mathias Lambert and Ruhi Sarikaya*

Amazon Alexa Machine Learning

**Mon-P-2-1-9, Time: 16:30-18:30**

In the slot-filling paradigm, where a user can refer back to slots in the context during the conversation, the goal of the contextual understanding system is to resolve the referring expressions to the appropriate slots in the context. In large-scale multi-domain systems, this presents two challenges - scaling to a very large and potentially unbounded set of slot values and dealing with diverse schemas. We present a neural network architecture that addresses the slot value scalability challenge by reformulating the contextual interpretation as a decision to carryover a slot from a set of possible candidates. To deal with heterogenous schemas, we introduce a simple data-driven method for transforming the candidate slots. Our experiments show that our approach can scale to multiple domains and provides competitive results over a strong baseline.

**Notes**

### Capsule Networks for Low Resource Spoken Language Understanding

*Vincent Renkens and Hugo van Hamme*
Department Electrical Engineering-ESAT, KULeuven
Kasteelpark Arenberg 10, Bus 2441, B-3001 Leuven Belgium

**Mon-P-2-1-10, Time: 16:30-18:30**

Designing a spoken language understanding system for command-and-control applications can be challenging because of a wide variety of domains and users or because of a lack of training data. In this paper we discuss a system that learns from scratch from user demonstrations. This method has the advantage that the same system can be used for many domains and users without modifications and that no training data is required prior to deployment. The user is required to train the system, so for a user friendly experience it is crucial to minimize the required amount of data. In this paper we investigate whether a capsule network can make efficient use of the limited amount of available training data. We compare the proposed model to an approach based on Non-negative Matrix Factorisation which is the state-of-the-art in this setting and another deep learning approach that was recently introduced for end-to-end spoken language understanding. We show that the proposed model outperforms the baseline models for three command-and-control applications: controlling a small robot, a vocally guided card game and a home automation task.

### Intent Discovery Through Unsupervised Semantic Text Clustering

*Padmasundari and Srinivas Bangalore*
Interactions LLC

**Mon-P-2-1-11, Time: 16:30-18:30**

Conversational systems need to understand spoken language to be able to converse with a human in a meaningful coherent manner. This understanding (Spoken Language understanding - SLU) of the human language is operationalized through identifying intents and entities. While classification methods that rely on labeled data are often used for SLU, creating large supervised data sets is extremely tedious and time consuming. This paper presents a practical approach to automate the process of intent discovery on unlabeled data sets of human language text through clustering techniques. We explore a range of representations for the texts and various clustering methods to validate the clustering stability through quantitative metrics like Adjusted Random Index (ARI). A final alignment of the clusters to the semantic intent is determined through consensus labelling. Our experiments on public datasets demonstrate the effectiveness of our approach generating homogeneous clusters with 89% cluster accuracy, leading to better semantic intent alignments. Furthermore, we illustrate that the clustering offer an alternate and effective way to mine sentence variants that can aid the bootstrapping of SLU models.

### Multimodal Polynomial Fusion for Detecting Driver Distraction

*Yulun Du, Alan W Black, Louis-Philippe Morency and Maxine Eskenazi*
Language Technologies Institute, Carnegie Mellon University, USA

**Mon-P-2-1-12, Time: 16:30-18:30**

Distracted driving is deadly, claiming 3,477 lives in the U.S. in 2015 alone. Although there has been a considerable amount of research on modeling the distracted behavior of drivers under various conditions, accurate automatic detection using multiple modalities and especially the contribution of using the speech modality to improve accuracy has received little attention. This paper introduces a new multimodal dataset for distracted driving behavior and discusses automatic distraction detection using features from three modalities: facial expression, speech and car signals. Detailed multimodal feature analysis shows that adding more modalities monotonically increases the predictive accuracy of the model. Finally, a simple and effective multimodal fusion technique using a polynomial fusion layer shows superior distraction detection results compared to the baseline SVM and neural network models.

### Engagement Recognition in Spoken Dialogue via Neural Network by Aggregating Different Annotators' Models

*Koji Inoue, Divesh Lala, Katsuya Takanashi and Tatsuya Kawahara*
Graduate School of Informatics, Kyoto University, Japan

**Mon-P-2-1-13, Time: 16:30-18:30**

This paper addresses engagement recognition based on four multimodal listener behaviors - backchannels, laughing, eye-gaze and head nodding. Engagement is an indicator of how much a user is interested in the current dialogue. Multiple third-party annotators give ground truth labels of engagement in a human-robot interaction corpus. Since perception of engagement is subjective, the annotations are sometimes different between individual annotators. Conventional methods directly use integrated labels, such as those generated through simple majority voting and do not consider each annotator's recognition. We propose a two-step engagement recognition where each annotator's recognition is modeled and the different annotators' models are aggregated to recognize the integrated label. The proposed neural network consists of two parts. The first part corresponds to each annotator's model which is trained with the corresponding labels independently. The second part aggregates the different annotators' models to obtain one integrated label. After each part is pre-trained, the whole network is fine-tuned through back-propagation of prediction errors. Experimental results show that the proposed network outperforms baseline models which directly recognize the integrated label without considering differing annotations.

### A First Investigation of the Timing of Turn-taking in Ruuli

*Tuarik Buanzur[1], Margaret Zellers[1], Saudah Namyalo[2] and Alena Witzlack-Makarevich[1]*
[1]Institut für Skandinavistik, Frisistik, und Allgemeine Sprachwissenschaft, University of Kiel, Germany
[2]Department of African Languages, Makerere University, Kampala, Uganda

**Mon-P-2-1-14, Time: 16:30-18:30**

Turn-taking behavior in conversation is reported to be universal among cultures, although the language-specific means used to accomplish smooth turn-taking are likely to differ. Previous studies investigating turn-taking have primarily focused on languages which are already heavily-studied. The current work investigates the timing of turn-taking in question-response sequences in naturalistic conversations in Ruuli, an under-studied Bantu language spoken in Uganda. We extracted sequences involving wh-questions and polar questions and measured the duration of the gap or overlap between questions and their following responses, additionally differentiating between different response types such as affirmative (i.e. type-conforming) or negative (i.e. non-type-conforming) responses to polar questions. We find that the timing of responses to various question types in Ruuli is consistent with timings that have been reported for a variety of other languages, with a mean gap duration between questions and responses of around 259 ms. Our findings thus emphasize the universal nature of turn-taking behavior in human interaction, despite Ruuli's substantial structural differences from languages in which turn-taking has been previously studied.

## Mon-P-2-2: Spoofing Detection

Hall 4-6: Poster-2; 16:30-18:30; Monday, 3 September, 2018
Chair: Emre Yılmaz

### Spoofing Detection Using Adaptive Weighting Framework and Clustering Analysis

*Yuanjun Zhao, Roberto Togneri and Victor Sreeram*
School of Electrical, Electronic and Computer Engineering,
The University of Western Australia (UWA), Australia

**Notes**

**Mon-P-2-2-1, Time: 16:30-18:30**

Security of Automatic Speaker Verification (ASV) systems against imposters are now focusing on anti-spoofing countermeasures. Under the severe threat of various speech spoofing techniques, ASV systems can easily be 'fooled' by spoofed speech which sounds as real as human-beings. As two effective solutions, the Constant Q Cepstral Coefficients (CQCC) and the Scattering Cepstral Coefficients (SCC) perform well on the detection of artificial speech signals, especially for attacks from speech synthesis (SS) and voice conversion (VC). However, for spoofing subsets generated by different approaches, a low Equal Error Rate (EER) cannot be maintained. In this paper, an adaptive weighting based standalone detector is proposed to address the selective detection degradation. The clustering property of the genuine and the spoofed subsets are analysed for the selection of suitable weighting factors. With a Gaussian Mixture Model (GMM) classifier as the back-end, the proposed detector is evaluated on the ASVspoof 2015 database. The EERs of 0.01% and 0.20% are obtained on the known and the unknown attacks, respectively. This presents an essential complementation between the CQCC and the SCC and also promotes the future research on generalized countermeasures.

## Exploration of Compressed ILPR Features for Replay Attack Detection

*Sarfaraz Jelil[1], Sishir Kalita[1], S R Mahadeva Prasanna[1,2] and Rohit Sinha[1]*

[1]Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati-781039, India

[2]Department of Electrical Engineering, Indian Institute of Technology Dharwad, Dharwad-580011, India

**Mon-P-2-2-2, Time: 16:30-18:30**

This paper deals with the problem of detecting replay attacks on speaker verification systems. In literature, apart from the acoustic features, source features have also been successfully used for this task. In existing source features, only the information around glottal closure instants (GCIs) have been utilized. We hypothesize that the feature derived by capturing the temporal dynamics between two GCIs would be more discriminative for such task. Motivated by that, in this work we explore the use of discrete cosine transform compressed integrated linear prediction residual (ILPR) features for discriminating between genuine and replayed signals. A spoof detection system is built using the compressed ILPR feature and a Gaussian mixture model (GMM) classifier. A baseline system is also built using constant-Q cepstral coefficient feature with GMM back-end. These systems are tested on the ASVSpoof 2017 Version 2.0 database. On fusing the systems developed using acoustic and proposed source features an equal error rate of 9.41% is achieved on the evaluation set.

## Detection of Replay-Spoofing Attacks Using Frequency Modulation Features

*Tharshini Gunendradasan[1], Buddhi Wickramasinghe[1,2], Ngoc Phu Le[1,2], Eliathamby Ambikairajah[1,2] and Julien Epps[1,2]*

[1]School of Electrical Engineering and Telecommunications, UNSW, Australia

[2]ATP Research Laboratory, DATA61, CSIRO, Australia

**Mon-P-2-2-3, Time: 16:30-18:30**

Prevention of malicious spoofing attacks is currently acknowledged as a priority area of investigation for the deployment of automatic speaker verification systems. Various features of speech signals have been used to fight counterfeit attacks. Among the different spoofing attack variants, replay attacks pose a significant threat as they do not require any expert knowledge and are difficult to detect. This paper proposes the use of a spectral centroid based frequency modulation (FM) features that we term spectral centroid deviation (SCD) for replay attack detection. Spectral centroid frequency (SCF) and spectral centroid magnitude coefficient (SCMC) features extracted from the same front-end as SCD are also investigated as complementary features. Evaluations on the ASVspoof 2017 dataset indicate that the proposed SCD features with a Gaussian Mixture Model (GMM) back-end is highly capable of discriminating genuine from replay spoofed speech, providing an equal error rate improvement greater than 60% relative to the CQCC baseline system from the ASVspoof 2017 challenge. Interestingly, experiments also reveal that the proposed SCD features exhibit an increased variance for replay spoofed speech relative to genuine speech, particularly for the lowest and highest frequency subbands.

## Effectiveness of Speech Demodulation-Based Features for Replay Detection

*Madhu Kamble, Hemlata Tak and Hemant Patil*

Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, Gujarat, India

**Mon-P-2-2-4, Time: 16:30-18:30**

Replay attack presents a great threat to Automatic Speaker Verification (ASV) system. The speech can be modeled as amplitude and frequency modulated (AM-FM) signals. In this paper, we explore speech demodulation-based features using Hilbert transform (HT) and Teager Energy Operator (TEO) for replay detection. In particular, we propose features, namely, HT-based Instantaneous Amplitude (IA) and Instantaneous Frequency (IF) Cosine Coefficients (i.e., HT-IACC and HT-IFCC) and Energy Separation Algorithm (ESA)-based features (i.e., ESA-IACC and ESA-IFCC). For adapting instantaneous energy w.r.t given sampling frequency, ESA requires 3 samples whereas HT requires relatively large number of samples and thus, ESA gives high time resolution.The experiments were performed on ASV spoof 2017 Challenge database for replay spoof speech detection (SSD).The experimental results shows that ESA-based features gave lower EER. In addition, linearly-spaced Gabor filterbank gave lower EER than Butterworth filterbank. To explore possible complementary information using amplitude and frequency, we have used score-level fusion of IA and IF. With HT-based feature set, the score-level fusion gave EER of 5.24% (dev) and 10.03% (eval), whereas ESA-based feature set reduced the EER to 2.01% (dev) and 9.64% (eval).

## Novel Variable Length Energy Separation Algorithm Using Instantaneous Amplitude Features for Replay Detection

*Madhu Kamble and Hemant Patil*

Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, Gujarat, India

**Mon-P-2-2-5, Time: 16:30-18:30**

Voice-based speaker authentication or Automatic Speaker Verification (ASV) system is now becoming practical reality after several decades of research. However, still this technology is very much susceptible to various spoofing attacks. Among various spoofing attacks, replay is the most challenging attack. In this paper, we propose a novel feature set based on our recently introduced Variable length Energy Separation Algorithm (VESA) during INTERSPEECH 2017. The key idea of this paper is to capture the Instantaneous Amplitude (IA) obtained from the instantaneous energy fluctuations. The replay speech is affected by acoustic environment and distortions of intermediate device. Thus, the noise added in replayed speech is important to detect. The Amplitude Modulations (AM) are more susceptible to noise and multipath interferences that may result due to replay mechanism. The experiments are performed on various dependency index (DI) and lower EER of 6.12% and 11.94% is found on dev and eval set, respectively, of ASV Spoof 2017 Challenge database. Furthermore, we compare our results with CQCC, LFCC, MFCC and VESA-IFCC feature sets. The score-level fusion VESA-IFCC and proposed feature set further reduced the EER to 0.19% and 7.11% on dev and eval set, respectively.

## Feature with Complementarity of Statistics and Principal Information for Spoofing Detection

*Jichen Yang[1], Changhuai You[2] and Qianhua He[1]*

[1]School of Electronic and Information Engineering, South China University of Technology, China

[2]Institute for Infocomm Research, A*STAR, Singapore

**Mon-P-2-2-6, Time: 16:30-18:30**

**Notes**

Constant-Q transform (CQT) has demonstrated its effectiveness in anti-spoofing feature analysis for automatic speaker verification. This paper introduces a statistics-plus-principal information feature where a short-term spectral statistics information (STSSI), octave-band principal information (OPI) and full-band principal information (FPI) are proposed on the basis of CQT. Firstly, in contrast to conventional utterance-level long-term statistic information, STSSI reveals the spectral statistics at frame-level, moreover it provides a feasibility condition for model training while only small training database is available. Secondly, OPI emphasizes the principal information for octave-bands, STSSI and OPI creates a strong complementarity to enhance the anti-spoofing feature. Thirdly, FPI is also of complementary effect with OPI. With the statistical property over CQT spectral domain and the principal information through discrete cosine transform (DCT), the proposed statistics-plus-principal feature shows reasonable advantage of the complementary trait for spoofing detection. In this paper, we setup deep neural network (DNN) classifiers for evaluation of the features. Experiments show the effectiveness of the proposed feature as compared to many conventional features on ASVspoof 2017 and ASVspoof 2015 corpus.

## Multiple Phase Information Combination for Replay Attacks Detection

*Dongbo Li [1], Longbiao Wang [1], Jianwu Dang[1,2], Meng Liu[1], Zeyan Oo[3], Seiichi Nakagawa[4], Haotian Guan[5] and Xiangang Li[6]*

[1]Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin University, Tianjin, China

[2]Japan Advanced Institute of Science and Technology, Ishikawa, Japan

[3]Nagaoka University of Technology, Nagaoka, Japan

[4]Chubu University, Kasugai, Japan

[5]Intelligent Spoken Language Technology (Tianjin) Co., Ltd., Tianjin, China

[6]AI Labs, Didi Chuxing, Beijing, China

**Mon-P-2-2-7, Time: 16:30-18:30**

In recent years, the performance of Automatic Speaker Verification (ASV) systems has been improved significantly. However, they are still affected by different kind of spoofing attacks. In this paper, we propose a method that fused different phase features and amplitude features to detect replay attacks. We propose the mel-scale relative phase feature and apply source-filter vocal tract feature in phase domain for replay attacks detection. These two phase-based features are combined to get complementary information. In addition to these phase haracteristics, constant Q cepstral coefficients (CQCCs) are used. The proposed methods are evaluated using the ASVspoof 2017 challenge database and Gaussian mixture model was used as the back-end model. The proposed approach achieved 55.6% relative error reduction rate than the conventional magnitude-based feature.

## Frequency Domain Linear Prediction Features for Replay Spoofing Attack Detection

*Buddhi Wickramasinghe[1,2], Saad Irtza[1], Eliathamby Ambikairajah[1,2] and Julien Epps[1,2]*

[1]School of Electrical Engineering and Telecommunications, UNSW Australia

[2]Data61, CSIRO, Sydney, Australia

**Mon-P-2-2-8, Time: 16:30-18:30**

Automatic speaker verification (ASV) systems are vulnerable to various types of spoofing attacks such as speech synthesis, voice conversion and replay attacks. Recent research has highlighted the need for more effective countermeasures for replay attacks, which can be very challenging to detect, however replayed speech has previously shown frequency band-specific differences when compared with genuine speech. In this paper, we propose the use of long-term temporal envelopes of subband signals using a frequency domain linear prediction (FDLP) framework. This flexible framework makes use of temporal envelope information, which has not previously been investigated for replay spoofing detection. Evaluations of the proposed system and its fusion with other subsystems were carried out on the ASVspoof 2017 database. Interestingly, smoother temporal envelopes, based on very long windows of up to 1 second, seem to be most successful and show good prospects for performance improvements via fusion.

## Auditory Filterbank Learning for Temporal Modulation Features in Replay Spoof Speech Detection

*Hardik Sailor, Madhu Kamble and Hemant Patil*

Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar-382007, Gujarat, India

**Mon-P-2-2-9, Time: 16:30-18:30**

In this paper, we present a standalone replay spoof speech detection (SSD) system to classify the natural vs. replay speech. The replay speech spectrum is known to be affected in the higher frequency range. In this context, we propose to exploit an auditory filterbank learning using Convolutional Restricted Boltzmann Machine (ConvRBM) with the pre-emphasized speech signals. Temporal modulations in amplitude (AM) and frequency (FM) are extracted from the ConvRBM subbands using the Energy Separation Algorithm (ESA). ConvRBM-based short-time AM and FM features are developed using cepstral processing, denoted as AM-ConvRBM-CC and FM-ConvRBM-CC. Proposed temporal modulation features performed better than the baseline Constant-Q Cepstral Coefficients (CQCC) features. On the evaluation set, an absolute reduction of 7.48% and 5.28% in Equal Error Rate (EER) is obtained using AM-ConvRBM-CC and FM-ConvRBM-CC, respectively compared to our CQCC baseline. The best results are achieved by combining scores from AM and FM cues (0.82% and 8.89% EER for development and evaluation set, respectively). The statistics of AM-FM features are analyzed to understand the performance gap and complementary information in both the features.

## Deep Siamese Architecture Based Replay Detection for Secure Voice Biometric

*Kaavya Sriskandaraja[1,2], Vidhyasaharan Sethu[1] and Eliathamby Ambikairajah[1,2]*

[1]School of Electrical Engineering and Telecommunications, UNSW Australia

[2]DATA61, CSIRO, Sydney, Australia

**Mon-P-2-2-10, Time: 16:30-18:30**

Replay attacks are the simplest and the most easily accessible form of spoofing attacks on voice biometric systems and can be hard to detect by systems designed to identify spoofing attacks based on synthesised speech. In this paper, we propose a novel approach to evaluate the similarities between pairs of speech samples to detect replayed speech based on a suitable embedding learned by deep Siamese architectures. Specifically, we train a deep Siamese network to identify pairs of genuine speech samples and pairs of replayed speech samples as being 'similar' and mixed pairs of genuine and replayed speech to be identified as 'dissimilar'. Siamese networks are particularly suited to this task and have been shown to be effective in problems where intra-class variability is large and the number of training samples per class is relatively small. The internal low-dimensional embedding learnt by the Siamese network to accomplish this task is then used as the basis for replay detection. The proposed approach outperforms state-of-the-art systems when evaluated on the ASVspoof 2017 challenge corpus without relying on fusion with other sub-systems.

## A Deep Identity Representation for Noise Robust Spoofing Detection

*Alejandro Gómez Alanís[1], Antonio M. Peinado[1], Jose A. Gonzalez[2] and Angel Gomez[1]*

[1]University of Granada, Granada, Spain

[2]University of Malaga, Malaga, Spain

**Mon-P-2-2-11, Time: 16:30-18:30**

The issue of the spoofing attacks which may affect automatic speaker verification systems (ASVs) has recently received an increased attention,

**Notes**

so that a number of countermeasures have been developed for detecting high technology attacks such as speech synthesis and voice conversion. However, the performance of anti-spoofing systems degrades significantly in noisy conditions. To address this issue, we propose a deep learning framework to extract spoofing identity vectors, as well as the use of soft missing-data masks. The proposed feature extraction employs a convolutional neural network (CNN) plus a recurrent neural network (RNN) in order to provide a single deep feature vector per utterance. Thus, the CNN is treated as a convolutional feature extractor that operates at the frame level. On top of the CNN outputs, the RNN is employed to obtain a single spoofing identity representation of the whole utterance. Experimental evaluation is carried out on both a clean and a noisy version of the ASVSpoof2015 corpus. The experimental results show that our proposals clearly outperforms other methods recently proposed such as the popular CQCC+GMM system or other similar deep feature systems for both seen and unseen noisy conditions.

### End-To-End Audio Replay Attack Detection Using Deep Convolutional Networks with Attention

*Francis Tom[1,2], Mohit Jain[1,3] and Prasenjit Dey[1]*

[1]IBM Research, India

[2]IIT Kharagpur, India

[3]University of Washington, Seattle USA

**Mon-P-2-2-12, Time: 16:30-18:30**

With automatic speaker verification (ASV) systems becoming increasingly popular, the development of robust countermeasures against spoofing is needed. Replay attacks pose a significant threat to the reliability of ASV systems because of the relative difficulty in detecting replayed speech and the ease with which such attacks can be mounted. In this paper, we propose an end-to-end deep learning framework for audio replay attack detection. Our proposed approach uses a novel visual attention mechanism on time-frequency representations of utterances based on group delay features, via deep residual learning (an adaptation of ResNet-18 architecture). Using a single model system, we achieve a perfect Equal Error Rate (EER) of 0% on both the development as well as the evaluation set of the ASVspoof 2017 dataset, against a previous best of 0.12% on the development set and 2.76% on the evaluation set reported in the literature. This highlights the efficacy of our feature representation and attention-based architecture in tackling the challenging task of audio replay attack detection.

### Decision-level Feature Switching as a Paradigm for Replay Attack Detection

*Saranya M S and Hema Murthy*

Indian Institute of Technology Madras

**Mon-P-2-2-13, Time: 16:30-18:30**

A pre-recorded audio sample of an authentic speaker presented to a voice-based biometric system is termed as a replay attack. Such attacks can be detected by identifying the characteristics of the recording device and environment. An analysis of different recording devices indicates that each recording device affects the spectrum differently. It is also observed that each feature captures specific characteristics of recording devices. In particular, Mel Filterbank Slope (MFS) captures low-frequency information corresponding to that of the low-quality recording devices, while Linear Filterbank Slope (LFS) captures high-frequency information which corresponds to that of a high-quality recording device. The proposed approach uses MFS and LFS along with Mel Frequency Cepstral Coefficients (MFCC) and Constant-Q Cepstral Coefficients (CQCC) in a Decision-level Feature Switching (DLFS) paradigm to determine whether a given utterance is spoofed. The obtained results surpass the state-of-the-art Light Convolutional Neural Network (LCNN) based replay detection system with a relative improvement of 7.43% on the ASV-spoof-2017 evaluation dataset.

### Modulation Dynamic Features for the Detection of Replay Attacks

*Gajan Suthokumar[1,2], Vidhyasaharan Sethu[1], Chamith Wijenayake[1] and Eliathamby Ambikairajah[1,2]*

[1]School of Electrical Engineering and Telecommunications, UNSW Australia

[2]DATA61, CSIRO, Sydney, Australia

**Mon-P-2-2-14, Time: 16:30-18:30**

The development of automatic systems that can detect replayed speech has emerged as a significant research challenge for securing voice biometric systems and is the focus of this paper. Specifically, this paper proposes two novel features to capture the static and dynamic characteristics of the signal from the modulation spectrum, which complement short term spectral features for use in replay detection. The modulation spectral centroid frequency feature is proposed as a vector representation of the first order spectral moments of the modulation spectrum. In conjunction to this, the long term spectral average serves to capture the static characteristics of the modulation spectrum. The proposed system, employing a GMM back-end, was evaluated on the ASVSpoof 2017 dataset and found to yield an EER of 6.54%.

## Mon-P-2-3: Speech Analysis and Representation

Hall 4-6: Poster-3; 16:30-18:30; Monday, 3 September, 2018

Chair: Haizhou Li

### On the Usefulness of the Speech Phase Spectrum for Pitch Extraction

*Erfan Loweimi, Jon Barker and Thomas Hain*

Speech and Hearing Research Group (SPandH), University of Sheffield

**Mon-P-2-3-1, Time: 16:30-18:30**

Most frequency domain techniques for pitch extraction such as cepstrum, harmonic product spectrum (HPS) and summation residual harmonics (SRH) operate on the magnitude spectrum and turn it into a function in which the fundamental frequency emerges as argmax. In this paper, we investigate the extension of these three techniques to the phase and group delay (GD) domains. Our extensions exploit the observation that the bin at which F (magnitude) becomes maximum, for some monotonically increasing function F, is equivalent to bin at which F (phase) has maximum negative slope and F (group delay) has the maximum value. To extract the pitch track from speech phase spectrum, these techniques were coupled with the source-filter model in the phase domain that we proposed in earlier publications and a novel voicing detection algorithm proposed here. The accuracy and robustness of the phase-based pitch extraction techniques are illustrated and compared with their magnitude-based counterparts using six pitch evaluation metrics. On average, it is observed that the phase spectrum can be successfully employed in pitch tracking with comparable accuracy and robustness to the speech magnitude spectrum.

### Time-regularized Linear Prediction for Noise-robust Extraction of the Spectral Envelope of Speech

*Manu Airaksinen, Lauri Juvela, Okko Räsänen and Paavo Alku*

Aalto University, Finland

**Mon-P-2-3-2, Time: 16:30-18:30**

Feature extraction of speech signals is typically performed in short-time frames by assuming that the signal is stationary within each frame. For the extraction of the spectral envelope of speech, which conveys the formant frequencies produced by the resonances of the slowly varying vocal tract, an often used frame length is within 20-30 ms. However, this kind of conventional frame-based spectral analysis is oblivious of the broader temporal context of the signal and is prone to degradation by, for example, environmental noise.

In this paper, we propose a new frame-based linear prediction (LP) analysis method that includes a regularization term that penalizes energy differences in consecutive frames of an all-pole spectral envelope model.

**Notes**

This integrates the slowly varying nature of the vocal tract as a part of the analysis. Objective evaluations related to feature distortion and phonetic representational capability were performed by studying the properties of the mel-frequency cepstral coefficient (MFCC) representations computed from different spectral estimation methods under noisy conditions using the TIMIT database. The results show that the proposed time-regularized LP approach exhibits superior MFCC distortion behavior while simultaneously having the greatest average separability of different phoneme categories in comparison to the other methods.

## Auditory Filterbank Learning Using ConvRBM for Infant Cry Classification

*Hardik B. Sailor and Hemant Patil*

Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication (DA-IICT), Gandhinagar-382007

**Mon-P-2-3-3, Time: 16:30-18:30**

The infant cry classification is a socially-relevant problem where the task is to classify the normal vs. pathological cry signals. Since the cry signals are very different from the speech signals in terms of temporal and spectral content, there is a need for better feature representation for infant cry signals. In this paper, we propose to use unsupervised auditory filterbank learning using Convolutional Restricted Boltzmann Machine (ConvRBM). Analysis of the subband filters shows that most of the subband filters are Fourier-like basis functions. The infant cry classification experiments were performed on the two databases, namely, DA-IICT Cry and Baby Chillanto. The experimental results show that the proposed features perform better than the standard Mel Frequency Cepstral Coefficients (MFCC) using various statistically meaningful performance measures. In particular, our proposed ConvRBM-based features obtained an absolute improvement of 2% and 0.58% in the classification accuracy on the DA-IICT Cry and the Baby Chillanto database, respectively.

## Effectiveness of Dynamic Features in INCA and Temporal Context-INCA

*Nirmesh Shah and Hemant Patil*

Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar-382007, India

**Mon-P-2-3-4, Time: 16:30-18:30**

Non-parallel Voice Conversion (VC) has gained significant attention since last one decade. Obtaining corresponding speech frames from both the source and target speakers before learning the mapping function in the non-parallel VC is a key step in the standalone VC task. Obtaining such corresponding pairs, is more challenging due to the fact that both the speakers may have uttered different utterances from same or the different languages. Iterative combination of a Nearest Neighbor search step and a Conversion step Alignment (INCA) and its variant Temporal Context (TC)-INCA are popular unsupervised alignment algorithms. The INCA and TC-INCA iteratively learn the mapping function after getting the Nearest Neighbor (NN) aligned pairs from the intermediate converted and the target spectral features. In this paper, we propose to use dynamic features along with static features to calculate the NN aligned pairs in both the INCA and TC-INCA algorithms (since the dynamic features are known to play a key role to differentiate major phonetic categories). We obtained on an average relative improvement of 13.75% and 5.39% with our proposed Dynamic INCA and Dynamic TC-INCA, respectively. This improvement is also positively reflected in the quality of converted voices.

## Singing Voice Phoneme Segmentation by Hierarchically Inferring Syllable and Phoneme Onset Positions

*Rong Gong and Xavier Serra*

Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain

**Mon-P-2-3-5, Time: 16:30-18:30**

In this paper, we tackle the singing voice phoneme segmentation problem in the singing training scenario by using language-independent information - onset and prior coarse duration. We propose a two-step method. In the first step, we jointly calculate the syllable and phoneme onset detection functions (ODFs) using a convolutional neural network (CNN). In the second step, the syllable and phoneme boundaries and labels are inferred hierarchically by using a duration-informed hidden Markov model (HMM). To achieve the inference, we incorporate the a priori duration model as the transition probabilities and the ODFs as the emission probabilities into the HMM. The proposed method is designed in a language-independent way such that no phoneme class labels are used. For the model training and algorithm evaluation, we collect a new jingju (also known as Beijing or Peking opera) solo singing voice dataset and manually annotate the boundaries and labels at phrase, syllable and phoneme levels. The dataset is publicly available. The proposed method is compared with a baseline method based on hidden semi-Markov model (HSMM) forced alignment. The evaluation results show that the proposed method outperforms the baseline by a large margin regarding both segmentation and onset detection tasks.

## Novel Empirical Mode Decomposition Cepstral Features for Replay Spoof Detection

*Prasad Tapkir and Hemant Patil*

Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, India

**Mon-P-2-3-6, Time: 16:30-18:30**

The advances in Automatic Speaker Verification (ASV) system for voice biometric purpose comes with the danger of spoofing attacks. The replay attack is the most accessible attack, where the attacker imitates speaker's identity by replaying the pre-recorded speech samples of the target speaker. Most of the conventional features, such as Mel Frequency Cepstral Coefficients (MFCC), Instantaneous Frequency Cepstral Coefficients (IFCC), etc. uses filterbank structure for feature extraction purpose. In this paper, we propose a novel Empirical Mode Decomposition Cepstral Coefficient (EMDCC) feature set, where the filterbank in MFCC is replaced with the Empirical Mode Decomposition (EMD) to obtain the subband signals. The propose feature set takes an advantage of using EMD that acts as a dyadic filterbank and handles the nonlinear and non-stationary nature of the speech signal. The stand-alone EMDCC feature set gives the Equal Error Rate (EER) of 28.06% compared to the baseline CQCC and MFCC system with EER of 29.18% and 31.3%, respectively on the evaluation set of ASV Spoof 2017 Challenge database. Furthermore, the proposed feature set is fused with the Linear Frequency Modified Group Delay Cepstral Coefficient (LFMGDCC) at score-level and we obtain a reduced EER of 18.36% on evaluation set.

## Novel Linear Frequency Residual Cepstral Features for Replay Attack Detection

*Hemlata Tak and Hemant Patil*

Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, Gujarat, India

**Mon-P-2-3-7, Time: 16:30-18:30**

Replay attack poses the most difficult challenge for the development of countermeasures for spoofed speech detection (SSD) system. Earlier researchers mainly used vocal tract-based (segmental) information for replay detection. However, during replay, excitation source-based information also gets affected (in particular, degradation in pitch source harmonics at higher frequency regions) due to recording environment and replay devices. Hence, in addition to the vocal tract-based system information, we have also explored the excitation source-based informations for SSD. In particular, we have used Linear Frequency Residual Cepstral Coefficients (LFRCC) for replay detection. The objective of this paper is to explore possible complementary excitation (glottal) source information present in the Linear Prediction residual-based features. Experiments performed on the ASV Spoof 2017 Challenge database with Gaussian Mixture Model (GMM) and Convolutional Neural Network (CNN) classifiers. When we combined the source and system-based information, we obtained on an average 28.77% and 42.72% relative decrease in Equal Error Rate (EER) on development and evaluation set, respectively. Furthermore, when we perform score-level fusion of feature sets (for a fixed classifier) followed by a classifier-level fusion of GMM and CNN (for a fixed feature set), we obtained reduced EER of 2.40% and 9.06% on dev and eval set, respectively.

**Notes**

## Analysis of sparse representation based feature on speech mode classification

*Kumud Tripathi and K. Sreenivasa Rao*

Dept. of Computer Science and Engineering,

Indian Institute of Technology Kharagpur, India

**Mon-P-2-3-8, Time: 16:30-18:30**

Traditional phone recognition systems are developed using read speech. But, in reality, the speech that needs to be processed by machine is not always in read mode. Therefore to handle the phone recognition in realistic scenarios, three broad modes of speech: read, conversation and extempore are considered in this study. The conversation mode includes informal communication in an unconstrained environment between two or more individuals. In the extempore mode, a person speaks with confidence without the help of notes. Read mode is a formal type of speech in a rigid environment. In this work, we have proposed a sparse based feature for speech mode classification. The effectiveness of sparse representation depends on the dictionary. Therefore, we have learned multiple overcomplete dictionaries by using parallel atom-update dictionary learning (PAU-DL) technique to capture the discrimination characteristics present in the considered speech modes. Further, sparse features correspond to the sequence of speech frames are derived using the learned dictionary by applying the orthogonal matching pursuit (OMP) algorithm. The proposed sparse features are evaluated on speech corpora consisting of six Indian languages by performing classification of speech modes. The results with the proposed sparse features outperform the standard spectral, excitation source and prosodic features.

## Multicomponent 2-D AM-FM Modeling of Speech Spectrograms

*Jitendra Kumar Dhiman, Neeraj Sharma and Chandra Sekhar Seelamantula*

Department of Electrical Engineering, Indian Institute of Science,

Bangalore - 560 012, India

**Mon-P-2-3-9, Time: 16:30-18:30**

In contrast to 1-D short-time analysis of speech, 2-D modeling of spectrograms provides a characterization of speech attributes directly in the joint time-frequency plane. Building on existing 2-D models to analyze a spectrogram patch, we propose a multicomponent 2-D AM-FM representation for spectrogram decomposition. The components of the proposed representation comprise a DC, a fundamental frequency carrier and its harmonics and a spectrotemporal envelope, all in 2-D. The number of harmonics required is patch-dependent. The estimation of the AM and FM is done using the Riesz transform and the component weights are estimated using a least-squares approach. The proposed representation provides an improvement over existing state-of-the-art approaches, for both male and female speakers. This is quantified using reconstruction SNR and perceptual evaluation of speech quality (PESQ) metric. Further, we perform an overlap-add on the DC component, pooling all the patches and obtain a time-frequency (t-f) aperiodicity map for the speech signal. We verify its effectiveness in improving speech synthesis quality by using it in an existing state-of-the-art vocoder.

## An Optimization Framework for Recovery of Speech from Phase-Encoded Spectrograms

*Abhilash Sainathan[1], Sunil Rudresh[2] and Chandra Sekhar Seelamantula[2]*

[1]Department of Electrical Communication Engineering

[2]Department of Electrical Engineering

Indian Institute of Science, Bangalore - 560 012, India

**Mon-P-2-3-10, Time: 16:30-18:30**

In general, reconstruction of a speech signal from the spectrogram is non-unique because of the unavailability of the phase spectrum. Considering zero phase would result in a minimum-phase reconstruction. This limitation is overcome by computing the recently introduced phase-encoded spectrogram. In this approach, one modifies each frame of a speech signal to possess the causal, delta-dominant (CDD) property prior to computing the spectrogram. In an earlier publication, we showed that finite-length CDD sequences can be retrieved exactly from their magnitude spectra using a cepstrum technique. Although exactness is guaranteed in principle, practical implementations result in a limited, but high, reconstruction accuracy. In this paper, we focus on increasing the reconstruction accuracy. We formulate the reconstruction problem within an optimization framework and deploy a recently proposed iterative, alternating direction method of multipliers (ADMM) algorithm called autocorrelation retrieval—Kolmogorov factorization (CoRK). Experimental validations show that the CoRK algorithm results in a reconstruction accurate up to machine precision. We also show that both CoRK and cepstrum techniques are robust and invariant to the choice of the window duration, the amount of overlap between consecutive speech frames, the strength of the delta used to impart the CDD property and the presence of noise.

## Speaker Recognition with Nonlinear Distortion: Clipping Analysis and Impact

*Wei Xia and John H.L. Hansen*

Center for Robust Speech Systems

University of Texas at Dallas, Richardson, TX 75080

**Mon-P-2-3-11, Time: 16:30-18:30**

Speech, speaker and language systems have traditionally relied on carefully collected speech material for training acoustic models. There is an overwhelming abundance of publicly accessible audio material available for training. A major challenge, however, is that such found data is not professionally recorded and therefore may contain a wide diversity of background noise, nonlinear distortions, or other unknown environmental based contamination or mismatch. There is a critical need for automatic analysis to screen such unknown data sets before acoustic model development, or to perform input audio purity screening prior to classification. In this study, we propose a waveform based clipping detection algorithm for naturalistic audio streams and analyze the impact of clipping at different severities on speech quality measures and automatic speaker recognition systems. We use the TIMIT and NIST SRE-08 corpora as case studies. The results show, as expected, that clipping introduces a nonlinear distortion into clean speech data, which reduces both speech quality and speaker recognition performance. We also investigate what degree of clipping can be present to sustain effective speech system performance. The proposed detection system, which will be released, could contribute to massive new audio collections for speech and language technology development.

## Linear Prediction Residual based Short-term Cepstral Features for Replay Attacks Detection

*Madhusudan Singh and Debadatta Pati*

Department of Electronics and Communication Engineering

National Institute of Technology Nagaland, Chumukedima,

Dimapur-797103, India

**Mon-P-2-3-12, Time: 16:30-18:30**

Modern automatic speaker verification (ASV) systems are highly vulnerable to spoof attacks and developing ASV anti-spoofing algorithms to protect ASV systems form these attacks is currently a part of active research. Contrarily to current trends on development of stand-alone spoof detection system, this work aims detection of replay attacks directly on the ASV system. The claim made through replay spoofing trials is rejected as impostors directly by ASV system. The objective here is to model the changes in the excitation signal characteristics caused by playback devices for replay detection. Accordingly, two linear prediction (LP) residual based source features are proposed for rejecting replay spoofing trials namely, RMFCC (residual mel-frequency cepstral coefficients) and LPRHEMFCC (LP residual Hilbert envelope MFCC). A comparative analysis between these two source features has been performed through speaker verification experiments to evaluate their effectiveness for ASV anti-spoofing applications. The comparison between the two has been made in the form of (source feature + MFCC) combination. The experiments are conducted using self-developed IITG-MV replay database. From the experimental results, it has been observed that 'LPRHEMFCC+MFCC' combination outperforms 'RMFCC+MFCC' combination, under replay attacks. Finally, the experiments are repeated on ASVspoof2017 database to validate the efficacy of proposed work.

**Notes**

### Analysis of Variational Mode Functions for Robust Detection of Vowels

*Surbhi Sakshi, Avinash Kumar and Gayadhar Pradhan*
Department of Electronics and Communication Engineering
National Institute of Technology Patna, India

**Mon-P-2-3-13, Time: 16:30-18:30**

In this work, initially the speech signal is decomposed into variational mode functions (VMFs) with the aid of variational mode decomposition (VMD). Each decomposed VMF represents different frequency band of the input speech signal. An approximate speech signal is then reconstructed by using a set of selected VMFs whose center frequency predominantly corresponds to the frequency range of the vowels. In the reconstructed speech signal, energy due to the high frequency unvoiced sound units and noises is suppressed. Consequently, over an analysis frame, the mean of the square magnitude (MSM) of the sample points is significantly higher for the vowels than other sound units. Further, the MSM at each time instant is non-linearly mapped (NLM) using a negative exponential functions to enhance the transitions at the onset and the offset points of vowels and suppress small fluctuations. The NLM-MSM is used as a front-end feature for discriminating vowels in a given speech signal. The experiments conducted on TIMIT database show that, the proposed approach outperforms the existing methods for the task of detecting vowels in a given speech signal under clean and noisy test scenarios.

## Mon-P-2-4: Sequence Models for ASR

Hall 4-6: Poster-4; 16:30-18:30; Monday, 3 September, 2018
Chair: Kate Knill

### Improving Attention Based Sequence-to-Sequence Models for End-to-End English Conversational Speech Recognition

*Chao Weng[1], Jia Cui[1], Guangsen Wang[2], Jun Wang[2], Chengzhu Yu[1], Dan Su[1] and Dong Yu[1]*
[1]Tencent AI Lab, Bellevue, USA
[2]Tencent AI Lab, Shenzhen, China

**Mon-P-2-4-1, Time: 16:30-18:30**

In this work, we propose two improvements to attention based sequence-to-sequence models for end-to-end speech recognition systems. For the first improvement, we propose to use an input-feeding architecture which feeds not only the previous context vector but also the previous decoder hidden state information as inputs to the decoder. The second improvement is based on a better hypothesis generation scheme for sequential minimum Bayes risk (MBR) training of sequence-to-sequence models where we introduce softmax smoothing into N-best generation during MBR training. We conduct the experiments on both Switchboard-300hrs and Switchboard+Fisher-2000hrs datasets and observe significant gains from both proposed improvements. Together with other training strategies such as dropout and scheduled sampling, our best model achieved WERs of 8.3%/15.5% on the Switchboard/CallHome subsets of Eval2000 without any external language models which is highly competitive among state-of-the-art English conversational speech recognition systems.

### Segmental Encoder-Decoder Models for Large Vocabulary Automatic Speech Recognition

*Eugen Beck, Mirko Hannemann, Patrick Dötsch, Ralf Schlüter and Hermann Ney*
Human Language Technology and Pattern Recognition, Computer Science Department, RWTH Aachen University, 52062 Aachen, Germany

**Mon-P-2-4-2, Time: 16:30-18:30**

It has been known for a long time that the classic Hidden-Markov-Model (HMM) derivation for speech recognition contains assumptions such as independence of observation vectors and weak duration modeling that are practical but unrealistic. When using the hybrid approach this is amplified by trying to fit a discriminative model into a generative one. Hidden Conditional Random Fields (CRFs) and segmental models (e.g. Semi-Markov CRFs / Segmental CRFs) have been proposed as an alternative, but for a long time have failed to get traction until recently. In this paper we explore different length modeling approaches for segmental models, their relation to attention-based systems. Furthermore we show experimental results on a handwriting recognition task and to the best of our knowledge the first reported results on the Switchboard 300h speech recognition corpus using this approach.

### Acoustic Modeling with DFSMN-CTC and Joint CTC-CE Learning

*ShiLiang Zhang and Ming Lei*
Machine Intelligence Technology, Alibaba Group

**Mon-P-2-4-3, Time: 16:30-18:30**

Recently, the connectionist temporal classification (CTC) based acoustic models have achieved comparable or even better performance, with much higher decoding efficiency, than the conventional hybrid systems in LVCSR tasks. For CTC-based models, it usually uses the LSTM-type networks as acoustic models. However, LSTMs are computationally expensive and sometimes difficult to train with CTC criterion. In this paper, inspired by the recent DFSMN works, we propose to replace the LSTMs with DFSMN in CTC-based acoustic modeling and explore how this type of non-recurrent models behave when trained with CTC loss. We have evaluated the performance of DFSMN-CTC using both context-independent (CI) and context-dependent (CD) phones as target labels in many LVCSR tasks with various amount of training data. Experimental results shown that DFSMN-CTC acoustic models using either CI-Phones or CD-Phones can significantly outperform the conventional hybrid models that trained with CD-Phones and cross-entropy (CE) criterion. Moreover, a novel joint CTC and CE training method is proposed, which enables to improve the stability of CTC training and performance. In a 20000 hours Mandarin recognition task, joint CTC-CE trained DFSMN can achieve a 11.0% and 30.1% relative performance improvement compared to DFSMN-CE models in a normal and fast speed test set respectively.

### End-to-End Speech Command Recognition with Capsule Network

*Jaesung Bae and Dae-Shik Kim*
School of Electrical Engineering,
Korea Advanced Institute of Science and Technology (KAIST)

**Mon-P-2-4-4, Time: 16:30-18:30**

In recent years, neural networks have become one of the common approaches used in speech recognition(SR), with SR systems based on Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) achieving the state-of-the-art results in various SR benchmarks. Especially, since CNNs are capable of capturing the local features effectively, they are applied to tasks which have relatively short-term dependencies, such as keyword spotting or phoneme-level sequence recognition. However, one limitation of CNNs is that, with max-pooling, they do not consider the pose relationship between low-level features. Motivated by this problem, we apply the capsule network to capture the spatial relationship and pose information of speech spectrogram features in both frequency and time axes. We show that our proposed end-to-end SR system with capsule networks on one-second speech commands dataset achieves better results on both clean and noise-added test than baseline CNN models.

### End-to-End Speech Recognition from the Raw Waveform

*Neil Zeghidour[1,2], Nicolas Usunier[1], Gabriel Synnaeve[1], Ronan Collobert[1] and Emmanuel Dupoux[2]*
[1]Facebook A.I. Research, Paris, France; New York & Menlo Park, USA
[2]CoML, ENS/CNRS/EHESS/INRIA/PSL Research University, Paris, France

**Notes**

**Mon-P-2-4-5, Time: 16:30-18:30**

State-of-the-art speech recognition systems rely on fixed, hand-crafted features such as mel-filterbanks to preprocess the waveform before the training pipeline. In this paper, we study end-to-end systems trained directly from the raw waveform, building on two alternatives for trainable replacements of mel-filterbanks that use a convolutional architecture. The first one is inspired by gammatone filterbanks (Hoshen et al., 2015; Sainath et al, 2015) and the second one by the scattering transform (Zeghidour et al., 2017). We propose two modifications to these architectures and systematically compare them to mel-filterbanks, on the Wall Street Journal dataset. The first modification is the addition of an instance normalization layer, which greatly improves on the gammatone-based trainable filterbanks and speeds up the training of the scattering-based filterbanks. The second one relates to the low-pass filter used in these approaches. These modifications consistently improve performances for both approaches and remove the need for a careful initialization in scattering-based trainable filterbanks. In particular, we show a consistent improvement in word error rate of the trainable filterbanks relatively to comparable mel-filterbanks. It is the first time end-to-end models trained from the raw signal significantly outperform mel-filterbanks on a large vocabulary task with clean recording conditions.

## A Multistage Training Framework for Acoustic-to-Word Model

*Chengzhu Yu, Chunlei Zhang, Chao Weng, Jia Cui and Dong Yu*

Tencent AI Lab, Bellevue, USA

**Mon-P-2-4-6, Time: 16:30-18:30**

Acoustic-to-word (A2W) prediction model based on Connectionist Temporal Classification (CTC) criterion has gained increasing interest in recent studies. Although previous studies have shown that A2W system could achieve competitive Word Error Rate (WER), there is still performance gap compared with the conventional speech recognition system when the amount of training data is not exceptionally large. In this study, we empirically investigate advanced model initializations and training strategies to achieve competitive speech recognition performance on 300 hour subset of the Switchboard task (SWB-300Hr). We first investigate the use of hierarchical CTC pretraining for improved model initialization. We also explore curriculum training strategy to gradually increase the target vocabulary size from 10k to 20k. Finally, joint CTC and Cross Entropy (CE) training techniques are studied to further improve the performance of A2W system. The combination of hierarchical-CTC model initialization, curriculum training and joint CTC-CE training translates to a relative of 12.1% reduction in WER. Our final A2W system evaluated on Hub5-2000 test sets achieves a WER of 11.4/20.8 for Switchboard and CallHome parts without using language model and decoder.

## Syllable-Based Sequence-to-Sequence Speech Recognition with the Transformer in Mandarin Chinese

*Shiyu Zhou[1,2], Linhao Dong[1,2], Shuang Xu[1] and Bo Xu[1]*

[1]Institute of Automation, Chinese Academy of Sciences

[2]University of Chinese Academy of Sciences

**Mon-P-2-4-7, Time: 16:30-18:30**

Sequence-to-sequence attention-based models have recently shown very promising results on automatic speech recognition (ASR) tasks, which integrate an acoustic, pronunciation and language model into a single neural network. In these models, the Transformer, a new sequence-to-sequence attention based model relying entirely on self-attention without using RNNs or convolutions, achieves a new single-model state-of-the-art BLEU on neural machine translation (NMT) tasks. Since the outstanding performance of the Transformer, we extend it to speech and concentrate on it as the basic architecture of sequence-to-sequence attention-based model on Mandarin Chinese ASR tasks. Furthermore, we investigate a comparison between syllable based model and context-independent phoneme (CI-phoneme) based model with the Transformer in Mandarin Chinese. Additionally, a greedy cascading decoder with the Transformer is proposed for mapping CI-phoneme sequences and syllable sequences into word sequences. Experiments on HKUST datasets demonstrate that syllable based model with the Transformer performs better than CI-phoneme based counterpart, and achieves a character error rate (CER) of 28.77%, which is competitive to the state-of-the-art CER of 28.0% by the joint CTC-attention based encoder-decoder network.

## Densely Connected Networks for Conversational Speech Recognition

*Kyu Han, Akshay Chandrashekaran, Jungsuk Kim and Ian Lane*

Capio Inc., Belmont, CA, USA

**Mon-P-2-4-8, Time: 16:30-18:30**

In this paper we show how we have achieved the state-of-the-art performance on the industry-standard NIST 2000 Hub5 English evaluation set. We propose densely connected LSTMs (namely, dense LSTMs), inspired by the densely connected convolutional neural networks recently introduced for image classification tasks. It is shown that the proposed dense LSTMs would provide more reliable performance as compared to the conventional, residual LSTMs as more LSTM layers are stacked in neural networks. With RNN-LM rescoring and lattice combination on the 5 systems (including 2 dense LSTM based systems) trained across three different phone sets, Capio's conversational speech recognition system has obtained 5.0% and 9.1% on Switchboard and CallHome, respectively.

## Multi-Head Decoder for End-to-End Speech Recognition

*Tomoki Hayashi[1], Shinji Watanabe[2], Tomoki Toda[1] and Kazuya Takeda[1]*

[1]Nagoya University

[2]Johns Hopkins University

**Mon-P-2-4-9, Time: 16:30-18:30**

This paper presents a new network architecture called multi-head decoder for end-to-end speech recognition as an extension of a multi-head attention model. In the multi-head attention model, multiple attentions are calculated and then, they are integrated into a single attention. On the other hand, instead of the integration in the attention level, our proposed method uses multiple decoders for each attention and integrates their outputs to generate a final output. Furthermore, in order to make each head to capture the different modalities, different attention functions are used for each head, leading to the improvement of the recognition performance with an ensemble effect. To evaluate the effectiveness of our proposed method, we conduct an experimental evaluation using Corpus of Spontaneous Japanese. Experimental results demonstrate that our proposed method outperforms the conventional methods such as location-based and multi-head attention models and that it can capture different speech/linguistic contexts within the attention-based encoder-decoder framework.

## Compressing End-to-End ASR Networks by Tensor-Train Decomposition

*Takuma Mori[1], Andros Tjandra[1,2], Sakriani Sakti[1,2] and Satoshi Nakamura[1,2]*

1Graduate School of Information Science, Nara Institute of Science and Technology, Japan

[2]RIKEN, Center for Advanced Intelligence Project AIP, Japan

**Mon-P-2-4-10, Time: 16:30-18:30**

End-to-end deep learning has become a popular framework for automatic speech recognition (ASR) tasks and it has proven itself to be a powerful solution. Unfortunately, network structures commonly have millions of parameters and large computational resources are required to make this approach feasible for training and running such networks. Moreover, many applications still prefer lightweight models of ASR that can run efficiently on mobile or wearable devices. To address this challenge, we propose an approach that can reduce the number of ASR parameters. Specifically, we perform Tensor-Train decomposition on the weight matrix of the gated recurrent unit (TT-GRU) in the end-to-end ASR framework. Experimental results on LibriSpeech data reveal that the compressed ASR with TT-GRU can maintain good performance while greatly reducing the number of parameters.

**Notes**

### Speech2Vec: A Sequence-to-Sequence Framework for Learning Word Embeddings from Speech

*Yu-An Chung and James Glass*

Computer Science and Artificial Intelligence Laboratory,

Massachusetts Institute of Technology, Cambridge, MA 02139, USA

**Mon-P-2-4-11, Time: 16:30-18:30**

In this paper, we propose a novel deep neural network architecture, Speech2Vec, for learning fixed-length vector representations of audio segments excised from a speech corpus, where the vectors contain semantic information pertaining to the underlying spoken words and are close to other vectors in the embedding space if their corresponding underlying spoken words are semantically similar. The proposed model can be viewed as a speech version of Word2Vec. Its design is based on a RNN Encoder-Decoder framework and borrows the methodology of skipgrams or continuous bag-of-words for training. Learning word embeddings directly from speech enables Speech2Vec to make use of the semantic information carried by speech that does not exist in plain text. The learned word embeddings are evaluated and analyzed on 13 widely used word similarity benchmarks and outperform word embeddings learned by Word2Vec from the transcriptions.

### Extending Recurrent Neural Aligner for Streaming End-to-End Speech Recognition in Mandarin

*Linhao Dong[1,2], Shiyu Zhou[1,2], Wei Chen[1] and Bo Xu[1]*

[1]Institute of Automation, Chinese Academy of Sciences, China

[2]University of Chinese Academy of Sciences, China

**Mon-P-2-4-12, Time: 16:30-18:30**

End-to-end models have been showing superiority in Automatic Speech Recognition (ASR). At the same time, the capacity of streaming recognition has become a growing requirement for end-to-end models. Following these trends, an encoder-decoder recurrent neural network called Recurrent Neural Aligner (RNA) has been freshly proposed and shown its competitiveness on two English ASR tasks. However, it is not clear if RNA can be further improved and applied to other spoken language. In this work, we explore the applicability of RNA in Mandarin Chinese and present four effective extensions: In the encoder, we redesign the temporal down-sampling and introduce a powerful convolutional structure. In the decoder, we utilize a regularizer to smooth the output distribution and conduct joint training with a language model. On two Mandarin Chinese conversational telephone speech recognition (MTS) datasets, our Extended-RNA obtains promising performance. Particularly, it achieves 27.7% character error rate (CER), which is superior to current state-of-the-art result on the popular HKUST task.

## Mon-P-2-5: Source Separation and Spatial Analysis

Hall 4-6: Poster-5; 16:30-18:30; Monday, 3 September, 2018

Chair: Mathew Magimai Doss

### Joint Noise and Reverberation Adaptive Learning for Robust Speaker DOA Estimation with an Acoustic Vector Sensor

*Disong Wang and Yuexian Zou*

ADSPLAB/Intelligent Lab, School of ECE, Peking University,

Shenzhen, 518055, China

**Mon-P-2-5-1, Time: 16:30-18:30**

Deep neural network (DNN) based DOA estimation (DNN-DOAest) methods report superior performance but the degradation is observed under stronger additive noise and room reverberation conditions. Motivated by our previous work with an acoustic vector sensor (AVS) and the great success of DNN based speech denoising and dereverberation (DNN-SDD), a unified DNN framework for robust DOA estimation task is thoroughly investigated in this paper. First, a novel DOA cue termed as sub-band inter-sensor data ratio (Sb-ISDR) is proposed to efficiently represent DOA information for training a DNN-DOAest model. Second, a speech-aware DNN-SDD is presented, where coherence vectors denoting the probability of time-frequency points dominated by speech signals are used as additional input to facilitate the training to predict complex ideal ratio masks. Last, by stacking the DNN-DOAest on the DNN-SDD with a joint part, the unified network is jointly fine-tuned, which enables DNN-SDD to serve as a pre-processing front-end to adaptively generate 'clean' speech features that are easier to be correctly classified by the following DNN-DOAest for robust DOA estimation. Experimental results on simulated and recorded data confirm the effectiveness and superiority of our proposed methods under different noise and reverberations compared with baseline methods.

### Multiple Concurrent Sound Source Tracking Based on Observation-Guided Adaptive Particle Filter

*Hong Liu, Haipeng Lan, Bing Yang and Cheng Pang*

Key Laboratory of Machine Perception, Peking University,

Shenzhen Graduate School

**Mon-P-2-5-2, Time: 16:30-18:30**

Particle filter (PF) has been proved to be an effective tool to track sound sources. In traditional PF, a pre-defined dynamic model is used to model source motion, which tends to be mismatched due to the uncertainty of source motion. Besides, non-stationary interferences pose a severe challenge to source tracking. To this end, an observation-guided adaptive particle filter (OAPF) is proposed for multiple concurrent sound source tracking. Firstly, sensor signals are processed in the time-frequency domain to obtain the direction of arrival (DOA) observations of sources. Then, by updating particle states with these DOA observations, angular distances between particles and observations are reduced to guide particles to directions of sources. Thirdly, particle weights are updated by an interference-adaptive likelihood function to reduce the impacts of interferences. At last, with the updated particles and the corresponding weights, OAPF is utilized to determine the final DOAs of sources. Experimental results demonstrate that our method achieves favorable performance for multiple concurrent sound source tracking in noisy environments.

### Harmonic-Percussive Source Separation of Polyphonic Music by Suppressing Impulsive Noise Events

*Gurunath Reddy M, K. Sreenivasa Rao and Partha Pratim Das*

Indian Institute of Technology, Kharagpur, India

**Mon-P-2-5-3, Time: 16:30-18:30**

In recent years, harmonic-percussive source separation methods are gaining importance because of their potential applications in many music information retrieval tasks. The goal of the decomposition methods is to achieve near real-time separation, distortion and artifact free component spectrograms and their equivalent time domain signals for potential music applications. In this paper, we propose a decomposition method based on filtering/suppressing the impulsive interference of percussive source on the harmonic components and impulsive interference of the harmonic source on the percussive components by modified moving average filter in the Fourier frequency domain. The significant advantage of the proposed method is that it minimizes the artifacts in the separated signal spectrograms. In this work, we have proposed Affine and Gain masking methods to separate the harmonic and percussive components to achieve minimal spectral leakage. The objective measures and separated spectrograms showed that the proposed method is better than the existing rank-order filtering based harmonic-percussive separation methods.

### Speaker Activity Detection and Minimum Variance Beamforming for Source Separation

*Enea Ceolini, Jithendar Anumula, Adrian Huber, Ilya Kiselev and Shih-Chii Liu*

Institute of Neuroinformatics, University of Zurich and ETH Zurich,

Zurich, Switzerland

**Mon-P-2-5-4, Time: 16:30-18:30**

This work proposes a framework that renders minimum variance

**Notes**

beamforming blind allowing for source separation in real world environments with an ad-hoc multi-microphone setup using no assumptions other than knowing the number of speakers. The framework allows for multiple active speakers at the same time and estimates the activity of every single speaker at flexible time resolution. These estimated speaker activities are subsequently used for the calibration of the beamforming algorithm. This framework is tested with three different speaker activity detection (SAD) methods, two of which use classical algorithms and one that is event-driven. Our methods, when tested in real world reverberant scenarios, can achieve very high signal-to-interference ratio (SIR) of around 20 dB and sound quality of 0.85 in short-time objective intelligibility (STOI) close to optimal beamforming results of 22 dB SIR and 0.89 in STOI.

## Sparsity-Constrained Weight Mapping for Head-Related Transfer Functions Individualization from Anthropometric Features

*Xiaoke Qi[1] and Jianhua Tao[1,2,3]*

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China

[2]CAS Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, China

[3]School of Artificial Intelligence, University of Chinese Academy of Sciences, China

**Mon-P-2-5-5, Time: 16:30-18:30**

Head-related transfer functions (HRTFs) describe the propagation of sound waves from the sound source to ear drums, which contain most of information for localization. However, HRTFs are highly individual-dependent and thus because of the difference of anthropometric features between subjects, individualization of HRTFs is a great challenge for accurate localization perception in virtual auditory displays (VAD). In this paper, we propose a sparsity-constrained weight mapping method termed SWM to obtain individual HRTFs. The key idea behind SWM is to obtain optimal weights to combine HRTFs from the training subjects based on the relationship of anthropometric features between the target subject and the training subjects. To this end, SWM learns two sparse representations between the target subject and the training subjects in terms of anthropometric features and HRTFs, respectively. A non-negative sparse model is used for this purpose when considering the non-negative property of the anthropometric features. Then, we build a mapping between the two weight vectors using a nonlinear regression. Furthermore, an iterative data extension method is proposed in order to increase training samples for mapping model. The objective and subjective experimental results show that the proposed method outperforms other methods in terms of log-spectral distortion (LSD) and localization accuracy.

## Speech Source Separation Using ICA in Constant Q Transform Domain

*Dheeraj Sai D.V.L.N, Kishor K.S and Sri Rama Murty Kodukula*

Department of Electrical Engineering, Indian Institute of Technology, Hyderabad

**Mon-P-2-5-6, Time: 16:30-18:30**

In order to separate individual sources from convoluted speech mixtures, complex-domain independent component analysis (ICA) is employed on the individual frequency bins of time frequency representations of the speech mixtures, obtained using short-time Fourier transform (STFT). The frequency components computed using STFT are separated by constant frequency difference with a constant frequency resolution. However, it is well known that the human auditory mechanism offers better resolution at lower frequencies. Hence, the perceptual quality of the extracted sources critically depends on the separation achieved in the lower frequency components. In this paper, we propose to perform source separation on the time-frequency representation computed though constant Q transform (CQT), which offers non uniform logarithmic binning in the frequency domain. Complex-domain ICA is performed on the individual bins of the CQT in order to get separated components in each frequency bin which are suitably scaled and permuted to obtain separated sources in the CQT domain. The estimated sources are obtained by applying inverse constant Q transform to the scaled and permuted sources. In comparison with the STFT based frequency domain ICA methods, there has been a consistent improvement of 3 dB or more in the Signal to Interference Ratios of the extracted sources.

## Multi-talker Speech Separation Based on Permutation Invariant Training and Beamforming

*Lu Yin[1,2], Ziteng Wang[1,2], Risheng Xia[1], Junfeng Li[1,2] and Yonghong Yan1,2,3*

[1]Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences

[2]University of Chinese Academy of Sciences

[3]Xinjiang Laboratory of Minority Speech and Language Information Processing, Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences

**Mon-P-2-5-7, Time: 16:30-18:30**

The recently proposed Permutation Invariant Training (PIT) technique addresses the label permutation problem for multi-talker speech separation. It has shown to be effective for the single-channel separation case. In this paper, we propose to extend the PIT-based technique to the multichannel multi-talker speech separation scenario. PIT is used to train a neural network that outputs masks for each separate speaker which is followed by a Minimum Variance Distortionless Response (MVDR) beamformer. The beamformer utilizes the spatial information of different speakers and alleviates the performance degradation due to misaligned labels. Experimental results show that the proposed PIT-MVDR-based technique leads to higher Signal-to-Distortion Ratios (SDRs) compared to the single-channel speech separation method when tested on two-speaker and three-speaker mixtures.

## Expectation-Maximization Algorithms for Itakura-Saito Nonnegative Matrix Factorization

*Paul Magron and Tuomas Virtanen*

Laboratory of Signal Processing, Tampere University of Technology, Finland

**Mon-P-2-5-8, Time: 16:30-18:30**

This paper presents novel expectation-maximization (EM) algorithms for estimating the nonnegative matrix factorization model with Itakura-Saito divergence. Indeed, the common EM-based approach exploits the space-alternating generalized EM (SAGE) variant of EM but it usually performs worse than the conventional multiplicative algorithm. We propose to explore more exhaustively those algorithms, in particular the choice of the methodology (standard EM or SAGE variant) and the latent variable set (full or reduced). We then derive four EM-based algorithms, among which three are novel. Speech separation experiments show that one of those novel algorithms using a standard EM methodology and a reduced set of latent variables outperforms its SAGE variants and competes with the conventional multiplicative algorithm.

## Subband Weighting for Binaural Speech Source Localization

*Karthik Girija Ramesan[1], Parth Suresh[2] and Prasanta Kumar Ghosh[1]*

1Electrical Engineering, Indian Institute of Science (IISc), Bengaluru-560012, India

[2]Computer Science and Engineering, TKM College of Engineering, Kollam-691005, India

**Mon-P-2-5-9, Time: 16:30-18:30**

We consider the task of speech source localization from a binaural recording using interaural time difference (ITD). A typical approach is to process binaural speech using gammatone filters and calculate frame-level ITD in each subband. The ITDs in each gammatone subband are statistically modelled using Gaussian mixture models (GMMs) for every direction during training. Given a binaural test-speech, the source is localized using maximum likelihood (ML) criterion. In this work, we propose a subband weighting scheme where subband likelihoods are weighted based on their reliability. We measure the reliability of a subband using the average frame level localization error obtained for the respective subbands.

**Notes**

These reliability values are used as the weights for each subband likelihood prior to combining the likelihoods for ML estimation. We also introduce non-linear warping of these weights to accommodate and analyse a larger space of possible subband weights. Experiments on Subject_003 from the CIPIC database reveal that weighting the subbands is better than the unweighted scheme of combining likelihoods.

# Plenary Talk-1

Hall 3; 8:30-9:30; Tuesday, 4 September, 2018

Chair: Hema Murthy

### Universal Tendencies for Cross-Linguistic Prosodic Tendencies: A Review and Some New Proposals

*Jacqueline Vaissière*

Professor Emeritus, Université Sorbonne Nouvelle, France

**Tue-Plenary-1, Time: 08:30-09:30**

The present talk aims first to review the literature on similar tendencies regularly observed in typologically unrelated languages. The tendencies concern the use of fundamental frequency (F0, including declination line as the reference line, the top-line, up-stepping, down-stepping, register change and range widening-reducing), lengthening-shortening maneuvers and strengthening-weakening phenomena at the glottal and supraglottic levels, for instantiating acoustically the syllable, the word, the minor and major phrases, and the utterance. Our presentation concerns only attitudinally and emotionally neutral utterances. The second part of the talk will present particular aspects: 1) the different centers of articulatory "effort" at the syllable level; 2) the suggestion of the existence of an unmarked strong-long pattern, neither trochaic nor iambic, at the word level in languages where natives don't have the consciousness of a "lexical stress," or don't agree on its existence or position; 3) the regrouping of one or more words into a prosodic phrase by the application of two established principles: a) the "hat-pattern" principle (t'Hart) favoring initial high-rising and final low-falling F0, and b) the intensive or the temporal rhythmic basic tendencies (Woodrow, Fraisse) favoring a more intense, stronger, more precisely articulated beginning and a lengthened ending; 4) the existence of a multilayer rhythm at the utterance level composed by the repetition/alternation of integrated Gestalts at the levels of the syllable, word, and phrases. One or two Gestalts will prevail perceptually depending on a) the language, b) the style, and c) the rate of speech. The impressionistic evidence of a particular type of language-dependent "rhythm" depends on the listener's expectations, related to his maternal language and the languages he already masters, and up to a certain extent, to his pre-existing theoretical beliefs.

# Tue-O-1-1: Acoustic Model Adaptation

Hall 3; 10:00-12:00; Tuesday, 4 September, 2018

Chairs: Rich Schwartz and Thomas Hain

### Learning to Adapt: A Meta-learning Approach for Speaker Adaptation

*Ondřej Klejch, Joachim Fainberg and Peter Bell*

Centre for Speech Technology Research, University of Edinburgh, Edinburgh EH8 9AB, UK

**Tue-O-1-1-1, Time: 10:00-10:20**

The performance of automatic speech recognition systems can be improved by adapting an acoustic model to compensate for the mismatch between training and testing conditions, for example by adapting to unseen speakers. The success of speaker adaptation methods relies on selecting weights that are suitable for adaptation and using good adaptation schedules to update these weights in order not to overfit to the adaptation data. In this paper we investigate a principled way of adapting all the weights of the acoustic model using a meta-learning. We show that the meta-learner can learn to perform supervised and unsupervised speaker adaptation and that it outperforms a strong baseline adapting LHUC parameters when adapting a DNN AM with 1.5M parameters. We also report initial experiments on adapting TDNN AMs, where the meta-learner achieves comparable performance with LHUC.

### Speaker Adaptation and Adaptive Training for Jointly Optimised Tandem Systems

*Yu Wang, Chao Zhang, Mark Gales and Philip Woodland*

Cambridge University Engineering Dept, Trumpington St., Cambridge CB2 1PZ, U.K

**Tue-O-1-1-2, Time: 10:20-10:40**

Speaker independent (SI) Tandem systems trained by joint optimisation of bottleneck (BN) deep neural networks (DNNs) and Gaussian mixture models (GMMs) have been found to produce similar word error rates (WERs) to Hybrid DNN systems. A key advantage of using GMMs is that existing speaker adaptation methods, such as maximum likelihood linear regression (MLLR), can be used which to account for diverse speaker variations and improve system robustness. This paper investigates speaker adaptation and adaptive training (SAT) schemes for jointly optimised Tandem systems. Adaptation techniques investigated include constrained MLLR (CMLLR) transforms based on BN features for SAT as well as MLLR and parameterised sigmoid functions for unsupervised test-time adaptation. Experiments using English multi-genre broadcast (MGB3) data show that CMLLR SAT yields a 4% relative WER reduction over jointly trained Tandem and Hybrid SI systems and further reductions in WER are obtained by system combination.

### Comparison of BLSTM-Layer-Specific Affine Transformations for Speaker Adaptation

*Markus Kitza, Ralf Schlüter and Hermann Ney*

Lehrstuhl Informatik 6 - Human Language Technology and Pattern Recognition, RWTH Aachen University

**Tue-O-1-1-3, Time: 10:40-11:00**

Bidirectional Long Short-Term Memory (BLSTM) Recurrent Neural Networks (RNN) acoustic models have demonstrated superior performance over Deep feed-forward Neural Networks (DNN) models in speech recognition and many other tasks. Although, a lot of work has been reported on DNN model adaptation, very little has been done on BLSTM model adaptation.

This work presents a systematic study on the adaptation of BLSTM acoustic models by means of learning affine transformations within the neural network on small amounts of unsupervised adaptation data.

Through a series of experiments on two major speech recognition benchmarks (Switchboard and CHiME-4), we investigate the significance of the position of the transformation in a BLSTM Network using a separate transformation for the forward- and backward-direction. We observe that applying affine transformations result in consistent relative word error rate reductions ranging from 6% to 11% depending on the task and the degree of mismatch between training and test data.

### Correlational Networks for Speaker Normalization in Automatic Speech Recognition

*Rini A Sharon, Sandeep Reddy Kothinti and Umesh Srinivasan*

Indian Institute of Technology Madras, India

**Tue-O-1-1-4, Time: 11:00-11:20**

In this paper, we propose using common representation learning(CRL) for speaker normalization in automatic speech recognition (ASR). Conventional methods like feature space maximum likelihood linear regression (fMLLR) require two pass decode and their performance is often limited by the amount of data during test. While i-vectors do not require two-pass decode, a significant number of input frames are required for estimation. Hence, as an alternative, a regression model employing correlational neural networks (CorrNet) for multi-view CRL is proposed. In this approach, the CorrNet training methodology treats normalized and un-normalized features as two parallel views of the same speech data. Once trained, this network generates frame-wise fMLLR-like features, thus overcoming the limitations

**Notes**

of fMLLR/i-vectors. The recognition accuracy using the proposed CorrNet-generated features is comparable with the i-vector model counterparts and significantly better than the un-normalized features like filterbank. With CorrNet-features, we get an absolute improvement in word error rate of 2.5% for TIMIT, 2.69% for WSJ84 and 3.2% for Switchboard-33hour over un-normalized features.

### Machine Speech Chain with One-shot Speaker Adaptation

*Andros Tjandra, Sakriani Sakti and Satoshi Nakamura*
Nara Institute of Science and Technology, Graduate School of Information Science, Japan
RIKEN, Center for Advanced Intelligence Project AIP, Japan

**Tue-O-1-1-5, Time: 11:20-11:40**

In previous work, we developed a closed-loop speech chain model based on deep learning, in which the architecture enabled the automatic speech recognition (ASR) and text-to-speech synthesis (TTS) components to mutually improve their performance. This was accomplished by the two parts teaching each other using both labeled and unlabeled data. This approach could significantly improve model performance within a single-speaker speech dataset, but only a slight increase could be gained in multi-speaker tasks. Furthermore, the model is still unable to handle unseen speakers. In this paper, we present a new speech chain mechanism by integrating a speaker recognition model inside the loop. We also propose extending the capability of TTS to handle unseen speakers by implementing one-shot speaker adaptation. This enables TTS to mimic voice characteristics from one speaker to another with only a one-shot speaker sample, even from a text without any speaker information. In the speech chain loop mechanism, ASR also benefits from the ability to further learn an arbitrary speaker's characteristics from the generated speech waveform, resulting in a significant improvement in the recognition rate.

### Domain Adaptation Using Factorized Hidden Layer for Robust Automatic Speech Recognition

*Khe Chai Sim, Arun Narayanan, Ananya Misra, Anshuman Tripathi, Golan Pundak, Tara Sainath, Parisa Haghani, Bo Li and Michiel Bacchiani*
Google, USA

**Tue-O-1-1-6, Time: 11:40-12:00**

Domain robustness is a challenging problem for automatic speech recognition (ASR). In this paper, we consider speech data collected for different applications as separate domains and investigate the robustness of acoustic models trained on multi-domain data on unseen domains. Specifically, we use Factorized Hidden Layer (FHL) as a compact low-rank representation to adapt a multi-domain ASR system to unseen domains. Experimental results on two unseen domains show that FHL is a more effective adaptation method compared to selectively fine-tuning part of the network, without dramatically increasing the model parameters. Furthermore, we found that using singular value decomposition to initialize the low-rank bases of an FHL model leads to a faster convergence and improved performance.

## Tue-O-1-2: Statistical Parametric Speech Synthesis

Hall 1; 10:00-12:00; Tuesday, 4 September, 2018
Chairs: K Sreenivasa Rao and Andrew Breen

### Waveform-Based Speaker Representations for Speech Synthesis

*Moquan Wan, Gilles Degottex and Mark J.F. Gales*
Cambridge University Engineering Department, UK

**Tue-O-1-2-1, Time: 10:00-10:20**

Speaker adaptation is a key aspect of building a range of speech processing systems, for example personalised speech synthesis. For deep-learning based approaches, the model parameters are hard to interpret, making speaker adaptation more challenging. One widely used method to address this problem is to extract a fixed length vector as speaker representation and use this as an additional input to the task-specific model. This allows speaker-specific output to be generated, without modifying the model parameters. However, the speaker representation is often extracted in a task-independent fashion. This allows the same approach to be used for a range of tasks, but the extracted representation is unlikely to be optimal for the specific task of interest. Furthermore, the features from which the speaker representation is extracted are usually pre-defined, often a standard speech representation. This may limit the available information that can be used. In this paper, an integrated optimisation framework for building a task specific speaker representation, making use of all the available information, is proposed. Speech synthesis is used as the example task. The speaker representation is derived from raw waveform, incorporating text information via an attention mechanism. This paper evaluates and compares this framework with standard task-independent forms.

### Incremental TTS for Japanese Language

*Tomoya Yanagita[1], Sakriani Sakti[1,2] and Satoshi Nakamura[1,2]*
[1]Graduate School of Information Science, Nara Institute of Science and Technology, Japan
[2]RIKEN, Center for Advanced Intelligence Project AIP, Japan

**Tue-O-1-2-2, Time: 10:20-10:40**

Simultaneous lecture translation requires speech to be translated in real time before the speaker has spoken an entire sentence since a long delay will create difficulties for the listeners trying to follow the lecture. The challenge is to construct a full-fledged system with speech recognition, machine translation and text-to-speech synthesis (TTS) components that could produce high-quality speech translations on the fly. Specifically for a TTS, this poses problems as a conventional framework commonly requires the language-dependent contextual linguistics of a full sentence to produce a natural-sounding speech waveform. Several studies have proposed ways for an incremental TTS (ITTS), in which it can estimate the target prosody from only partial knowledge of the sentence. However, most investigations are being done only in French, English and German. French is a syllable-timed language and the others are stress-timed languages. The Japanese language, which is a mora-timed language, has not been investigated so far. In this paper, we evaluate the quality of Japanese synthesized speech based on various linguistic and temporal incremental units. Experimental results reveal that an accent phrase incremental unit (a group of moras) is essential for a Japanese ITTS as a trade-off between quality and synthesis units.

### Transfer Learning Based Progressive Neural Networks for Acoustic Modeling in Statistical Parametric Speech Synthesis

*Ruibo Fu[1,2], Jianhua Tao[1,2,3], Yibin Zheng[1,2] and Zhengqi Wen[1]*
[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3]CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China

**Tue-O-1-2-3, Time: 10:40-11:00**

The fundamental frequency and the spectrum parameters of the speech are correlated thus one of their learned mapping from the linguistic features can be leveraged to help determine the other. The conventional methods treated all the acoustic features as one stream for acoustic modeling. And the multi-task learning methods were applied to acoustic modeling with several targets in a global cost function. To improve the accuracy of the acoustic model, the progressive deep neural networks (PDNN) is applied for acoustic modeling in statistical parametric speech synthesis (SPSS) in our method. Each type of the acoustic features is modeled in different sub-

**Notes**

networks with its own cost function and the knowledge transfers through lateral connections. Each sub-network in the PDNN can be trained step by step to reach its own optimum. Experiments are conducted to compare the proposed PDNN-based SPSS system with the standard DNN methods. The multi-task learning (MTL) method is also applied to the structure of PDNN and DNN as the contrast experiment of the transfer learning. The computational complexity, prediction sequences and quantity of hierarchies of the PDNN are investigated. Both objective and subjective experimental results demonstrate the effectiveness of the proposed technique.

### A Unified Framework for the Generation of Glottal Signals in Deep Learning-based Parametric Speech Synthesis Systems

*Min-Jae Hwang[1,2], Eunwoo Song[1,2], Jin-Seob Kim[2] and Hong-Goo Kang[1]*

[1]Department of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea

[2]NAVER Corp., Seongnam, Korea

**Tue-O-1-2-4, Time: 11:00-11:20**

In this paper, we propose a unified training framework for the generation of glottal signals in deep learning (DL)-based parametric speech synthesis systems.

The glottal vocoding-based speech synthesis system, especially the modeling-by-generation (MbG) structure that we proposed recently, significantly improves the naturalness of synthesized speech by faithfully representing the noise component of the glottal excitation with an additional DL structure.

Because the MbG method introduces a multistage processing pipeline, however, its training process is complicated and inefficient.

To alleviate this problem, we propose a unified training approach that directly generates speech parameters by merging all the required models, such as acoustic, glottal and noise models into a single unified network.

Considering the fact that noise analysis should be performed after training the glottal model, we also propose a stochastic noise analysis method that enables noise modeling to be included in the unified training process by iteratively analyzing the noise component in every epoch.

Both objective and subjective test results verify the superiority of the proposed algorithm compared to conventional methods.

### Acoustic Modeling Using Adversarially Trained Variational Recurrent Neural Network for Speech Synthesis

*Joun Yeop Lee[1], Sung Jun Cheon[1], Byoung Jin Choi[1], Nam Soo Kim[1] and Eunwoo Song[2]*

[1]Department of Electrical and Computer Engineering and INMC, Seoul National University, Korea

[2]NAVER Corp., Seongnam, Korea

**Tue-O-1-2-5, Time: 11:20-11:40**

In this paper, we propose a variational recurrent neural network (VRNN) based method for modeling and generating speech parameter sequences. In recent years, the performance of speech synthesis systems has been improved over conventional techniques thanks to deep learning-based acoustic models. Among the popular deep learning techniques, recurrent neural networks (RNNs) has been successful in modeling time-dependent sequential data efficiently. However, due to the deterministic nature of RNNs prediction, such models do not reflect the full complexity of highly structured data, like natural speech. In this regard, we propose adversarially trained variational recurrent neural network (AdVRNN) which use VRNN to better represent the variability of natural speech for acoustic modeling in speech synthesis. Also, we apply adversarial learning scheme in training AdVRNN to overcome oversmoothing problem. We conducted comparative experiments for the proposed VRNN with the conventional gated recurrent unit which is one of RNNs, for speech synthesis system. It is shown that the proposed AdVRNN based method performed better than the conventional GRU technique.

### On the Application and Compression of Deep Time Delay Neural Network for Embedded Statistical Parametric Speech Synthesis

*Yibin Zheng[1,2], Jianhua Tao[1,2], Zhengqi Wen[1] and Ruibo Fu[1,2]*

[1]National Laboratory of Pattern Recognition, Institute of Automation, CAS, China

[2]School of Artificial Intelligence, University of Chinese Academy of Science, China

**Tue-O-1-2-6, Time: 11:40-12:00**

Acoustic models based on long short-term memory (LSTM) recurrent neural networks (RNNs) were applied to statistical parametric speech synthesis (SPSS) and shown significant improvements. However, the model complexity and inference time cost of RNNs are much higher than feed-forward neural networks (FNN) due to the sequential nature of the learning algorithm, thus limiting its usage in many runtime applications. In this paper, we explore a novel application of deep time delay neural network (TDNN) for embedded SPSS, which requires low disk footprint, memory and latency. The TDNN could model long short-term temporal dependencies with inference cost comparable to standard FNN. Temporal subsampling enabled by TDNN could reduce computational complexity. Then we compress deep TDNN using singular value decomposition (SVD) to further reduce model complexity, which are motivated by the goal of building embedded SPSS systems which can be run efficiently on mobile devices. Both objective and subjective experimental results show that, the proposed deep TDNN with SVD compression could generate synthesized speech with better speech quality than FNN and comparable speech quality to LSTM, while drastically reduce model complexity and speech parameter generation time.

## Tue-O-1-3: Emotion Modeling

Hall 2; 10:00-12:00; Tuesday, 4 September, 2018

Chairs: Carlos Busso and Elmar Noeth

### Integrating Recurrence Dynamics for Speech Emotion Recognition

*Efthymios Tzinis[1,2], Georgios Paraskevopoulos[1,2], Christos Baziotis[1] and Alexandros Potamianos[1,2]*

[1]School of Electrical & Computer Engineering, National Technical University of Athens, Greece

[2]Behavioral Signal Technologies, Los Angeles, CA, USA

**Tue-O-1-3-1, Time: 10:00-10:20**

We investigate the performance of features that can capture nonlinear recurrence dynamics embedded in the speech signal for the task of Speech Emotion Recognition (SER). Reconstruction of the phase space of each speech frame and the computation of its respective Recurrence Plot (RP) reveals complex structures which can be measured by performing Recurrence Quantification Analysis (RQA). These measures are aggregated by using statistical functionals over segment and utterance periods. We report SER results for the proposed feature set on three databases using different classification methods. When fusing the proposed features with traditional feature sets, e.g., [1], we show an improvement in unweighted accuracy of up to 5.7% and 10.7% on Speaker-Dependent (SD) and Speaker-Independent (SI) SER tasks, respectively, over the baseline [1]. Following a segment-based approach we demonstrate state-of-the-art performance on IEMOCAP using a Bidirectional Recurrent Neural Network.

### Towards Temporal Modelling of Categorical Speech Emotion Recognition

*Wenjing Han[1], Huabin Ruan[2], Xiaomin Chen[3], Zhixiang Wang[1],*

**Notes**

*Haifeng Li[3] and Björn Schuller[3,4,5]*
[1]Samsung Research Institute China Beijing (SRC-B), Beijing, China
[2]Protein Research Technology Center, Tsinghua University, Beijing, China
[3]School of Computing Science and Technology, Harbin Institute of Technology, Harbin, China
[4]GLAM - Group on Language, Audio & Music, Imperial College London, U.K.
[5]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany

**Tue-O-1-3-2, Time: 10:20-10:40**

To model the categorical speech emotion recognition task in a temporal manner, the first challenge arising is how to transfer the categorical label for each utterance into a label sequence. To settle this, we make a hypothesis that an utterance is consisting of emotional and non-emotional segments and these non-emotional segments correspond to silent regions, short pauses, transitions between phonemes, unvoiced phonemes, etc.

With this hypothesis, we propose to treat an utterance's label sequence as a chain of two states: the emotional state denoting the emotional frame and Null denoting the non-emotional frame. Then, we exploit a recurrent neural network based connectionist temporal classification model to automatically label and align an utterance's emotional segments with emotional labels, while non-emotional segments with Nulls. Experimental results on the IEMOCAP corpus validate our hypothesis and also demonstrate the effectiveness of our proposed method compared to the state-of-the-art algorithms.

### Emotion Recognition from Human Speech Using Temporal Information and Deep Learning

*John Kim[1] and Rif A. Saurous[2]*
[1]Menlo School, Atherton, CA, USA
[2]Google Inc. Mountain View, CA, USA

**Tue-O-1-3-3, Time: 10:40-11:00**

Emotion recognition by machine is a challenging task, but it has great potential to make empathic human-machine communications possible. In conventional approaches that consist of feature extraction and classifier stages, extensive studies have devoted their effort to developing good feature representations, but relatively little effort was made to make proper use of the important temporal information in these features. In this paper, we propose a model combining features known to be useful for emotion recognition and deep neural networks to exploit temporal information when recognizing emotion status. A benchmark evaluation on EMO-DB demonstrates that the proposed model achieves a state-of-the-art performance of 88.9% recognition rate.

### Role of Regularization in the Prediction of Valence from Speech

*Kusha Sridhar, Srinivas Parthasarathy and Carlos Busso*
Multimodal Signal Processing(MSP) lab, Department of Electrical and Computer Engineering, The University of Texas at Dallas, Richardson TX 75080, USA

**Tue-O-1-3-4, Time: 11:00-11:20**

Regularization plays a key role in improving the prediction of emotions using attributes such as arousal, valence and dominance. Regularization is particularly important with deep neural networks (DNNs), which have millions of parameters. While previous studies have reported competitive performance for arousal and dominance, the prediction results for valence using acoustic features are significantly lower. We hypothesize that higher regularization can lead to better results for valence. This study focuses on exploring the role of dropout as a form of regularization for valence suggesting the need for higher regularization. We analyze the performance of regression models for valence, arousal and dominance as a function of the dropout probability. We observe that the optimum dropout rates are consistent for arousal and dominance. However, the optimum dropout rate for valence is higher. To understand the need for higher regularization for valence, we perform an empirical analysis to explore the nature of emotional cues conveyed in speech. We compare regression models with speaker-dependent and speaker-independent partitions for training and testing. The experimental evaluation suggests stronger speaker dependent traits for valence. We conclude that higher regularization is needed for valence to force the network to learn global patterns that generalize across speakers.

### Learning Spontaneity to Improve Emotion Recognition in Speech

*Karttikeya Mangalam and Tanaya Guha*
Indian Institute of Technology, Kanpur

**Tue-O-1-3-5, Time: 11:20-11:40**

We investigate the effect and usefulness of spontaneity (i.e. whether a given speech is spontaneous or not) in speech in the context of emotion recognition. We hypothesize that emotional content in speech is interrelated with its spontaneity and use spontaneity classification as an auxiliary task to the problem of emotion recognition. We propose two supervised learning settings that utilize spontaneity to improve speech emotion recognition: a hierarchical model that performs spontaneity detection before performing emotion recognition and a multitask learning model that jointly learns to recognize both spontaneity and emotion. Through various experiments on the well-known IEMOCAP database, we show that by using spontaneity detection as an additional task, significant improvement can be achieved over emotion recognition systems that are unaware of spontaneity. We achieve state-of-the-art emotion recognition accuracy (4-class, 69.1%) on the IEMOCAP database outperforming several relevant and competitive baselines.

### Predicting Categorical Emotions by Jointly Learning Primary and Secondary Emotions through Multitask Learning

*Reza Lotfian and Carlos Busso*
Multimodal Signal Processing(MSP) lab, Department of Electrical and Computer Engineering, The University of Texas at Dallas, Richardson TX 75080, USA

**Tue-O-1-3-6, Time: 11:40-12:00**

Detection of human emotions is an essential part of affect-aware human-computer interaction (HCI). In daily conversations, the preferred way of describing affects is by using categorical emotion labels (e.g., sad, anger, surprise). In categorical emotion classification, multiple descriptors (with different degrees of relevance) can be assigned to a sample. Perceptual evaluations have relied on primary and secondary emotions to capture the ambiguous nature of spontaneous recordings. Primary emotion is the most relevant category felt by the evaluator. Secondary emotions capture other emotional cues also conveyed in the stimulus. In most cases, the labels collected from the secondary emotions are discarded, since assigning a single class label to a sample is preferred from an application perspective. In this work, we take advantage of both types of annotations to improve the performance of emotion classification. We collect the labels from all the annotations available for a sample and generate primary and secondary emotion labels. A classifier is then trained using multitask learning with both primary and secondary emotions. We experimentally show that considering secondary emotion labels during the learning process leads to relative improvements of 7.9% in F1-score for an 8-class emotion classification task.

## Tue-O-1-4: Models of Speech Perception

MR G.01-G.02; 10:00-12:00; Tuesday, 4 September, 2018
Chairs: Bernd T. Meyer and Rohit Sinha

### Picture Naming or Word Reading: Does the Modality Affect Speech Motor Adaptation and Its Transfer?

*Tiphaine Caudrelier, Pascal Perrier, Jean-Luc Schwartz and*

**Notes**

*Amélie Rochet-Capellan*
Univ. Grenoble Alpes, CNRS, Grenoble INP*, GIPSA-lab, 38000 Grenoble, France
*Institute of Engineering Univ. Grenoble Alpes

**Tue-O-1-4-1, Time: 10:00-10:20**

Auditory-motor adaptation and transfer paradigms are increasingly used to explore speech motor control as well as phonological representations underlying speech production. Auditory-motor adaptation is generally assumed to occur at the sensory-motor level. However, few studies suggested that linguistic or contextual factors such as the modality of presentation of stimuli influences adaptation.

The present study investigates the influence of the modality of stimuli presentation (written word vs. a picture representing the same word) on auditory-motor adaptation and transfer. In this speech production experiment, speakers' auditory feedback was altered online, inducing adaptation. We contrasted the magnitude of adaptation in these two different modalities and we assessed transfer from /pe/ to the French word /epe/ in the same vs. different modality of presentation, using a mixed 2*2 subject design.

The magnitude of adaptation was not different between modalities. This observation contrasts with recent findings showing an effect of the modality (a written word vs. a go signal) on adaptation. Moreover, transfer did occur from one modality to the other and transfer pattern depended on the modality of transfer stimuli. Overall, the results suggest that picture naming and word reading rely on sensory-motor representations that may be linked to contextual (or surface) characteristics.

## Measuring the Band Importance Function for Mandarin Chinese with a Bayesian Adaptive Procedure

*Yufan Du[1], Yi Shen[2], Hongying Yang[1], Xihong Wu[1] and Jing Chen[1]*
[1]Department of Machine Intelligence, Speech and Hearing Research Center and Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing, China.
[2]Department of Speech and Hearing Sciences, Indiana University Bloomington, 200 S Jordan Ave., Bloomington, IN 47405

**Tue-O-1-4-2, Time: 10:20-10:40**

A speech intelligibility index (SII) based band importance function (BIF) for Mandarin monosyllabic words spoken by a female speaker was derived with an adaptive procedure in this work. The adaptive procedure, namely the quick-band-importance-function (qBIF) procedure, optimized the stimulus on each trial according listeners' performance on proceeding trials in an iterative fashion. This method greatly improved the efficiency of data collection. Test-retest experiments were conducted and confirmed the reliability of this adaptive procedure at a group level. The BIF derived in this work showed generally consistence with the BIF derived with the traditional paradigm with noticeable differences at certain frequencies.

## Wide Learning for Auditory Comprehension

*Elnaz Shafaei-Bajestan and R. Harald Baayen*
Quantitative Linguistics, Eberhard Karls Universität Tübingen, Tübingen, Germany

**Tue-O-1-4-3, Time: 10:40-11:00**

Classical linguistic, cognitive and engineering models for speech recognition and human auditory comprehension posit representations for sounds and words that mediate between the acoustic signal and interpretation. Recent advances in automatic speech recognition have shown, using deep learning, that state-of-the-art performance is obtained without such units. We present a cognitive model of auditory comprehension based on wide rather than deep learning that was trained on 20 to 80 hours of TV news broadcasts. Just as deep network models, our model is an end-to-end system that does not make use of phonemes and phonological wordform representations. Nevertheless, it performs well on the difficult task of single word identification (model accuracy 11.37%, Mozilla DeepSpeech: 4.45%). The architecture of the model is a simple two-layered wide neural network with weighted connections between the acoustic frequency band features as inputs and lexical outcomes (pointers to semantic vectors) as outputs. Model performance shows hardly any degredation when trained on speech in noise rather than on clean speech. Performance was further enhanced by adding a second network to a standard wide network. The present word recognition module is designed to become part of a larger system modeling the comprehension of running speech.

## Analyzing Reaction Time Sequences from Human Participants in Auditory Experiments

*Louis ten Bosch[1,2], Mirjam Ernestus[1,2] and Lou Boves[1]*
[1]Radboud University Nijmegen, NL
[2]Max Planck Institute for Psycholinguistics

**Tue-O-1-4-4, Time: 11:00-11:20**

Sequences of reaction times (RT) produced by participants in an experiment are not only influenced by the stimuli, but by many other factors as well, including fatigue, attention, experience, IQ, handedness, etc. These confounding factors result in longterm effects (such as a participant's overall reaction capability) and in short- and medium-time fluctuations in RTs (often referred to as 'local speed effects'). Because stimuli are usually presented in a random sequence different for each participant, local speed effects affect the underlying 'true' RTs of specific trials in different ways across participants. To be able to focus statistical analysis on the effects of the cognitive process under study, it is necessary to reduce the effect of confounding factors as much as possible. In this paper we propose and compare techniques and criteria for doing so, with focus on reducing ('filtering') the local speed effects. We show that filtering matters substantially for the significance analyses of predictors in linear mixed effect regression models. The performance of filtering is assessed by the average between-participant correlation between filtered RT sequences and by Akaike's Information Criterion, an important measure of the goodness-of-fit of linear mixed effect regression models.

## Prediction of Perceived Speech Quality Using Deep Machine Listening

*Jasper Ooster[1,3], Rainer Huber[2,3] and Bernd T. Meyer[1,3]*
[1]Medizinische Physik, Carl von Ossietzky Universität, Oldenburg, Germany
[2]Fraunhofer IDMT - Hearing, Speech and Audio Technology, Oldenburg, Germany
[3]Cluster of Excellence Hearing4all, Germany

**Tue-O-1-4-5, Time: 11:20-11:40**

Subjective ratings of speech quality (SQ) are essential for evaluating algorithms for speech transmission and enhancement. In this paper we explore a non-intrusive model for SQ prediction based on the output of a deep neural net (DNN) from a regular automatic speech recognizer. The degradation of phoneme probabilities obtained from the net is quantified with the mean temporal distance proposed earlier for multi-stream ASR. The SQ predicted with this method is compared with average subject ratings from the TCD-VoIP speech quality database that covers several effects of SQ degradation that can occur in VoIP applications such as clipping, packet loss, echo effects, background noise and competing speakers. Our approach is tailored to speech and therefore not applicable when quality is degraded by a competing speaker, which is reflected by an insignificant correlation between model output and subjective SQ. In all other conditions mentioned above, the model reaches an average correlation of r=0.87, which is higher than the correlation achieved with the baseline ITU-T P.563 (r=0.71) and the American National Standard ANIQUE+ (r=0.75). Since the most robust ASR system is not necessarily the best model to predict SQ, we investigate the effect of the amount of training data on quality prediction.

## Prediction of Subjective Listening Effort from Acoustic Data with Non-Intrusive Deep Models

*Paul Kranzusch[1,3], Rainer Huber[2,3], Melanie Krüger[4], Birger Kollmeier[1,3] and Bernd T. Meyer[1,3]*
[1]Medizinische Physik, Carl von Ossietzky Universität, Oldenburg, Germany
[2]Fraunhofer IDMT - Hearing, Speech and Audio Technology,

**Notes**

Oldenburg, Germany
[3]Cluster of Excellence Hearing4all, Germany
[4]Hörzentrum Oldenburg GmbH, Oldenburg, Germany

**Tue-O-1-4-6, Time: 11:40-12:00**

The effort of listening to spoken language is a highly important perceptive measure for the design of speech enhancement algorithms and hearing-aid processing. In previous research, we proposed a model that quantifies the phoneme output probabilities obtained from a deep neural net (DNN), which resulted in accurate predictions for unseen speech samples. However, high correlations between subjective ratings and model output were observed in known noise types, which is an unrealistic assumption in real-life scenarios. This paper explores non-intrusive listening effort prediction in unseen noisy environments. A set of different noise types are used for training a standard automatic speech recognition (ASR) system. Model predictions are produced by measuring the mean temporal distance of phoneme vectors from the DNN and compared to subjective ratings of hearing-impaired and normal-hearing listener responses group in three databases that cover a variety of noise types and signal enhancement algorithms. We obtain an average correlation of 0.88 and outperform three baseline measures in most conditions.

# Tue-O-1-5: Multimodal Dialogue Systems

MR G.03-G.04; 10:00-12:00; Tuesday, 4 September, 2018
Chairs: Alexandros Potamianos and Helen Meng

## A Case Study on the Importance of Belief State Representation for Dialogue Policy Management

*Margarita Kotti[1], Vassilios Diakoloukas[2], Alexandros Papangelis[1], Michail Lagoudakis[2] and Yannis Stylianou[1,3]*
[1]Speech Technology Group, Toshiba Research Cambridge, UK
[2]School of Electrical & Computer Engineering, Technical University of Crete, Greece
[3]Department of Computer Science, University of Crete, Greece

**Tue-O-1-5-1, Time: 10:00-10:20**

A key component of task-oriented dialogue systems is the belief state representation, since it directly affects the policy learning efficiency. In this paper, we propose a novel, binary, compact, yet scalable belief state representation. We compare the standard verbose belief state representation (268 dimensions) with the domain-independent representation (57 dimensions) and the proposed representation (13 or 4 dimensions). To test those representations, the recently introduced Advantage Actor Critic (A2C) algorithm is exploited. The latter has not been tested before for any representation apart from the verbose one. We study the effect of the belief state representation within A2C under 0%, 15%, 30% and 45% semantic error rate and conclude that the novel binary representation in general outperforms both the domain-independent and the verbose belief state representation. Further, the robustness of the binary representation is tested under more realistic scenarios with mismatched semantic error rates, within the A2C and DQN algorithms. The results indicate that the proposed compact, binary representation performs better or similarly to the other representations, being an efficient and promising alternative to the full belief.

## Prediction of Turn-taking Using Multitask Learning with Prediction of Backchannels and Fillers

*Kohei Hara, Koji Inoue, Katsuya Takanashi and Tatsuya Kawahara*
School of Informatics, Kyoto University, Japan

**Tue-O-1-5-2, Time: 10:20-10:40**

We address prediction of turn-taking considering related behaviors such as backchannels and fillers. Backchannels are used by listeners to acknowledge that the current speaker can hold the turn. On the other hand, fillers are used by prospective speakers to indicate a will to take a turn. We propose a turn-taking model based on multitask learning in conjunction with prediction of backchannels and fillers. The multitask learning of LSTM neural networks shared by these tasks allows for efficient and generalized learning and thus improves prediction accuracy. Evaluations with two kinds of dialogue corpora of human-robot interaction demonstrate that the proposed multitask learning scheme outperforms the conventional single-task learning.

## Conversational Analysis Using Utterance-level Attention-based Bidirectional Recurrent Neural Networks

*Chandrakant Bothe, Sven Magg, Cornelius Weber and Stefan Wermter*
Knowledge Technology, Department of Informatics, University of Hamburg, Vogt-Koelln-Str. 30, 22527 Hamburg, Germany

**Tue-O-1-5-3, Time: 10:40-11:00**

Recent approaches for dialogue act recognition have shown that context from preceding utterances is important to classify the subsequent one. It was shown that the performance improves rapidly when the context is taken into account. We propose an utterance-level attention-based bidirectional recurrent neural network (Utt-Att-BiRNN) model to analyze the importance of preceding utterances to classify the current one. In our setup, the BiRNN is given the input set of current and preceding utterances. Our model outperforms previous models that use only preceding utterances as context on the used corpus. Another contribution of our research is a mechanism to discover the amount of information in each utterance to classify the subsequent one and to show that context-based learning not only improves the performance but also achieves higher confidence in the recognition of dialogue acts. We use character- and word-level features to represent the utterances.

The results are presented for character and word feature representations and as an ensemble model of both representations. We found that when classifying short utterances, the closest preceding utterances contribute to a higher degree.

## A Comparative Study of Statistical Conversion of Face to Voice Based on Their Subjective Impressions

*Yasuhito Ohsugi, Daisuke Saito and Nobuaki Minematsu*
Graduate School of Engineering, The University of Tokyo
Vogt-Koelln-Str. 30, 22527 Hamburg, Germany

**Tue-O-1-5-4, Time: 11:00-11:20**

Recently, various types of Voice-based User Interfaces (VUIs) including smart speakers have been developed to be on the market. However, many of the VUIs use only synthetic voices to provide information for users. To realize a more natural interface, one feasible solution will be personifying VUIs by adding visual features such as face, but what kind of face is suited to a given quality of voice or what kind of voice quality is suited to a given face? In this paper, we test methods of statistical conversion from face to voice based on their subjective impressions. To this end, six combinations of two types of face features, one type of speech features, and three types of conversion models are tested using a parallel corpus developed based on subjective mapping from face features to voice features. The experimental results show that each subject judge one specific and subject-dependent voice quality as suited to different faces and that the optimal number of mixtures of face features is different from the numbers of mixtures of voice features tested.

## Follow-up Question Generation Using Pattern-based Seq2seq with a Small Corpus for Interview Coaching

*Ming-Hsiang Su[1], Chung-Hsien Wu[1], Kun-Yi Huang[1], Qian-Bei Hong[2] and Huai-Hung Huang[1]*
[1]Department of Computer Science and Information Engineering, National Cheng Kung University, Taiwan
[2]PhD Program for Multimedia Systems and Intelligent Computing,

**Notes**

National Cheng Kung University and Academia Sinica, Taiwan

**Tue-O-1-5-5, Time: 11:20-11:40**

Interview is a vital part of recruitment process and is especially challenging for the beginners. In an interactive and natural interview, the interviewers would ask follow-up questions or request further elaborations when they are not satisfied with the interviewee's initial response. In this study, as only a small interview corpus is available, a pattern-based sequence to sequence (Seq2seq) model is adopted for follow-up question generation. First, word clustering is employed to automatically transform the question/answer sentences into sentence patterns, in which each sentence pattern is composed of word classes, to decrease the complexity of the sentence structures. Next, the convolutional neural tensor network (CNTN) is used to select a target sentence in an interviewee's answer turn for follow-up question generation. In order to generate the follow-up question pattern, the selected target sentence pattern is fed to a Seq2seq model to obtain the corresponding follow-up question pattern. Then the word class positions in the generated follow-up question sentence pattern is filled in with the words using a word class table obtained from the training corpus. Finally, the n-gram language model is used to rank the candidate follow-up questions and choose the most suitable one as the response to the interviewee. This study collected 3390 follow-up question and answer sentence pairs for training and evaluation. Five-fold cross validation was employed and the experimental results show that the proposed method outperformed the traditional word-based method and achieved a more favorable performance based on a statistical significance test.

### Coherence Models for Dialogue

*Alessandra Cervone, Evgeny Stepanov and Giuseppe Riccardi*

Signals and Interactive Systems Lab, DISI, University of Trento, Italy

**Tue-O-1-5-6, Time: 11:40-12:00**

Coherence across multiple turns is a major challenge for state-of-the-art dialogue models. Arguably the most successful approach to automatically learning text coherence is the entity grid, which relies on modelling patterns of distribution of entities across multiple sentences of a text. Originally applied to the evaluation of automatic summaries and the news genre, among its many extensions, this model has also been successfully used to assess dialogue coherence.

Nevertheless, both the original grid and its extensions do not model intents, a crucial aspect that has been studied widely in the literature in connection to dialogue structure.

We propose to augment the original grid document representation for dialogue with the intentional structure of the conversation.

Our models outperform the original grid representation on both text discrimination and insertion, the two main standard tasks for coherence assessment across three different dialogue datasets, confirming that intents play a key role in modelling dialogue coherence.

.....

## Tue-SS-1-1: Special Session: Speech Recognition for Indian Languages

MR 1.01-1.02; 10:00-12:00; Tuesday, 4 September, 2018

Chairs: Pedro Moreno, Eugene Weinstein, Sarah Abu Sharkh, Haruko Ishikawa, Daan van Esch, Rita Singh and Preethi Jyothi

.....

### Introduction

**Tue-SS-1-1-1, Time: 10:00-10:10**

### Indian Languages ASR: A Multilingual Phone Recognition Framework with IPA Based Common Phone-set, Predicted Articulatory Features and Feature Fusion

*Manjunath K E[1], K. Sreenivasa Rao[2], Dinesh Babu Jayagopi[1] and V Ramasubramanian[1]*

[1]International Institute of Information Technology - Bangalore (IIIT-B), Bangalore, India

[2]Indian Institute of Technology Kharagpur, Kharagpur, India

**Tue-SS-1-1-2, Time: 10:10-10:15**

In this study, a multilingual phone recognition system for four Indian languages - Kannada, Telugu, Bengali and Odia - is described. International phonetic alphabets are used to derive the transcription. Multilingual Phone Recognition System (MPRS) is developed using the state-of-the-art DNNs. The performance of MPRS is improved using the Articulatory Features (AFs). DNNs are used to predict the AFs for place, manner, roundness, frontness and height AF groups. Further, the MPRS is also developed using oracle AFs and their performance is compared with that of predicted AFs. Oracle AFs are used to set the best performance realizable by AFs predicted from MFCC features by DNNs. In addition to the AFs, we have also explored the use of phone posteriors to further boost the performance of MPRS.We show that oracle AFs by feature fusion with MFCCs offer a remarkably low target of PER of 10.4%, which is 24.7% absolute reduction compared to baseline MPRS with MFCCs alone. The best performing system using predicted AFs has shown 2.8% reduction in absolute PER (8% reduction in relative PER) compared to baseline MPRS.

### Rapid Collection of Spontaneous Speech Corpora Using Telephonic Community Forums

*Agha Ali Raza[1], Awais Athar[2], Shan Randhawa[1], Zain Tariq[1], Muhammad Bilal Saleem[1], Haris Bin Zia [1], Umar Saif[1] and Roni Rosenfeld[3]*

[1]Information Technology University, Lahore, Pakistan

[2]European Bioinformatics Institute (EMBL-EBI), Cambridge, UK

[3]Carnegie Mellon University, Pittsburgh, PA, USA

**Tue-SS-1-1-3, Time: 10:15-10:20**

We present a novel technique for rapid collection of spontaneous speech data over mobile phone channel using telephonic community forums. Our public forum allows users to post audio messages, listen to messages posted by others, post votes and audio comments and share content with friends through subsidized phone calls. The entertainment aspects and sharing features of the forum lead to its viral spread in Pakistan. Within 8 months, it reached 11,017 users and gathered 1,207 hours of speech data comprising 57,454 audio-posts and 130,685 audio-comments, spanning Urdu and 9 regional languages. We trained an ASR using just 9.5 hours of the corpus to obtain 24.19% WER. Community forums automatically overcome common spontaneous speech data collection challenges like speaker recruitment, natural speech elicitation, content diversity, informed consent, sampling real-world ambient noise and reach (for geographically remote linguistic communities). This technique is especially useful for gathering speech corpora for under-resourced languages hence enabling the development of speech recognition, keyword spotting, speaker ID and noise classification systems (among others) for such languages. It also allows rapid, automatic preservation of spoken languages and oral aspects of culture. This technique can be extended to collect speech data for endangered languages, oral cultures and linguistic minorities.

### Effect of TTS Generated Audio on OOV Detection and Word Error Rate in ASR for Low-resource Languages

*Savitha Murthy[1], Dinkar Sitaram[1] and Sunayana Sitaram[2]*

[1]PES University, Bangalore, India

[2]Microsoft Research, Bangalore, India

**Tue-SS-1-1-4, Time: 10:20-10:25**

Out-of-Vocabulary (OOV) detection and recovery is an important aspect of reducing Word Error Rate (WER) in Automatic Speech Recognition (ASR). In this paper, we evaluate the effect on WER for a low-resource language ASR system using OOV detection and recovery. We use a small seed corpus of continuous speech and improve the vocabulary by incorporating the detected OOV words. We use a syllable-model to detect and learn OOV words and, augment the word-model with these words leading to improved

**Notes**

recognition. Our research investigates the effect on OOV detection and recovery after adding missing syllable sounds in the syllable model using a Text-to-Speech (TTS) system. Our experiments are conducted using 5 hours of continuous speech Kannada corpus. We use an already available Festival TTS for Hindi to generate Kannada speech. Our initial experiments report an improvement in OOV detection due to addition of missing syllable sounds using a cross-lingual TTS system.

## Development of Large Vocabulary Speech Recognition System with Keyword Search for Manipuri

*Tanvina Patel, Krishna D N, Noor Fathima, Nisar Shah, Mahima C, Deepak Kumar and Anuroop Iyengar*

Cogknit Semantics, Bangalore, India

**Tue-SS-1-1-5, Time: 10:25-10:30**

Research in Automatic Speech Recognition (ASR) has witnessed a steep improvement in the past decade (especially for English language) where the variety and amount of training data available is huge. In this work, we develop an ASR and Keyword Search (KWS) system for Manipuri, a low-resource Indian Language. Manipuri (also known as Meitei), is a Tibeto-Burman language spoken predominantly in Manipur (a northeastern state of India). We collect and transcribe telephonic read speech data of 90+ hours from 300+ speakers for the ASR task. Both state-of-the-art Gaussian Mixture-Hidden Markov Model (GMM-HMM) and Deep Neural Network-Hidden Markov Model (DNN-HMM) based architectures are developed as a baseline. Using the collected data, we achieve better performance using DNN-HMM systems, i.e., 13.57% WER for ASR and 7.64% EER for KWS. The KALDI speech recognition tool-kit is used for developing the systems. The Manipuri ASR system along with KWS is integrated as a visual interface for demonstration purpose. Future systems will be improved with more amount of training data and advanced forms of acoustic models and language models.

## Robust Mizo Continuous Speech Recognition

*Abhishek Dey[1], Biswajit Dev Sarma[5], Wendy Lalhminghlui[3], Lalnunsiami Ngente[3], Parismita Gogoi[2,3], Priyankoo Sarmah[3], S R Mahadeva Prasanna[3,4], Rohit Sinha[3] and Nirmala S.R. [1]*

[1]GUIST, Gauhati University, Guwahati-781014, India

[2]DUIET, Dibrugarh University, Dibrugarh-786004, India

[3]Indian Institute of Technology Guwahati, Guwahati-781039, India

[4]Indian Institute of Technology Dharwad, Dharwad-580011, India

[5]Kaliber Labs Inc, Guwahati-781039, India

**Tue-SS-1-1-6, Time: 10:30-10:35**

Mizo is an under-resourced tonal language that is mainly spoken in North-East India. It has 4 canonical tones along with a tone-sandhi. In Mizo language, a majority of the words contain tone information. As a result of that, it exhibits higher acoustic variability like other tonal languages in the world. In this work, we investigate the impact of tonal information on robust Mizo continuous speech recognition (CSR). First, separate baseline CSR systems are developed employing the Mel-frequency cepstral coefficient (MFCC) based acoustic features and salient acoustic modeling paradigms. For further improvement, the tonal information has been incorporated in each of the CSR systems. For this purpose, 3-dimensional tonal features are derived which include pitch, pitch-difference and probability of voicing values. Our experimental study reveals that with the inclusion of tonal information, the robustness of Mizo CSR system gets enhanced across all acoustic modeling paradigms. This trend is attributed to lesser degradation in the fundamental frequency information than the vocal tract information under noisy conditions.

## Semi-supervised and Active-learning Scenarios: Efficient Acoustic Model Refinement for a Low Resource Indian Language

*Maharajan Chellapriyadharshini, Anoop Toffy, Srinivasa Raghavan K. M. and V Ramasubramanian*

International Institute of Information Technology Bangalore (IIIT-B), Bangalore, India

**Tue-SS-1-1-7, Time: 10:35-10:40**

We address the problem of efficient acoustic-model refinement (continuous retraining) using semi-supervised and active learning for a low resource Indian language, wherein the low resource constraints are having i) a small labeled corpus from which to train a baseline `seed' acoustic model and ii) a large training corpus without orthographic labeling or from which to perform a data selection for manual labeling at low costs. The proposed semi-supervised learning decodes the unlabeled large training corpus using the seed model and through various protocols, selects the decoded utterances with high reliability using confidence levels (that correlate to the WER of the decoded utterances) and iterative bootstrapping. The proposed active learning protocol uses confidence level based metric to select the decoded utterances from the large unlabeled corpus for further labeling. The semi-supervised learning protocols can offer a WER reduction, from a poorly trained seed model, by as much as 50% of the best WER-reduction realizable from the seed model's WER, if the large corpus were labeled and used for acoustic-model training. The active learning protocols allow that only 60% of the entire training corpus be manually labeled, to reach the same performance as the entire data.

## Automatic Speech Recognition with Articulatory Information and a Unified Dictionary for Hindi, Marathi, Bengali and Oriya

*Debadatta Dash[1], Myungjong Kim[1], Kristin Teplansky[1,2] and Jun Wang[1,2]*

[1]Speech Disorders & Technology Lab, Department of Bioengineering

[2]Callier Center for Communication Disorders, The University of Texas at Dallas, TX, USA

**Tue-SS-1-1-8, Time: 10:40-10:45**

Despite the continuous progress of Automatic Speech recognition (ASR) technologies, these systems for Indian languages are still in infancy stage due to a multitude of challenges involved, including resource deficiency. This paper addressed this challenge with four Indian languages, Hindi, Marathi, Bengali and Oriya by integrating articulatory information into acoustic features, thereby compensating the low resource property of these languages for improved performance. Articulatory movements were recorded during speech production using an electromagnetic articulograph and trained together with acoustic features to build automatic speech recognizers for these languages. Both speaker-dependent and -independent recognition experiments were conducted by adopting three ASR models: Gaussian Mixture Model (GMM)-Hidden Markov Model (HMM), Deep Neural Network (DNN)-HMM and Long Short Term Memory recurrent neural network (LSTM)-HMM. A cross-language similarity was discerned in both acoustic and articulatory domains in the pairs of Oriya-Bengali and Hindi-Marathi. Based on these observations, a multi-lingual, multi-modal speech recognizer was built by constructing a unified dictionary consisting of common and unique phonemes of all the four languages, which reduced the phoneme error rates.

## Poster Session

**Tue-SS-1-1-9, Time: 10:45-11:30**

## Discussion

**Tue-SS-1-1-10, Time: 11:30-12:00**

**Notes**

## Tue-S&T-1-1: Show and Tell 3

MR G.05-G.06; 10:00-12:00; Tuesday, 4 September, 2018

Chairs: Laxminarayana P and Niranjan Nayak

### Captaina: Integrated Pronunciation Practice and Data Collection Portal

*Aku Rouhe[1], Reima Karhila[1], Aija Elg[1], Minnaleena Toivola[2], Peter Smit[1], Anna-Riikka Smolander[2] and Mikko Kurimo[1]*

[1]Aalto University

[2]Helsinki University

**Tue-S&T-1-1-1, Time: 10:00-12:00**

We demonstrate Captaina, computer assisted pronunciation training portal. It is aimed at university students, who read passages aloud and receive automatic feedback based on speech recognition and phoneme classification. Later their teacher can provide more accurate feedback and comments through the portal.

The system enables better independent practice. It also acts as a data collection method. We aim to gather both good quality second language speech data with segmentations and the teacher given evaluations of pronunciation.

### auMina™ - Enterprise Speech Analytics

*Umesh Sachdev, Rajagopal Jayaraman and Zainab Millwala*

Uniphore Software Systems, Chennai, India

**Tue-S&T-1-1-2, Time: 10:00-12:00**

This paper gives an overview of a commercially viable product, auMina™ – an Enterprise Speech Analytics solution. It details out the features and capabilities of the product and the different business outcomes which can be derived from this Speech Analytics solution.

### HoloCompanion: An MR Friend for EveryOne

*Annam Naresh[1], Rushabh Gandhi[2], Mallikarjuna Rao Bellamkonda[1] and Mithun Das Gupta[1]*

[1]Microsoft, Hyderabad, India

[2]BITS, Hyderabad, India

**Tue-S&T-1-1-3, Time: 10:00-12:00**

Chat bots are becoming ubiquitous in our day to day life. The advent of the summer of AI has brought us all in close contact with intelligent agents such as Cortana, Siri and Alexa. We envisage a world, where these bots have there physical existence within the realm of Mixed Reality (MR). We present the first 3D chit-chat bot called the HoloCompanion. This bot has a personality, can chat with anyone and about any topic and has articulated lip, eye and head movements.

### akeira™ - Virtual Assistant

*Umesh Sachdev, Rajagopal Jayaraman and Zainab Millwala Bellamkonda[1] and Mithun Das Gupta[1]*

Uniphore Software Systems, Chennai, India

**Tue-S&T-1-1-4, Time: 10:00-12:00**

This paper gives an overview of a commercially viable product, akeira™ – Voice Virtual Assistant. It also details out the features, capabilities and benefits of this intuitive Virtual Assistant.

### Brain-Computer Interface using Electroencephalogram Signatures of Eye Blinks

*Srihari Maruthachalam[1], Sidharth Aggarwal[1], Mari Ganesh Kumar[1], Mriganka Sur[2] and Hema Murthy[1]*

[1]Department of Computer Science and Engineering, Indian Institute of Technology, Madras

[2]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge

**Tue-S&T-1-1-5, Time: 10:00-12:00**

The objective of this work is to develop a personalized eye blink based communicator device. The eye blink is detected using a single channel Electroencephalogram (EEG) system with a Bluetooth interface. Eye blinks have predominant signatures in EEG. Different patterns based on these signatures are used to map alphanumeric characters on a virtual keyboard and words are generated. Voice module is incorporated into the app on the android device for better accessibility of the device by including Text To Speech synthesizer (TTS) at the back-end to produce speech output.

## Tue-P-1-1: Speaker Verification II

Hall 4-6: Poster-1; 10:00-12:00; Tuesday, 4 September, 2018

Chair: Sriram Ganapathy

### Voice Comparison and Rhythm: Behavioral Differences between Target and Non-target Comparisons

*Moez Ajili[1], Jean-François Bonastre[1] and Solange Rossato[2]*

[1]LIA-CERI, University Of Avignon, Avignon, France

[2]LIG, Univ.Grenoble.Alpes, Grenoble, France

**Tue-P-1-1-1, Time: 10:00-12:00**

It is common to see voice recordings being presented as a forensic trace in court. Generally, a forensic expert is asked to analyze both suspect and criminal's voice samples in order to indicate whether the evidence supports the prosecution (same-speaker) or defence (different-speakers) hypotheses. This process is known as Forensic Voice Comparison (FVC). Since the emergence of the DNA typing model, the likelihood-ratio (LR) framework has become the golden standard in forensic sciences. The LR not only supports one of the hypotheses but also quantifies the strength of its support. However, the LR accepts somepractical limitations due to its estimation process itself. It is particularly true when Automatic Speaker Recognition (ASpR) systems are considered as they are outputting a score in all situations regardless of the case specific conditions. Indeed, several factors are not taken into account by the estimation process like the quality and quantity of information in both voice recordings, their phonological content or also the speakers intrinsic characteristics. In our recent study, we showed the importance of the phonemic content and we highlighted interesting differences between inter-speakers effects and intra-speaker's ones.

In this article, we wish to take our previous analysis a step farther and investigate the impact of rhythm variation separately on target and non-target trials.

### Co-whitening of I-vectors for Short and Long Duration Speaker Verification

*Longting Xu[1], Kong Aik Lee[2], Haizhou Li[1] and Zhen Yang[3]*

[1]Department of Electrical and Computer Engineering, National University of Singapore

[2]Data Science Research Laboratories, NEC Corporation, Japan

[3]Broadband Wireless Communication and Sensor Network Technology Key Lab, Nanjing University of Posts and Telecommunications, China

**Tue-P-1-1-2, Time: 10:00-12:00**

An i-vector is a fixed-length and low-rank representation of a speech utterance. It has been used extensively in text-independent speaker verification. Ideally, speech utterances from the same speaker would map to an unique i-vector. However, this is not the case due to some intrinsic and extrinsic factors like physical condition of the speaker, channel difference, noise and notably the duration of speech utterances. In particular, we found

**Notes**

that i-vectors extracted from short utterances exhibit larger variance than that of long utterances. To address the problem, we propose a co-whitening approach, taking into account the duration, while maximizing the correlation between the i-vectors of short and long duration. The proposed co-whitening method was derived based on canonical correlation analysis (CCA). Experimental results on NIST SRE 2010 show that co-whitening method is effective in compensating the duration mismatch, leading to a reduction of up to 13.07% in equal error rate (EER).

## Compensation for Domain Mismatch in Text-independent Speaker Recognition

*Fahimeh Bahmaninezhad and John H.L. Hansen*

Center for Robust Speech Systems (CRSS), University of Texas at Dallas, Richardson, TX 75080

**Tue-P-1-1-3, Time: 10:00-12:00**

Domain mismatch continues to be a major research challenge for speaker recognition in naturalistic audio streams. This study presents a new technique for domain mismatch compensation within a text-independent speaker recognition scenario. The proposed method is designed for the NIST speaker recognition evaluation 2016 (SRE16) task, where speakers from training, development and evaluation data belong to different sets of languages. An i-vector/PLDA speaker recognition system is adopted for this study. To address the mismatch problem, we propose to append auxiliary features to the i-vectors. These auxiliary features are adapted representations of the i-vectors to the specific in-domain data; therefore, the new feature vector has two parts: (1) i-vectors which represent speaker identity and (2) auxiliary features which are representations of i-vectors in the in-domain data feature space (and may not contain speaker identity information). This new concatenated feature vector (we call this a-vector) is then post-processed with support vector discriminant analysis (SVDA) for further domain compensation. Evaluations based on the SRE16 confirm the effectiveness of the proposed technique. In terms of minimum Cprimary cost, a-vector outperforms the i-vector consistently. Moreover, comparing to previous single systems introduced for SRE16, we achieved 8.5%-18% improvements in terms of equal error rate.

## Joint Learning of J-Vector Extractor and Joint Bayesian Model for Text Dependent Speaker Verification

*Ziqiang Shi, Liu Liu, Huibin Lin and Rujie Liu*

Fujitsu Research and Development Center

**Tue-P-1-1-4, Time: 10:00-12:00**

J-vector and joint Bayesian have been proved to be very effective in text dependent speaker verification with short-duration speech. However current state-of-the-art framework often consider training the J-vector extractor and the joint Bayesian classifier separately. Such an approach will result in information loss for j-vector learning and also fail to exploit an end-to-end framework. In this paper we present a integrated approach to text dependent speaker verification, which consists of a siamese deep neural network that takes two variable length speech segments and maps them to the likelihood score and speaker/phrase labels, where the likelihood score as a loss guide is computed by a variant joint Bayesian model. The likelihood loss guide can constrain the j-vector extractor for improving the verification performance. Since the strengths of j-vector and joint Bayesian analysis appear complementary the joint learning significantly outperforms traditional separate training scheme. Our experiments on the the public RSR2015 part I data corpus demonstrate that this new training scheme can produce more discriminative j-vectors and leading to performance improvement on the speaker verification task.

## Latent Factor Analysis of Deep Bottleneck Features for Speaker Verification with Random Digit Strings

*Ziqiang Shi, Huibin Lin, Liu Liu and Rujie Liu*

Fujitsu Research and Development Center

**Tue-P-1-1-5, Time: 10:00-12:00**

Speaker verification with prompted random digit strings has been a challenging task due to very short test utterance. This work investigates how to combine methods from deep bottleneck features (DBF) and latent factor analysis (LFA) to result in a new state-of-the-art approach for such task. In order to provide a wider temporal context, a stacked DBF is extracted to replace the traditional MFCC feature in the derivation of the supervector representations and leads to a significant improvement for the speaker verification. The LFA is used to model these stacked DBFs in both digit and utterance scales. Based on this learned LFA model, two kinds of supervector representations are extracted for utterance and local digits respectively.

Since the strengths of DBF and LFA appear complementary, the combination significantly outperforms either of its components.
Experiments have been conducted on the public RSR2015 part III data corpus, the results showed that our approach can achieve 1.40% EER and 1.55% EER on male and female respectively.

## VoxCeleb2: Deep Speaker Recognition

*Joon Son Chung, Arsha Nagrani and Andrew Zisserman*

Visual Geometry Group, Department of Engineering Science, University of Oxford, UK

**Tue-P-1-1-6, Time: 10:00-12:00**

The objective of this paper is speaker recognition under noisy and unconstrained conditions.

We make two key contributions. First, we introduce a very large-scale speaker recognition dataset collected from open-source media. Using a fully automated pipeline, we curate VoxCeleb2 which contains over a million utterances from over 6,000 speakers. This is several times larger than any publicly available speaker recognition dataset.

Second, we develop and compare Convolutional Neural Network (CNN) models and training strategies that can effectively recognise identities from voice under various conditions.

The models trained on the VoxCeleb2 dataset surpass the performance of previous works on a benchmark dataset by a significant margin.

## Supervised I-vector Modeling - Theory and Applications

*Shreyas Ramoji and Sriram Ganapathy*

Learning and Extraction of Acoustic Patterns (LEAP) Lab, Electrical Engineering, Indian Institute of Science, Bengaluru, India

**Tue-P-1-1-7, Time: 10:00-12:00**

Over the last decade, the factor analysis based modeling of a variable length speech utterance into a fixed dimensional vector (termed as i-vector) has been prominently used for many tasks like speaker recognition, language recognition and even in speech recognition. The i-vector model is an unsupervised learning paradigm where the data is initially clustered using a Gaussian Mixture Universal Background Model (GMM-UBM). The adapted means of the Gaussian mixture components are dimensionality reduced using the Total Variability Matrix (TVM) where the latent variables are modeled with a single Gaussian distribution. In this paper, we propose to rework the theory of i-vector modeling using a supervised framework where the speech utterances are associated with a label. Class labels are introduced in the i-vector model using a mixture Gaussian prior. We show that the proposed model is a generalized i-vector model and the conventional i-vector model turns out to be a special case of this model. This model is applied for a language recognition task using the NIST Language Recognition Evaluation (LRE) 2017 dataset. In these experiments, the supervised i-vector model provides significant improvements over the conventional i-vector model (average relative improvements of 5% in terms of C_{avg}.

## LOCUST - Longitudinal Corpus and Toolset for Speaker Verification

*Evgeny Dmitriev[1], Yulia Kim[2], Anastasia Matveeva2, Claude Montacié[1], Yannick Boulard[4], Yadviga Sinyavskaya[3], Yulia Zhukova[2], Adam Zarazinski[5], Egor Akhanov[5], Ilya Viksnin[2], Andrei Shlykov[2] and Maria Usova[2]*

[1]Sorbonne University, France

**Notes**

[2]ITMO University, Russia
[3]Higher School of Economics, Russia
[4]SynAck Conseil, France
[5]Inca Digital Securities, USA

**Tue-P-1-1-8, Time: 10:00-12:00**

In this paper, we set forth a new longitudinal corpus and a toolset in an effort to address the influence of voice-aging on speaker verification.

We have examined previous longitudinal research of age-related voice changes as well as its applicability to real world use cases. Our findings reveal that scientists have treated age-related voice changes as a hindrance instead of leveraging it to the advantage of the identity validator. Additionally, we found a significant dearth of publicly available corpora related to both the time span of and the number of participants in audio recordings. We also identified a significant bias toward the development of speaker recognition technologies applicable to government surveillance systems compared to speaker verification systems used in civilian IT security systems.

To solve the aforementioned issues, we built an open project with the largest publicly available longitudinal speaker database, which includes 229 speakers with an average talking time exceeding 15 hours spanning across an average of 21 years per speaker. We assembled, cleaned and normalized audio recordings and developed software tools for speech features extractions, all of which we are releasing to the public domain.

## Analysis of Language Dependent Front-End for Speaker Recognition

*Srikanth Madikeri[1], Subhadeep Dey[1,2] and Petr Motlicek[1]*
[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

**Tue-P-1-1-9, Time: 10:00-12:00**

In Deep Neural Network (DNN) i-vector based speaker recognition systems, acoustic models trained for Automatic Speech Recognition are employed to estimate sufficient statistics for i-vector modeling.

The DNN based acoustic model is typically trained on a well-resourced language like English. In evaluation conditions where the enrollment and test data are not in English, as in the NIST SRE 2016 dataset, a DNN acoustic model generalizes poorly. In such conditions, a conventional Universal Background Model/Gaussian Mixture Model (UBM/GMM) based i-vector extractor performs better than the DNN based i-vector system.

In this paper, we address the scenario in which one can develop a Automatic Speech Recognizer with limited resources for a language present in the evaluation condition, thus enabling the use of a DNN acoustic model instead of UBM/GMM.

Experiments are performed on the Tagalog subset of the NIST SRE 2016 dataset assuming an open training condition.

With a DNN i-vector system trained for Tagalog, a relative improvement of 12.1% is obtained over a baseline system trained for English.

## Robust Speaker Recognition from Distant Speech under Real Reverberant Environments Using Speaker Embeddings

*Mahesh Kumar Nandwana, Julien van Hout , Mitchell McLaren, Allen Stauffer, Colleen Richey, Aaron Lawson and Martin Graciarena*
Speech Technology and Research Laboratory, SRI International, Menlo Park, California, CA

**Tue-P-1-1-10, Time: 10:00-12:00**

This article focuses on speaker recognition using speech acquired using a single distant or far-field microphone in an indoors environment. This study differs from the majority of speaker recognition research, which focuses on speech acquisition over short distances, such as when using a telephone handset or mobile device or far-field microphone arrays, for which beamforming can enhance distant speech signals. We use two large-scale corpora collected by retransmitting speech data in reverberant environments with multiple microphones placed at different distances. We first characterize three different speaker recognition systems ranging from a traditional universal background model (UBM) i-vector system to a state-of-the-art deep neural network (DNN) speaker embedding system with a probabilistic linear discriminant analysis (PLDA) back-end. We then assess the impact of microphone distance and placement, background noise and loudspeaker orientation on the performance of speaker recognition system for distant speech data. We observe that the recently introduced DNN speaker embedding based systems are far more robust compared to i-vector based systems, providing a significant relative improvement of up to 54% over the baseline UBM i-vector system and 45.5% over prior DNN-based speaker recognition technology.

## Investigation on Bandwidth Extension for Speaker Recognition

*Phani Sankar Nidadavolu, Cheng-I Lai, Jesús Villalba and Najim Dehak*
Center for Language and Speech Processing, The Johns Hopkins University

**Tue-P-1-1-11, Time: 10:00-12:00**

In this work, we investigate training speaker recognition systems on wideband (WB) features and compare their performance with narrowband (NB) baselines. NIST speaker recognition evaluations have mainly driven speaker recognition research in the past years. Because of the target application of these evaluations, most data available to train speaker recognition systems is NB telephone speech. Meanwhile, WB data have been more scarce not being enough to train factor analysis and PLDA models. Thus, the usual practice when dealing with WB speech consists in downsampling the signal to 8 kHz, which implies potential loss of useful information. Instead, we experimented upsampling the training telephone data and leaving the WB data unchanged. We adopt two techniques to upsample telephone data: (1) using a feed-forward neural network, termed Bandwidth Extension (BWE) network, to predict WB features given NB features as input; and (2) using basic upsampling with a low-pass filter interpolator. While the former intends to estimate the high frequency information, the latter does not. The upsampled features are used to train state-of-the art i-vector and recently proposed x-vector models. We evaluated the systems on Speakers In The Wild (SITW) database obtaining 11.5% relative improvement in detection cost function (DCF) with x-vector model.

## On Learning Vocal Tract System Related Speaker Discriminative Information from Raw Signal Using CNNs

*Hannah Muckenhirn[1,2], Mathew Magimai Doss[1] and Sebastien Marcel[1]*
[1]Idiap Research Institute, Martigny, Switzerland
[2]École polytechnique fédérale de Lausanne, Switzerland

**Tue-P-1-1-12, Time: 10:00-12:00**

In a recent work, we have shown that speaker verification systems can be built where both features and classifiers are directly learned from the raw speech signal with convolutional neural networks (CNNs). In this framework, the training phase also decides the block processing through cross validation. It was found that the first convolution layer, which processes about 20 ms speech, learns to model fundamental frequency information.

In the present paper, inspired from speech recognition studies, we build further on that framework to design a CNN-based system, which models sub-segmental speech (about 2ms speech) in the first convolution layer, with an hypothesis that such a system should learn vocal tract system related speaker discriminative information. Through experimental studies on Voxforge corpus and analysis on American vowel dataset, we show that the proposed system (a) indeed focuses on formant regions, (b) yields competitive speaker verification system and (c) is complementary to the CNN-based system that models fundamental frequency information.

**Notes**

### On Convolutional LSTM Modeling for Joint Wake-Word Detection and Text Dependent Speaker Verification

*Rajath Kumar[1], Vaishnavi Yeruva[2] and Sriram Ganapathy[2]*

[1]Department of Electrical Engineering, Columbia University, New York, NY

[2]Learning and Extraction of Acoustic Pattern Lab, Indian Institute of Science

**Tue-P-1-1-13, Time: 10:00-12:00**

The task of personalized keyword detection system which also performs text dependent speaker verification (TDSV) has received substantial interest recently. Conventional approaches to this task involve the development of the TDSV and wake-up-word detection systems separately. In this paper, we show that TDSV and keyword spotting (KWS) can be jointly modeled using the convolutional long short term memory (CLSTM) model architecture, where an initial convolutional feature map is further processed by a LSTM recurrent network. Given a small amount of training data for developing the CLSTM system, we show that the model provides accurate detection of the presence of the keyword in spoken utterance. For the TDSV task, the MTL model can be well regularized using the CLSTM training examples for personalized wake up task. The experiments are performed for KWS wake up detection and TDSV using the combined speech recordings from Wall Street Journal (WSJ) and LibriSpeech corpus. In these experiments with multiple keywords, we illustrate that the proposed approach of MTL significantly improves the performance of previously proposed neural network based text dependent SV systems. We also experimentally illustrate that the CLSTM model provides significant improvements over previously proposed keyword detection systems as well (average relative improvements of 30% over previous approaches).

### Cosine Metric Learning for Speaker Verification in the I-vector Space

*Zhongxin Bai, Xiao-Lei Zhang and Jingdong Chen*

Center for Intelligent Acoustics and Immersive Communications and School of Marine Science and Technology, Northwestern Polytechnical University

**Tue-P-1-1-14, Time: 10:00-12:00**

It is known that the equal-error-rate (EER) performance of a speaker verification system is determined by the overlap region of the decision scores of true and imposter trials. Also, the cosine similarity scores of the true or imposter trials produced by the state-of-the-art i-vector front-end approximate to a Gaussian distribution and the overlap region of the two classes of trials depends mainly on their between-class distance. Motivated by the above facts, this paper presents a cosine similarity learning (CML) framework for speaker verification, which combines classical compensation techniques and the cosine similarity scoring for improving the EER performance. CML minimizes the overlap region by enlarging the between-class distance while introducing a regularization term to control the within-class variance, which is initialized by a traditional channel compensation technique such as linear discriminant analysis. Experiments are carried out to compare the proposed CML framework with several traditional channel compensation baselines on the NIST speaker recognition evaluation data sets. The results show that CML outperforms all the studied initialization compensation techniques.

### An Unsupervised Neural Prediction Framework for Learning Speaker Embeddings Using Recurrent Neural Networks

*Arindam Jati and Panayiotis Georgiou*

University of Southern California, Los Angeles, CA, USA

**Tue-P-1-1-15, Time: 10:00-12:00**

This paper presents an unsupervised training framework for learning a speaker-specific embedding using a Neural Predictive Coding (NPC) technique. We employ a Recurrent Neural Network (RNN) trained on unlabeled audio with multiple and unknown speaker change points. We assume short-term speaker stationarity and hence that speech frames in close temporal proximity originated from a single speaker. In contrast, two random short speech segments from different audio streams are assumed to originate from two different speakers. Based on this hypothesis, a binary classification scenario of predicting whether an input pair of short speech segments comes from the same speaker or not, is developed. An RNN based deep siamese network is trained and the resulting embeddings, extracted from a hidden layer representation of the network, are employed as speaker embeddings. The experimental results on speaker change points detection show the efficacy of the proposed method to learn short-term speaker-specific features. We also show the consistency of these features via a simple statistics-based utterance-level speaker classification task. The proposed method outperforms the MFCC baseline for speaker change detection and both MFCC and i-vector baselines for speaker classification.

## Tue-P-1-2: Novel Approaches to Enhancement

Hall 4-6: Poster-2; 10:00-12:00; Tuesday, 4 September, 2018

Chair: Rajbabu Velmurugan

### A New Framework for Supervised Speech Enhancement in the Time Domain

*Ashutosh Pandey[1] and DeLiang Wang[1,2]*

[1]Department of Computer Science and Engineering, The Ohio State University, USA

[2]Center for Cognitive and Brain Sciences, The Ohio State University, USA

**Tue-P-1-2-1, Time: 10:00-12:00**

This work proposes a new learning framework that uses a loss function in the frequency domain to train a convolutional neural network (CNN) in the time domain. At the training time, an extra operation is added after the speech enhancement network to convert the estimated signal in the time domain to the frequency domain. This operation is differentiable and is used to train the system with a loss in the frequency domain. This proposed approach replaces learning in the frequency domain, i.e., short-time Fourier transform (STFT) magnitude estimation, with learning in the original time domain. The proposed method is a spectral mapping approach in which the CNN first generates a time domain signal then computes its STFT that is used for spectral mapping. This way the CNN can exploit the additional domain knowledge about calculating the STFT magnitude from the time domain signal. Experimental results demonstrate that the proposed method substantially outperforms the other methods of speech enhancement. The proposed approach is easy to implement and applicable to related speech processing tasks that require spectral mapping or time-frequency (T-F) masking.

### Speech Enhancement Using the Minimum-probability-of-error Criterion

*Jishnu Sadasivan[1], Subhadip Mukherjee[2] and Chandra Sekhar Seelamantula[2]*

[1]Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore 560012, India

[2]Department of Electrical Engineering, Indian Institute of Science, Bangalore 560012, India

**Tue-P-1-2-2, Time: 10:00-12:00**

We propose a novel speech denoising framework by minimizing the probability of error (PE), which measures the deviation probability of the estimate from its true value. To develop the minimum PE (MPE) criterion, one requires the knowledge of the noise probability density function (p.d.f.), which may not be available in a parametric form in speech denoising applications. Therefore, we adopt two approaches for modeling the noise p.d.f.: (i) Gaussian modeling based on adaptive variance estimation; and (ii) a Gaussian mixture model (GMM) in view of its approximation capabilities.

**Notes**

We consider discrete cosine transform (DCT) domain shrinkage, where the optimum shrinkage parameter is obtained by minimizing an estimate of the PE. A performance assessment for real-world noise types shows that for input signal-to-noise ratios (SNR) greater than 5 dB, the proposed MPE-based point-wise shrinkage estimators outperform three benchmark techniques in terms of segmental SNR and short-time objective intelligibility (STOI) scores.

## Exploring the Relationship between Conic Affinity of NMF Dictionaries and Speech Enhancement Metrics

*Pavlos Papadopoulos, Colin Vaz and Shrikanth Narayanan*

Signal Analysis and Interpretation Lab, University of Southern California, USA

**Tue-P-1-2-3, Time: 10:00-12:00**

Nonnegative Matrix Factorization(NMF) has been successfully used in speech enhancement. In the training phase NMF produces speech and noise dictionaries, whose elements are non-negative, while in the testing phase it estimates a non-negative activation matrix to express the enhanced speech signal as a conic combination of those dictionaries. This nonnegativity property enables us to interpret them as convex polyhedral cones that lie in the positive orthant. Conic affinity could be useful when designing NMF-based systems for unseen noise conditions, which operate by selecting an appropriate noise dictionary amongst a pool of potential candidates. To that end, we examine two conic affinity measures, one based on cosine similarity, while the other is based on euclidean distance from a point to a cone. Moreover, we construct an algorithm to show that conic affinity correlates with speech enhancement performance metrics.

## Using Shifted Real Spectrum Mask as Training Target for Supervised Speech Separation

*Yun Liu, Hui Zhang and Xueliang Zhang*

Inner Mongolia Key Laboratory of Mongolian Information Processing Technology, Inner Mongolia University, Hohhot, China

**Tue-P-1-2-4, Time: 10:00-12:00**

Deep learning-based speech separation has been widely studied in recent years. Most of these kind approaches focus on recovering the magnitude spectrum of the target speech, but ignore the phase estimation. Recently, a method called shifted real spectrum (SRS) is proposed. Unlike the short-time Fourier transform (STFT), the SRS contains only real components which encode the phase information. In this paper, we propose several SRS-based masks and use them as the training target of deep neural networks. Experimental results show that the proposed target outperforms the commonly used masks computed on STFT in general.

## Enhancement of Noisy Speech Signal by Non-Local Means Estimation of Variational Mode Functions

*Nagapuri Srinivas, Gayadhar Pradhan and Syed Shahnawazuddin*

Department of Electronics and Communication Engineering National Institute of Technology Patna, India

**Tue-P-1-2-5, Time: 10:00-12:00**

In this paper, a speech enhancement approach exploiting the efficacy of non-local means (NLM) estimation and variational mode decomposition (VMD) is proposed. The NLMestimation is effective in removing noises whenever non-local similarities are present among the samples of the signal under consideration. However, it suffers from the issue of under-averaging in those regions where amplitude and frequency variations are abrupt. Since speech is a non-stationary signal, the magnitude and frequency vary over the time. Consequently, NLM is not that effective in removing the noise components from the speech signal as observed in the case of image enhancement. To address this issue, the noisy speech signal is first decomposed into variational mode functions (VMFs) using VMD. Each of the VMFs represents a small portion of the overall frequency components of the signal. The VMFs are then combined into different groups depending on their similarities to reduce computational cost. Next, the non-local similarity present in each group of VMFs is exploited for an effective speech enhancement through NLM estimation. The enhancement performance of the proposed method is compared with two existing speech enhancement techniques. The experimental results presented in this study show that, the proposed method provides better speech enhancement performance.

## Phase-locked Loop (PLL) Based Phase Estimation in Single Channel Speech Enhancement

*Priya Pallavi and Ch V Rama Rao*

Department of E C E, National Institute of Technology Meghalaya

**Tue-P-1-2-6, Time: 10:00-12:00**

Conventional speech enhancement techniques are based on the modification of noisy spectral magnitude. In the reconstruction of the enhanced signal, noisy phase is combined with the modified noisy spectral magnitude. Recent studies on the importance of phase in enhancement process shows that the clean speech phase improves the quality of the enhanced signal. This work focused on Phase-Locked Loop (PLL) based time-domain approach for estimating the clean speech phase from noisy speech signal. The proposed technique is compared with the conventional approaches where noisy phase is used in the reconstruction of the enhanced signal. Here, Log-Likelihood Ratio (LLR), Weighted Spectral Slope (WSS) distance and Perceptual Evaluation of Speech Quality (PESQ) are used as performance measures. From experimental results, it is observed that the speech quality and intelligibility improved significantly with the proposed method over existing methods.

## Cycle-Consistent Speech Enhancement

*Zhong Meng[1,2], Jinyu Li[1], Yifan Gong[1] and Biing-Hwang (Fred) Juang[2]*

[1]Microsoft AI and Research, Redmond, WA, USA

[2]Georgia Institute of Technology, Atlanta, GA, USA

**Tue-P-1-2-7, Time: 10:00-12:00**

Feature mapping using deep neural networks is an effective approach for single-channel speech enhancement. Noisy features are transformed to the enhanced ones through a mapping network and the mean square errors between the enhanced and clean features are minimized. In this paper, we propose a cycle-consistent speech enhancement (CSE) in which an additional inverse mapping network is introduced to reconstruct the noisy features from the enhanced ones. A cycle-consistent constraint is enforced to minimize the reconstruction loss. Similarly, a backward cycle of mappings is performed in the opposite direction with the same networks and losses. With cycle-consistency, the speech structure is well preserved in the enhanced features while noise is effectively reduced such that the feature mapping network generalizes better to unseen data. In cases where only unparalleled noisy and clean data is available for training, two discriminator networks are used to distinguish the reconstructed clean and noisy features from the real ones. The discrimination losses are jointly optimized with reconstruction losses through adversarial multi-task learning. Evaluated on the CHiME-3 dataset, the proposed CSE achieves 19.60% and 6.69% relative word error rate improvements respectively when using or without using parallel clean and noisy speech data.

## Visual Speech Enhancement

*Aviv Gabbay, Asaph Shamir and Shmuel Peleg*

School of Computer Science and Engineering The Hebrew University of Jerusalem, Jerusalem, Israel

**Tue-P-1-2-8, Time: 10:00-12:00**

When video is shot in noisy environment, the voice of a speaker seen in the video can be enhanced using the visible mouth movements, reducing background noise.

While most existing methods use audio-only inputs, improved performance is obtained with our visual speech enhancement, based on an audio-visual neural network.

We include in the training data videos to which we added the voice of the target speaker as background noise. Since the audio input is not sufficient to separate the voice of a speaker from his own voice, the trained model better exploits the visual input and generalizes well to different noise types.

The proposed model outperforms prior audio visual methods on two public lipreading datasets. It is also the first to be demonstrated on a dataset not designed for lipreading, such as the weekly addresses of Barack Obama.

**Notes**

### Implementation of Digital Hearing Aid as a Smartphone Application

*Saketh Sharma, Nitya Tiwari and Prem C. Pandey*
Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, India

**Tue-P-1-2-9, Time: 10:00-12:00**

Hearing aids for persons with sensorineural hearing loss aim to compensate for degraded speech perception caused by frequency-dependent elevation of hearing thresholds, reduced dynamic range, abnormal loudness growth and increased temporal and spectral masking. A digital hearing aid is implemented as a smartphone application as an alternative to ASIC-based hearing aids. The implementation provides user-configurable processing for background noise suppression and dynamic range compression. Speech enhancement technique using spectral subtraction based on geometric approach and noise spectrum estimation based on dynamic quantile tracking is implemented to improve speech perception. To compensate for reduced dynamic range and frequency-dependent elevation of hearing thresholds, a sliding-band dynamic range compression technique is used. Both processing blocks are implemented for real-time processing using single FFT-based analysis-synthesis. Implementation as a smartphone application has been carried out using Nexus 5X with Android 7.1 Nougat OS. A touch-controlled graphical user interface enables the user to fine tune the processing parameters in an interactive and real-time mode. The audio latency is 45 ms, making it suitable for face-to-face communication.

### Bone-Conduction Sensor Assisted Noise Estimation for Improved Speech Enhancement

*Ching-Hua Lee, Bhaskar D. Rao and Harinath Garudadri*
Department of Electrical and Computer Engineering
University of California, San Diego

**Tue-P-1-2-10, Time: 10:00-12:00**

State-of-the-art noise power spectral density (PSD) estimation techniques for speech enhancement utilize the so-called speech presence probability (SPP). However, in highly non-stationary environments, SPP-based techniques could still suffer from inaccurate estimation, leading to significant amount of residual noise or speech distortion. In this paper, we propose to improve speech enhancement by deploying the bone-conduction (BC) sensor, which is known to be relatively insensitive to the environmental noise compared to the regular air-conduction (AC) microphone. A strategy is suggested to utilized the BC sensor characteristics for assisting the AC microphone in better SPP-based noise estimation. To our knowledge, no previous work has incorporated the BC sensor in this noise estimation aspect. Consequently, the proposed strategy can possibly be combined with other BC sensor assisted speech enhancement techniques. We show the feasibility and potential of the proposed method for improving the enhanced speech quality by both objective and subjective tests.

### Artificial Bandwidth Extension with Memory Inclusion Using Semi-supervised Stacked Auto-encoders

*Pramod Bachhav, Massimiliano Todisco and Nicholas Evans*
EURECOM, Sophia Antipolis, France

**Tue-P-1-2-11, Time: 10:00-12:00**

Artificial bandwidth extension (ABE) algorithms have been developed to improve quality when wideband devices receive speech signals from narrowband devices or infrastructure. The utilisation of contextual information in the form of dynamic features or explicit memory captured from neighbouring frames is common to ABE research, however the use of additional cues augments complexity and can introduce latency. Previous work shows that unsupervised, linear dimensionality reduction techniques help to reduce complexity. This paper reports a semisupervised, non-linear approach to dimensionality reduction using a stacke auto-encoder. In further contrast to previous work, it operates on raw spectra from which a low dimensional narrowband representation is learned in a data-driven manner. Three different objective speech quality measures show that the new features can be used with a standard regression model to improve ABE performance. Improvements in the mutual information between learned features and missing higher frequency components are also observed whereas improvements in speech quality are corroborated by informal listening tests.

### Large Vocabulary Concatenative Resynthesis

*Soumi Maiti[1], Joey Ching[2] and Michael Mandel[1,2]*
[1]Computer Science Program, the Graduate Center, City University of New York
[2]Computer and Information Science, Brooklyn College, City University of New York

**Tue-P-1-2-12, Time: 10:00-12:00**

Traditional speech enhancement systems reduce noise by modifying the noisy signal, which suffer from two problems: under-suppression of noise and over-suppression of speech.

As an alternative, in this paper, we use the recently introduced concatenative resynthesis approach where we replace the noisy speech with its clean resynthesis.

The output of such a system can produce speech that is both noise-free and high quality. This paper generalizes our previous small-vocabulary system to large vocabulary. To do so, we employ efficient decoding techniques using fast approximate nearest neighbor (ANN) algorithms. Firstly, we apply ANN techniques on the original small vocabulary task and get 5X speedup. We then apply the techniques to the construction of a large vocabulary concatenative resynthesis system and scale the system up to 12X larger dictionary. We perform listening tests with five participants to measure subjective quality and intelligibility of the output speech.

### Concatenative Resynthesis with Improved Training Signals for Speech Enhancement

*Ali Raza Syed[1], Viet Anh Trinh[1] and Michael Mandel[1,2]*
[1]The Graduate Center, CUNY, New York, NY, USA
[2]Brooklyn College, CUNY, New York, NY, USA

**Tue-P-1-2-13, Time: 10:00-12:00**

Noise reduction in speech signals remains an important area of research with potential for high impact in speech processing domains such as voice communication and hearing prostheses. We extend and demonstrate significant improvements to our previous work in synthesis-based speech enhancement, which performs concatenative resynthesis of speech signals for the production of noiseless, high quality speech. Concatenative resynthesis methods perform unit selection through learned non-linear similarity functions between short chunks of clean and noisy signals. These mappings are learned using deep neural networks (DNN) trained to predict high similarity for the exact chunk of speech that is contained within a chunk of noisy speech and low similarity for all other pairings. We find here that more robust mappings can be learned with a more efficient use of the available data by selecting pairings that are not exact matches, but contain similar clean speech that matches the original in terms of acoustic, phonetic and prosodic content. The resulting output is evaluated on the small vocabulary CHiME2-GRID corpus and outperforms our original baseline system in terms of intelligibility by combining phonetic similarity with similarity of acoustic intensity, fundamental frequency and periodicity.

## Tue-P-1-3: Syllabification, Rhythm and Voice Activity Detection

Hall 4-6: Poster-3; 10:00-12:00; Tuesday, 4 September, 2018
Chair: Ajay Srinivasamurthy

### Comparison of Syllabification Algorithms and Training Strategies for Robust Word Count Estimation across Different Languages and

**Notes**

## Recording Conditions

*Okko Räsänen[1], Seshadri Shreyas[1] and Marisa Casillas[2]*

[1]Department of Signal Processing and Acoustics, Aalto University, Finland

[2]Max Planck Institute for Psycholinguistics, Netherlands

**Tue-P-1-3-1, Time: 10:00-12:00**

Word count estimation (WCE) from audio recordings has a number of applications, including quantifying the amount of speech that language-learning infants hear in their natural environments, as captured by daylong recordings made with devices worn by infants. To be applicable in a wide range of scenarios and also low-resource domains, WCE tools should be extremely robust against varying signal conditions and require minimal access to labeled training data in the target domain. For this purpose, earlier work has used automatic syllabification of speech, followed by a least-squares-mapping of syllables to word counts. This paper compares a number of previously proposed syllabifiers in the WCE task, including a supervised bi-directional long short-term memory (BLSTM) network that is trained on a language for which high quality syllable annotations are available (a "high resource language") and reports how the alternative methods compare on different languages and signal conditions. We also explore additive noise and varying-channel data augmentation strategies for BLSTM training and show how they improve performance in both matching and mismatching languages. Intriguingly, we also find that even though the BLSTM works on languages beyond its training data, the unsupervised algorithms can still outperform it in challenging signal conditions on novel languages.

## A Comparison of Input Types to a Deep Neural Network-based Forced Aligner

*Matthew C. Kelley and Benjamin V. Tucker*

Department of Linguistics, University of Alberta, Edmonton, AB, Canada

**Tue-P-1-3-2, Time: 10:00-12:00**

The present paper investigates the effect of different inputs on the accuracy of a forced alignment tool built using deep neural networks. Both raw audio samples and Mel-frequency cepstral coefficients were compared as network inputs. A set of experiments were performed using the TIMIT speech corpus as training data and its accompanying test data set. The networks consisted of a series of convolutional layers followed by a series of bidirectional long short-term memory (LSTM) layers. The convolutional layers were trained first to act as feature detectors, after which their weights were frozen. Then, the LSTM layers were trained to learn the temporal relations in the data. The current results indicate that networks using raw audio perform better than those using Mel-frequency cepstral coefficients and an off-the-shelf forced aligner. Possible explanations for why the raw audio networks perform better are discussed. We then lay out potential ways to improve the results of the networks and conclude with a comparison of human cognition to network architecture.

## Joint Learning Using Denoising Variational Autoencoders for Voice Activity Detection

*Youngmoon Jung, Younggwan Kim, Yeunju Choi and Hoirin Kim*

School of Electrical Engineering, KAIST, Daejeon, South Korea

**Tue-P-1-3-3, Time: 10:00-12:00**

Voice activity detection (VAD) is a challenging task in very low signal-to-noise ratio (SNR) environments. To address this issue, a promising approach is to map noisy speech features to corresponding clean features and to perform VAD using the generated clean features. This can be implemented by concatenating a speech enhancement (SE) and a VAD network, whose parameters are jointly updated. In this paper, we propose denoising variational autoencoder-based (DVAE) speech enhancement in the joint learning framework. Moreover, we feed not only the enhanced feature but also the latent code from the DVAE into the VAD network. We show that the proposed joint learning approach outperforms conventional denoising autoencoder-based joint learning approach.

## Information Bottleneck Based Percussion Instrument Diarization System for Taniavartanam Segments of Carnatic Music Concerts

*Nauman Dawalatabad, Jom Kuriakose, Chandra Sekhar Chellu and Hema Murthy*

Indian Institute of Technology Madras, India

**Tue-P-1-3-4, Time: 10:00-12:00**

An approach to diarize taniavartanam segments of a Carnatic music concert is proposed in this paper. Information bottleneck (IB) based approach used for speaker diarization is applied for this task. IB system initializes the segments to be clustered uniformly with fixed duration. The issue with diarization of percussion instruments in taniavartanam is that the stroke rate varies highly across the segments. It can double or even quadruple within a short duration, thus leading to variable information rate in different segments. To address this issue, the IB system is modified to use the stroke rate information to divide the audio into segments of varying durations. These varying duration segments are then clustered using the IB approach which is then followed by Kullback-Leibler hidden Markov model (KL-HMM) based realignment of the instrument boundaries. Performance of the conventional IB system and the proposed system is evaluated on standard Carnatic music dataset. The proposed technique shows a best case absolute improvement of 8.2% over the conventional IB based system in terms of diarization error rate.

## Robust Voice Activity Detection Using Frequency Domain Long-Term Differential Entropy

*Debayan Ghosh[1], Muralishankar R[2] and Sanjeev Gurugopinath[1]*

[1]Dept. of ECE, PES University, Bengaluru 560085, India

[2]Dept. of ECE, CMR Institute of Technology, Bengaluru 560037, India

**Tue-P-1-3-5, Time: 10:00-12:00**

We propose a novel voice activity detection (VAD) scheme employing differential entropy at each frequency bin of power spectral estimates of past and present overlapping speech frames. Here, the power spectral estimate is obtained by employing the Bartlett-Welch method. Later, we add entropies across frequency bins and denote this as the frequency domain long-term differential entropy (FLDE). Long-term averaging enhances VAD performance under low signal-to-noise-ratio (SNR). We evaluate the performance of proposed FLDE scheme, considering 12 types of noises and 5 different SNRs which are artificially added to speech samples from the SWITCHBOARD corpus. We present VAD performance of FLDE and compare with existing VAD algorithms, such as ITU-T G.729B, likelihood ratio test, long-term signal variability and long-term spectral flatness measure based algorithms. Finally, we demonstrate that our FLDE-based VAD performs with best average accuracy and speech hit-rate among the VAD algorithms considered for evaluation.

## Device-directed Utterance Detection

*Sri Harish Mallidi, Roland Maas, Kyle Goehner,*
*Ariya Rastrow, Spyros Matsoukas and Björn Hoffmeister*

Amazon, USA

**Tue-P-1-3-6, Time: 10:00-12:00**

In this work, we propose a classifier for distinguishing device-directed queries from background speech in the context of interactions with voice assistants.

Applications include rejection of false wake-ups or unintended interactions as well as enabling wake-word free follow-up queries.
Consider the example interaction: "Computer, play music", "Computer, reduce the volume."

In this interaction, the user needs to repeat the wake-word (Computer) for the second query. To allow for more natural interactions, the device could immediately re-enter listening state after the first query (without wake-word repetition) and accept or reject a potential follow-up as device-directed or background speech.

The proposed model consists of two long short-term memory (LSTM) neural networks trained on acoustic features and automatic speech recognition

**Notes**

(ASR) 1-best hypotheses, respectively.

A feed-forward deep neural network (DNN) is then trained to combine the acoustic and 1-best embeddings, derived from the LSTMs, with features from the ASR decoder.

Experimental results show that ASR decoder, acoustic embeddings and 1-best embeddings yield an equal-error-rate (EER) of 9.3%, 10.9% and 20.1%, respectively.

Combination of the features resulted in a 44% relative improvement and a final EER of 5.2%.

## Acoustic-Prosodic Features of Tabla Bol Recitation and Correspondence with the Tabla Imitation

*Rohit M A and Preeti Rao*

Department of Electrical Engineering

Indian Institute of Technology, Bombay

**Tue-P-1-3-7, Time: 10:00-12:00**

In the Indian classical drumming tradition, the different strokes on the tabla are named by spoken syllables(bols) in a case of onomatopoeia. The recitation of a tabla composition using vocalic syllables(bols) plays an important role in the oral tradition of pedagogy in North Indian classical music. Previous studies have considered the phonetic features of isolated bol utterances with the corresponding isolated strokes produced on the tabla. In this work, we investigate the acoustic properties of bol recitation beyond the segmental measurements related to the phones or syllables. The recitation of a tabla composition, apart from conveying the basic "score" in terms of the sequence of stroke name and onset times,
is typically quite expressive in nature, being marked by pitch variations, loudness dynamics and voice quality variations across a sequence or phrase.

Given the distinct spaces of acoustic variation of the voice and tabla, we study acoustic-prosodic variations in the recitation and investigate the corresponding (time-aligned) supra-segmental acoustic variations in the drumming. An available large dataset of recordings of selected tabla compositions by an expert tabla player, each aligned with the corresponding bol recitation, is employed in the analyses. We find that while the recitation reliably encodes intensity variations across bols in a cycle, the observed pitch variations are meaningful only for the pitch-varying drum strokes of the left drum.

## Who Said That? a Comparative Study of Non-negative Matrix Factorization Techniques

*Teun Krikke, Frank Broz and David Lane*

Heriot-Watt University, Edinburgh, United Kingdom

**Tue-P-1-3-8, Time: 10:00-12:00**

In noisy environments it is difficult for a computer to understand what a person is saying especially when there are multiple speakers. In this paper we concentrate on separating overlapping speech. Non-negative matrix factorisation (NMF) is a method of doing source separation without needing a lot of data. The choice of cost function can have a significant impact on the performance of NMF. We evaluate NMF using three different cost functions (Euclidean, Itakura-Saito and Kullback-Leibler) including modifications using sparsity, convolution or additional information in the form of the direction of arrival. We conduct this evaluation on three different speech corpora. Adding directional information to NMF in the form of non-negative tensor factorisation (NTF) gives us the best result on the map task and vocalization corpora and the Itakura-Saito cost function performs best on the acoustic-camera corpus. In this paper, we show that the Itakura-Saito cost function is the most robust cost function when the recording contains noise. We do this by applying acoustic evaluation measurements.

## AVA-Speech: A Densely Labeled Dataset of Speech Activity in Movies

*Sourish Chaudhuri, Joseph Roth, Daniel P. W. Ellis, Andrew Gallagher, Liat Kaver, Radhika Marvin, Caroline Pantofaru, Nathan Reale, Loretta Guarino Reid, Kevin Wilson and Zhonghua Xi*

Google Inc.

**Tue-P-1-3-9, Time: 10:00-12:00**

Speech activity detection (or endpointing) is an important processing step for applications such as speech recognition, language identification and speaker diarization. Both audio- and vision-based approaches have been used for this task in various settings, often tailored toward end applications. However, much of the prior work reports results in synthetic settings, on task-specific datasets, or on datasets that are not openly available. This makes it difficult to compare approaches and understand their strengths and weaknesses. In this paper, we describe a new dataset which we will release publicly containing densely labeled speech activity in YouTube videos, with the goal of creating a shared, available dataset for this task. The labels in the dataset annotate three different speech activity conditions: clean speech, speech co-occurring with music and speech co-occurring with noise, which enable analysis of model performance in more challenging conditions based on the presence of overlapping noise. We report benchmark performance numbers on AVA-Speech using off-the-shelf, state-of-the-art audio and vision models that serve as a baseline to facilitate future research.

## Audiovisual Speech Activity Detection with Advanced Long Short-Term Memory

*Fei Tao and Carlos Busso*

Multimodal Signal Processing(MSP) lab, Department of Electrical and Computer Engineering, The University of Texas at Dallas, Richardson TX 75080, USA

**Tue-P-1-3-10, Time: 10:00-12:00**

Speech activity detection (SAD) is a key pre-processing step for a speech-based system. The performance of conventional audio-only SAD (A-SAD) systems is impaired by acoustic noise when they are used in practical applications. An alternative approach to address this problem is to include visual information, creating audiovisual speech activity detection (AV-SAD) solutions. In our previous work, we proposed to build an AV-SAD system using bimodal recurrent neural network (BRNN). This framework was able to capture the task-related characteristics in the audio and visual inputs and model the temporal information within and across modalities. The approach relied on long short-term memory (LSTM). Although LSTM can model longer temporal dependencies with the cells, the effective memory of the units is limited to a few frames, since the recurrent connection only considers the previous frame. For SAD systems, it is important to model longer temporal dependencies to capture the semi-periodic nature of speech conveyed in acoustic and orofacial features. This study proposes to implement a BRNN-based AV-SAD system with advanced LSTMs (A-LSTMs), which overcomes this limitation by including multiple connections to frames in the past. The results show that the proposed framework can significantly outperform the BRNN system trained with the original LSTM layers.

## Towards Automatic Speech Identification from Vocal Tract Shape Dynamics in Real-time MRI

*Pramit Saha, Praneeth Srungarapu and Sidney Fels*

Department of Electrical and Computer Engineering, University of British Columbia

**Tue-P-1-3-11, Time: 10:00-12:00**

Vocal tract configurations play a vital role in generating distinguishable speech sounds, by modulating the airflow and creating different resonant cavities in speech production. They contain abundant information that can be utilized to better understand the underlying speech production mechanism. As a step towards automatic mapping of vocal tract shape geometry to acoustics, this paper employs effective video action recognition techniques, like Long-term Recurrent Convolutional Networks (LRCN) models, to identify different vowel-consonant-vowel (VCV) sequences from dynamic shaping of the vocal tract. Such a model typically combines a CNN based deep hierarchical visual feature extractor with Recurrent Networks, that ideally makes the network spatio-temporally deep enough to learn the sequential dynamics of a short video clip for video classification tasks. We use a database consisting of 2D real-time MRI of vocal tract shaping during VCV utterances by 17 speakers. The comparative performances of this class of algorithms under various parameter settings and for various

**Notes**

classification tasks are discussed. Interestingly, the results show a marked difference in the model performance in the context of speech classification with respect to generic sequence or video classification tasks.

# Tue-P-1-4: Selected Topics in Neural Speech Processing

Hall 4-6: Poster-4; 10:00-12:00; Tuesday, 4 September, 2018

Chair: Wang Ye

## Structured Word Embedding for Low Memory Neural Network Language Model

*Kaiyu Shi and Kai Yu*

Key Lab of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering

SpeechLab, Department of Computer Science and Engineering

Shanghai Jiao Tong University, Shanghai, China

**Tue-P-1-4-1, Time: 10:00-12:00**

Neural network language model (NN LM), such as long short term memory (LSTM) LM, has been increasingly popular due to its promising performance. However, the model size of an uncompressed NN LM is still too large to be used in embedded or portable devices. The dominant part of memory consumption of NN LM is the word embedding matrix. Directly compressing the word embedding matrix usually leads to performance degradation. In this paper, a product quantization based structured embedding approach is proposed to significantly reduce memory consumption of word embeddings without hurting LM performance. Here, each word embedding vector is cut into partial embedding vectors which are then quantized separately. Word embedding matrix can then be represented by an index vector and a code-book tensor of the quantized partial embedding vectors. Experiments show that the proposed approach can achieve 10 to 20 times embedding parameter reduction rate with negligible performance loss.

## Role Play Dialogue Aware Language Models Based on Conditional Hierarchical Recurrent Encoder-Decoder

*Ryo Masumura, Tomohiro Tanaka, Atsushi Ando, Hirokazu Masataki and Yushi Aono*

NTT Media Intelligence Laboratories, NTT Corporation, Japan

**Tue-P-1-4-2, Time: 10:00-12:00**

We propose role play dialogue-aware language models (RPDA-LMs) that can leverage interactive contexts in role play multi-turn dialogues for estimating the generative probability of words. Our motivation is to improve automatic speech recognition (ASR) performance in role play dialogues such as contact center dialogues and service center dialogues. Although long short-term memory recurrent neural network based language models (LSTM-RNN-LMs) can capture long-range contexts within an utterance, they cannot utilize sequential interactive information between speakers in multi-turn dialogues. Our idea is to explicitly leverage speakers' roles of individual utterances, which are often available in role play dialogues, for neural language modeling. The RPDA-LMs are represented as a generative model conditioned by a role sequence of a target role play dialogue. We compose the RPDA-LMs by extending hierarchical recurrent encoder-decoder modeling so as to handle the role information. Our ASR evaluation in a contact center dialogue demonstrates that RPDA-LMs outperform LSTM-RNN-LMs and document-context LMs in terms of perplexity and word error rate. In addition, we verify the effectiveness of explicitly taking interactive contexts into consideration.

## Efficient Keyword Spotting Using Time Delay Neural Networks

*Samuel Myer and Vikrant Singh Tomar*

Fluent.ai Inc., Montreal, Canada

**Tue-P-1-4-3, Time: 10:00-12:00**

This paper describes a novel method of live keyword spotting using a two-stage time delay neural network. The model is trained using transfer learning: initial training with phone targets from a large speech corpus is followed by training with keyword targets from a smaller data set. The accuracy of the system is evaluated on two separate tasks. The first is the freely available Google Speech Commands dataset. The second is an in-house task specifically developed for keyword spotting. The results show significant improvements in false accept and false reject rates in both clean and noisy environments when compared with previously known techniques. Furthermore, we investigate various techniques to reduce computation in terms of multiplications per second of audio. Compared to recently published work, the proposed system provides up to 89% savings on computational complexity.

## Automatic DNN Node Pruning Using Mixture Distribution-based Group Regularization

*Tsukasa Yoshida[1,2], Takafumi Moriya[1], Kazuho Watanabe[2], Yusuke Shinohara[1], Yoshikazu Yamaguchi[1] and Yushi Aono[1]*

[1]NTT Media Intelligence Laboratories, NTT Corporation, Japan

[2]Department of Computer Science and Engineering, Toyohashi University of Technology, Japan

**Tue-P-1-4-4, Time: 10:00-12:00**

In this paper, we address a constrained training for deep neural network-based acoustic model size reduction. While the L2 regularizer is used as a modeling approach to shrinking parameters, we cannot cut down the unimportant parts because it does not assume any group structure.

The Group Lasso regularizer is used for the model size reduction approach. Group Lasso can set arbitrary group parameters (e.g. the column vector norms of the parameter matrices) as unimportant parts and make the parameters sparse. Therefore, we can prune the unimportant parameters whose group parameter norm is nearly zero. However, Group Lasso does not suggest a clear rule for separating parameters close to zero and large in the group parameter space and hence is unsuitable for the model size reduction.

To solve these problems, we propose a mixture distribution-based regularizer which assumes distributions of norms in the group parameter space. We evaluate our method on a NTT real recorded voice search data containing 1600 hours. Our proposal achieves 27.0% reduction compared to the pruned model by Group Lasso while keeping recognition performance.

## Conditional-Computation-Based Recurrent Neural Networks for Computationally Efficient Acoustic Modelling

*Raffaele Tavarone and Leonardo Badino*

Center for Translational Neurophysiology of Speech and Communication

**Tue-P-1-4-5, Time: 10:00-12:00**

The first step in Automatic Speech Recognition (ASR) is a fixed-rate segmentation of the acoustic signal into overlapping windows of fixed length. Although this procedure allows to achieve excellent recognition accuracy, it is far from being computationally efficient,
in that it may produce a highly redundant signal (i.e, almost identical spectral vectors may span many observation windows) that converts into computational overload. The reduction of such overload can be very beneficial for application such as offline ASR on mobile devices.

In this paper we present a principled way for saving numerical operations during ASR by using conditional-computation methods in deep bidirectional Recurrent Neural Networks (RNNs) for acoustic modelling.

The methods rely on learned binary neurons that allow hidden layers to be updated only when necessary or to keep their previous value.

We (i) evaluate, for the first time, conditional computation-based recurrent architectures on a speech recognition task and (ii) propose a novel model specifically designed for speech data that inherently builds a multi-scale temporal structure in the hidden layers.

**Notes**

Results on the TIMIT dataset show that conditional mechanisms in recurrent architectures can reduce hidden layer updates up to 40% at the cost of about 20% relative phone error rate increase.

## Leveraging Translations for Speech Transcription in Low-resource Settings

*Antonios Anastasopoulos and David Chiang*

Department of Computer Science and Engineering

University of Notre Dame, IN, USA

**Tue-P-1-4-6, Time: 10:00-12:00**

Recently proposed data collection frameworks for endangered language documentation aim not only to collect speech in the language of interest, but also to collect translations into a high-resource language that will render the collected resource interpretable. We focus on this scenario and explore whether we can improve transcription quality under these extremely low-resource settings with the assistance of text translations. We present a neural multi-source model and evaluate several variations of it on three low-resource datasets. We find that our multi-source model with shared attention outperforms the baselines, reducing transcription character error rate by up to 12.3%.

## Sequence-to-sequence Neural Network Model with 2D Attention for Learning Japanese Pitch Accents

*Antoine Bruguier[1], Heiga Zen[2] and Arkady Arkhangorodsky[3]*

[1]Google, Mountain View, USA

[2]Google, London, UK

[3]University of Toronto, Toronto, Canada

**Tue-P-1-4-7, Time: 10:00-12:00**

Many Japanese text-to-speech (TTS) systems use word-level pitch accents as one of the prosodic features. Combination of a pronunciation dictionary including lexical pitch accents and a statistical model representing the word accent sandhi is often used to predict pitch accents from a text. However, using human transcribers to build the dictionary and training data for the model is tedious and expensive. This paper proposes a neural pitch accent recognition model. This model combines the information from audio and its transcription (word sequence in hiragana characters) via two-dimensional attention and outputs word-level pitch accents. Experimental results show a reduction in the word pitch accent prediction error rate over that with text only. It lowers the load of human annotators when building a pronunciation dictionary. As the approach is general, it can be used to do pronunciation learning in other languages as well.

## Task Specific Sentence Embeddings for ASR Error Detection

*Sahar Ghannay, Yannick Estève and Nathalie Camelin*

LIUM - University of Le Mans, France

**Tue-P-1-4-8, Time: 10:00-12:00**

This paper presents a study on the modeling of automatic speech recognition errors at the sentence level. We aim in this study to compensate certain phenomena highlighted by the analysis of the outputs generated by the ASR error detection system we previously proposed. We investigated three different approaches, that are based respectively on the use of sentence embeddings dedicated to ASR error detection task, a probabilistic contextual model and a bidirectional long short term memory (BLSTM) architecture. An approach to build task-specific sentence embeddings is proposed and compared to the Doc2vec approach. Experiments are performed on transcriptions generated by the LIUM ASR system applied to the ETAPE corpus. They show that the proposed sentence embeddings dedicated to ASR error detection achieve better results than generic sentence embeddings and that the integration of task-specific sentence embeddings in our system achieves better results than the probabilistic contextual model and BLSTM models.

## Low-Latency Neural Speech Translation

*Jan Niehues, Ngoc-Quan Pham, Thanh-Le Ha, Matthias Sperber and Alex Waibel*

Institute for Anthropomatics and Robotics, KIT - Karlsruhe Institute of Technology, Germany

**Tue-P-1-4-9, Time: 10:00-12:00**

Through the development of neural machine translation, the quality of machine translation systems has been improved significantly. By exploiting advancements in deep learning, systems are now able to better approximate the complex mapping from source sentences to target sentences. But with this ability, new challenges also arise. An example is the translation of partial sentences in low-latency speech translation. Since the model has only seen complete sentences in training, it will always try to generate a complete sentence, though the input may only be a partial sentence. We show that NMT systems can be adapted to scenarios where no task-specific training data is available. Furthermore, this is possible without losing performance on the original training data. We achieve this by creating artificial data and by using multi-task learning. After adaptation, we are able to reduce the number of corrections displayed during incremental output construction by 45%, without a decrease in translation quality.

## Low-Resource Speech-to-Text Translation

*Sameer Bansal[1], Herman Kamper[2], Karen Livescu[3], Adam Lopez[1] and Sharon Goldwater[1]*

[1]School of Informatics, University of Edinburgh, UK

[2]Stellenbosch University, South Africa

[3]Toyota Technological Institute at Chicago, USA

**Tue-P-1-4-10, Time: 10:00-12:00**

Speech-to-text translation has many potential applications for low-resource languages, but the typical approach of cascading speech recognition with machine translation is often impossible, since the transcripts needed to train a speech recognizer are usually not available for low-resource languages. Recent work has found that neural encoder-decoder models can learn to directly translate foreign speech in high-resource scenarios, without the need for intermediate transcription. We investigate whether this approach also works in settings where both data and computation are limited. To make the approach efficient, we make several architectural changes, including a change from character-level to word-level decoding. We find that this choice yields crucial speed improvements that allow us to train with fewer computational resources, yet still performs well on frequent words. We explore models trained on between 20 and 160 hours of data and find that although models trained on less data have considerably lower BLEU scores, they can still predict words with relatively high precision and recall-around 50% for a model trained on 50 hours of data, versus around 60% for the full 160 hour model. Thus, they may still be useful for some low-resource scenarios.

## VoiceGuard: Secure and Private Speech Processing

*Ferdinand Brasser[1], Tommaso Frassetto[1], Korbinian Riedhammer[2], Ahmad-Reza Sadeghi[1], Thomas Schneider[1] and Christian Weinert[1]*

[1]Technische Universität Darmstadt, Germany

[2]University of Applied Sciences Rosenheim, Germany

**Tue-P-1-4-11, Time: 10:00-12:00**

With the advent of smart-home devices providing voice-based interfaces, such as Amazon Alexa or Apple Siri, voice data is constantly transferred to cloud services for automated speech recognition or speaker verification.

While this development enables intriguing new applications, it also poses significant risks: Voice data is highly sensitive since it contains biometric information of the speaker as well as the spoken words. This data may be abused if not protected properly, thus the security and privacy of billions of end-users is at stake.

We tackle this challenge by proposing an architecture, dubbed VoiceGuard, that efficiently protects the speech processing task inside a trusted execution environment (TEE). Our solution preserves the privacy of users while at the same time it does not require the service provider to reveal model parameters. Our architecture can be extended to enable user-specific models, such as feature transformations (including fMLLR), i-vectors, or model transformations (e.g., custom output layers). It also generalizes to secure on-premise solutions, allowing vendors to securely ship their models to customers.

**Notes**

We provide a proof-of-concept implementation and evaluate it on the Resource Management and WSJ speech recognition tasks isolated with Intel SGX, a widely available TEE implementation, demonstrating even real time processing capabilities.

## Perspective Talk-1

Hall 3; 12:00-12:30; Tuesday, 4 September, 2018
Chair: Ajay Srinivasamurthy

### Deep Learning based Situated Goal-oriented Dialogue Systems

*Dilek Hakkani-Tür*

Research Scientist, Amazon, USA

**Tue-Perspective-1, Time: 12:00-12:30**

Interacting with machines in natural language has been a holy grail since the beginning of computers. Given the difficulty of understanding natural language, only in the past couple of decades, we started seeing real user applications for targeted/limited domains. More recently, advances in deep learning-based approaches enabled exciting new research frontiers for end-to-end goal-oriented conversational systems. In this talk, I'll review end-to-end dialogue systems research, with components for situated language understanding, dialogue state tracking, policy, and language generation. The talk will highlight novel approaches where dialogue is viewed as a collaborative game between a user and an agent in the presence of visual information, and will aim to summarize challenges for future research.

## Industry Presentation-1

Hall 3; 12:30-13:00; Tuesday, 4 September, 2018
Chairs: Ananth Sankar and Hemant Patil

**Presenters:** Björn Hoffmeister and Sri Garimella
**Company:** Amazon

## Industry Presentation-2

Hall 1; 12:30-13:00; Tuesday, 4 September, 2018
Chairs: Dhananjaya N Gowda and Dileep A D

**Presenter:** Bowen Zhou
**Company:** JD

## Industry Presentation-3

Hall 2; 12:30-13:00; Tuesday, 4 September, 2018
Chairs: Gauri Deshpande and Rohit Sinha

**Presenter:** Samith Ramachandran
**Company:** Uniphore

## Tue-O-2-1: Dereverberation

Hall 3; 14:30-16:30; Tuesday, 4 September, 2018
Chairs: Tom Bäckström and Rajbabu Velmurugan

### Single-channel Speech Dereverberation via Generative Adversarial Training

*Chenxing Li[1,2], Tieqiang Wang[1,2], Shuang Xu[1] and Bo Xu[1]*

[1]Institute of Automation, Chinese Academy of Sciences, Beijing, P.R.China

[2]University of Chinese Academy of Sciences, Beijing, P.R.China

**Tue-O-2-1-1, Time: 14:30-14:50**

In this paper, we propose a single-channel speech dereverberation system (DeReGAT) based on convolutional, bidirectional long short-term memory and deep feed-forward neural network (CBLDNN) with generative adversarial training (GAT). In order to obtain better speech quality instead of only minimizing a mean square error (MSE), GAT is employed to make the dereverberated speech indistinguishable form the clean samples. Besides, our system can deal with wide range reverberation and be well adapted to variant environments. The experimental results show that the proposed model outperforms weighted prediction error (WPE) and deep neural network-based systems. In addition, DeReGAT is extended to an online speech dereverberation scenario, which reports comparable performance with the offline case.

### Single-channel Dereverberation Using Direct MMSE Optimization and Bidirectional LSTM Networks

*Wolfgang Mack, Soumitro Chakrabarty, Fabian-Robert Stöter, Sebastian Braun, Bernd Edler and Emanuël Habets*

International Audio Laboratories Erlangen, 91058 Erlangen, Germany

**Tue-O-2-1-2, Time: 14:50-15:10**

Dereverberation is useful in hands-free communication and voice controlled devices for distant speech acquisition. Single-channel dereverberation can be achieved by applying a time-frequency (TF) mask to the short-time Fourier transform (STFT) representation of a reverberant signal. Recent approaches have used deep neural networks (DNNs) to estimate such masks. Previously proposed DNN-based mask estimation methods train a DNN to minimize the mean-squared-error (MSE) between the desired and estimated masks. Recent TF mask estimation methods for signal separation directly minimize instead the MSE between the desired and estimated STFT magnitudes. We apply this direct optimization concept to dereverberation. Moreover, as reverberation exceeds the duration of a single STFT frame, we propose to use a bidirectional long short-term memory (LSTM) network which is able to take the relation between multiple STFT frames into account. We evaluated our method for different reverberation times and source-microphone distances using simulated as well as measured room impulse responses of different rooms. An evaluation of the proposed method and a comparison with a state-of-the-art method demonstrate the superiority of our approach and its robustness to different acoustic conditions.

**Notes**

### Single-channel Late Reverberation Power Spectral Density Estimation Using Denoising Autoencoders

*Ina Kodrasi and Hervé Bourlard*

Idiap Research Institute, Speech and Audio Processing Group, Martigny, Switzerland

**Tue-O-2-1-3, Time: 15:10-15:30**

In order to suppress the late reverberation in the spectral domain, many single-channel dereverberation techniques rely on an estimate of the late reverberation power spectral density (PSD). In this paper, we propose a novel approach to late reverberation PSD estimation using a denoising autoencoder (DA), which is trained to learn a mapping from the microphone signal PSD to the late reverberation PSD. Simulation results show that the proposed approach yields a high PSD estimation accuracy and generalizes well to unseen data. Furthermore, simulation results show that the proposed DA-based PSD estimate yields a higher PSD estimation accuracy and a similar dereverberation performance than a state-of-the-art statistical PSD estimate, which additionally also requires knowledge of the reverberation time.

### A Non-convolutive NMF Model for Speech Dereverberation

*Nikhil M, Rajbabu Velmurugan and Preeti Rao*

Indian Institute of Technology Bombay

**Tue-O-2-1-4, Time: 15:30-15:50**

Reverberation corrupts speech recorded using distant microphones, resulting in poor speech intelligibility. We propose a single-channel, supervised non-negative matrix factorization (NMF) based dereverberation method, in contrast to the convolutive NMF (CNMF) based methods in literature. Recent supervised approaches use a CNMF model for reverberation and a NMF model for clean speech spectrogram to obtain enhanced speech by directly estimating the clean speech activations. In the proposed method, with a separability assumption on the room impulse response (RIR) spectrogram, the reverb speech can be decomposed into bases and activations using conventional NMF. Using these reverb activations, the clean speech activations are estimated to obtain enhanced speech. The proposed model (i) helps in imposing meaningful constraints on the RIR in both frequency- and time-domains to achieve improved enhancement (ii) leads to a framework that can include a NMF model for noise. (iii) gives a better interpretation of the effects of reverberation in the NMF context. We evaluate and compare the enhancement performance of the algorithm on reverb and noisy conditions, simulated using TIMIT utterances and REVERB challenge RIRs. The proposed method performs better than existing C-NMF based methods in objective measures, such as cepstral distance (CD) and speech-to-reverberation modulation energy ratio (SRMR).

### Cross-Corpora Convolutional Deep Neural Network Dereverberation Preprocessing for Speaker Verification and Speech Enhancement

*Peter Guzewich, Stephen Zahorian, Xiao Chen and Hao Zhang*

Department of Electrical and Computer Engineering, Binghamton University, NY, USA

**Tue-O-2-1-5, Time: 15:50-16:10**

Deep neural network (DNN) dereverberation preprocessing has been shown to be a viable strategy for speech enhancement and increasing the accuracy of automatic speech recognition and automatic speaker verification. In this paper, an improved DNN technique based on convolutional neural networks is presented and compared to existing methods for speech enhancement and speaker verification in the presence of reverberation. This new technique is first shown to enhance speech quality as compared to other existing methods. Then, a more thorough set of experiments is presented that assesses cross-corpora speaker verification performance on data that contains real reverberation and noise. A discussion of the applicability and generalizability of such techniques is given.

### Dereverberation and Beamforming in Robust Far-Field Speaker Recognition

*Ladislav Mošner, Oldřich Plchot, Pavel Matějka, Ondřej Novotný and Jan Černocký*

Brno University of Technology, Speech@FIT and IT4I Center of Excellence, Czechia

**Tue-O-2-1-6, Time: 16:10-16:30**

This paper deals with robust speaker verification (SV) in far-field sensing. The robustness is verified on a subset of NIST SRE 2010 corpus retransmitted in multiple real rooms of different acoustics and captured with multiple microphones. We experimented with various data preprocessing steps including different approaches to dereverberation and beamforming applied to ad-hoc microphone arrays. We found that significant improvements in accuracy can be achieved with neural network based generalized eigenvalue beamformer preceded by weighted prediction error dereverberation. We also explored the effect of data augmentation by adding various real or simulated room acoustic properties to the Probabilistic Linear Discriminant Analysis (PLDA) training dataset. As a result, we developed a speaker recognition system whose performance is stable across different room acoustic conditions. It yields 41.4% relative improvement in performance over the system without multi-channel processing tested on the cleanest microphone data. With the best combination of data preprocessing and augmentation, we obtained a performance close to the one we achieved with the original clean test data.

## Tue-O-2-2: Audio Events and Acoustic Scenes

Hall 1; 14:30-16:30; Tuesday, 4 September, 2018

Chairs: Weiran Wang and Mark Hasegawa-Johnson

### Comparing the Max and Noisy-Or Pooling Functions in Multiple Instance Learning for Weakly Supervised Sequence Learning Tasks

*Yun Wang[1], Juncheng Li[1,2] and Florian Metze[1]*

[1]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, U.S.A.

[2]Research and Technology Center, Robert Bosch LLC, Pittsburgh, PA, U.S.A

**Tue-O-2-2-1, Time: 14:30-14:50**

Many sequence learning tasks require the localization of certain events in sequences. Because it can be expensive to obtain strong labeling that specifies the starting and ending times of the events, modern systems are often trained with weak labeling without explicit timing information. Multiple instance learning (MIL) is a popular framework for learning from weak labeling. In a common scenario of MIL, it is necessary to choose a pooling function to aggregate the predictions for the individual steps of the sequences. In this paper, we compare the "max" and "noisy-or" pooling functions on a speech recognition task and a sound event detection task. We find that max pooling is able to localize phonemes and sound events, while noisy-or pooling fails. We provide a theoretical explanation of the different behavior of the two pooling functions on sequence learning tasks.

### A Simple Model for Detection of Rare Sound Events

*Weiran Wang, Chieh-Chi Kao and Chao Wang*

Amazon Alexa, 101 Main St, Cambridge, MA 02142, USA

**Tue-O-2-2-2, Time: 14:50-15:10**

We propose a simple recurrent model for detecting rare sound events, when the time boundaries of events are available for training. Our model optimizes the combination of an utterance-level loss, which classifies whether an event occurs in an utterance and a frame-level loss, which

**Notes**

classifies whether each frame corresponds to the event when it does occur. The two losses make use of a shared vectorial representation the event and are connected by an attention mechanism. We demonstrate our model on Task 2 of the DCASE 2017 challenge and achieve competitive performance.

### Temporal Transformer Networks for Acoustic Scene Classification

*Teng Zhang, Kailai Zhang and Ji Wu*

Multimedia Signal and Intelligent Information Processing Lab

Department of Electronic Engineering, Tsinghua University, Beijing, P.R.China

**Tue-O-2-2-3, Time: 15:10-15:30**

Neural networks have been proven to be powerful models for acoustic scene classification tasks, but are still limited by the lack of ability to be temporally invariant to the audio data. In this paper, a novel temporal transformer module is proposed to allow the temporal manipulation of data in neural networks. This module is composed of a Fourier transform layer for feature maps and a learnable feature reduction layer and can be inserted into existing convolutional neural network (CNN) and Long short-term memory (LSTM) models. Experiments on LITIS Rouen dataset and DCASE2016 dataset show that the proposed method leads to a significant improvement when compared with the existing neural networks. Our approach is able to perform significantly better than the state-of-the-art result on LITIS Rouen dataset, obtaining a relative reduction of 23.6% on classification error.

### Temporal Attentive Pooling for Acoustic Event Detection

*Xugang Lu[1], Peng Shen[1], Sheng Li[1], Yu Tsao[2] and Hisashi Kawai[1]*

[1]National Institute of Information and Communications Technology, Japan

[2]Research Center for Information Technology Innovation, Academic Sinica, Taiwan

**Tue-O-2-2-4, Time: 15:30-15:50**

Deep convolutional neural network (DCNN) based model has been successfully applied to acoustic event detection (AED) due to its efficiency to explore temporal-frequency structure for feature representations. In most studies, the final representation either uses a temporal average- or max-pooling algorithm to accumulate local temporal features as a global representation for event classification. The temporal pooling algorithm in the DCNN is based on the assumption that the target label is assigned to all temporal locations (average pooling) or to only one temporal location with a maximum response (max-pooling). However, the acoustic event labels are holistic descriptions in a semantic level, it is difficult or even impossible to decide features from which temporal locations contribute to the event perception. In this study, we propose a weighted temporal-pooling algorithm to accumulate local temporal features for AED. The pooling algorithm integrates global and local attention modules in a convolutional recurrent neural network to integrate temporal features. Experiments on an AED task were carried out to evaluate the proposed model. Results showed that with the global and local attentions, a large gain was obtained.

### R-CRNN: Region-based Convolutional Recurrent Neural Network for Audio Event Detection

*Chieh-Chi Kao, Weiran Wang, Ming Sun and Chao Wang*

Amazon Alexa

**Tue-O-2-2-5, Time: 15:50-16:10**

This paper proposes a Region-based Convolutional Recurrent Neural Network (R-CRNN) for audio event detection (AED). The proposed network is inspired by Faster-RCNN [1], a well-known region-based convolutional network framework for visual object detection. Different from the original Faster-RCNN, a recurrent layer is added on top of the convolutional network to capture the long-term temporal context from the extracted high-level features. While most of the previous works on AED generate predictions at frame level first and then use post-processing to predict the onset/offset timestamps of events from a probability sequence; the proposed method generates predictions at event level directly and can be trained end-to-end with a multi-task loss, which optimizes the classification and localization of audio events simultaneously. The proposed method is tested on DCASE 2017 Challenge dataset [2]. To the best of our knowledge, R-CRNN is the best performing single-model method among all methods without using ensembles both on development and evaluation sets. Compared to the other region-based network for AED (R-FCN [3]) with an event-based error rate (ER) of 0.18 on the development set, our method reduced the ER to half.

### Detecting Media Sound Presence in Acoustic Scenes

*Constantinos Papayiannis[1,2], Justice Amoh[1,2], Viktor Rozgic[1], Shiva Sundaram[1] and Chao Wang[1]*

[1]Alexa Machine Learning, Amazon.com, Cambridge, MA, USA

[2]Department of Electrical and Electronic Engineering, Imperial College London, UK

[3]Thayer School of Engineering, Dartmouth College, Hanover, NH, USA

**Tue-O-2-2-6, Time: 16:10-16:30**

Using speech to interact with electronic devices and access services is becoming increasingly common. Using such applications in our households poses new challenges for speech and audio processing algorithms as these applications should perform robustly in a number of scenarios. Media devices are very commonly present in such scenarios and can interfere with the user-device communication by contributing to the noise or simply by being mistaken as user issued voice commands. Detecting the presence of media sounds in the environment can help avoid such issues. In this work we propose a method for this task based on a parallel CNN-GRU-FC classifier architecture which relies on multi-channel information to discriminate between media and live sources. Experiments performed using 378 hours of in-house audio recordings collected by volunteers show an F1 score of 71% with a recall of 72% in detecting active media sources. The use of information from multiple channels gave a relative improvement of 16% to the F1 score when compared to using information from only a single channel.

## Tue-O-2-3: Speaker Diarization

Hall 2; 14:30-16:30; Tuesday, 4 September, 2018

Chairs: Sriram Ganapathy and Xavier Anguera

### S4D: Speaker Diarization Toolkit in Python

*Pierre-Alexandre Broux[1,2], Florent Desnous[2], Anthony Larcher[2], Simon Petitrenaud[2], Jean Carrive[1] and Sylvain Meignier[2]*

[1]French National Audiovisual Institute (INA), Paris, France

[2]Computer Science Laboratory of Le Mans University (LIUM - EA 4023), Le Mans, France

**Tue-O-2-3-1, Time: 14:30-14:50**

In this paper, we present S4D, a new open-source Python toolkit dedicated to speaker diarization. S4D provides various state-of-the-art components and the possibility to easily develop end-to-end diarization prototype systems. S4D offers a large panel of clustering, segmentation, scoring and visualization algorithms. S4D has been thought to be easily understood, installed, modified and used in order to allow fast transfers of diarization technologies to industry and facilitate development of new approaches.

Examples, benchmarks on standard tasks and tutorials are provided in this paper.

S4D is an extension of the open-source toolkit for speaker recognition: SIDEKIT.

**Notes**

### Multimodal Speaker Segmentation and Diarization Using Lexical and Acoustic Cues via Sequence-to-sequence Neural Networks

*Tae Jin Park and Panayiotis Georgiou*

University of Southern California, Los Angeles, CA, USA

**Tue-O-2-3-2, Time: 14:50-15:10**

While there has been substantial amount of work in speaker diarization recently, there are few efforts in jointly employing lexical and acoustic information for speaker segmentation. Towards that, we investigate a speaker diarization system using a sequence-to-sequence neural network trained on both lexical and acoustic features. We also propose a loss function that allows for selecting not only the speaker change points but also the best speaker at any time by allowing for different speaker groupings. We incorporate Mel Frequency Cepstral Coefficients (MFCC) as an acoustic feature stream alongside lexical information that are obtained from conversations from the Fisher dataset. Thus, we show that acoustics provide complementary information to the lexical modality. The experimental results show that sequence-to-sequence system trained on both word sequences and MFCC can improve on speaker diarization result compared to the system that only relies on lexical modality or the baseline MFCC-based system. In addition, we test the performance of our proposed method with Automatic Speech Recognition (ASR) transcripts. While the performance on ASR transcripts drops, the Diarization Error Rate (DER) of our proposed method still outperforms the traditional method based on Bayesian Information Criterion (BIC).

### Combined Speaker Clustering and Role Recognition in Conversational Speech

*Nikolaos Flemotomos, Pavlos Papadopoulos, James Gibson and Shrikanth Narayanan*

Department of Electrical Engineering, University of Southern California, Los Angeles, CA, USA

**Tue-O-2-3-3, Time: 15:10-15:30**

Speaker Role Recognition (SRR) is usually addressed either as an independent classification task, or as a subsequent step after a speaker clustering module. However, the first approach does not take speaker-specific variabilities into account, while the second one results in error propagation. In this work we propose the integration of an audio-based speaker clustering algorithm with a language-aided role recognizer into a meta-classifier which takes both modalities into account. That way, we can treat separately any speaker-specific and role-specific characteristics before combining the relevant information together. The method is evaluated on two corpora of different conditions with interactions between a clinician and a patient and it is shown that it yields superior results for the SRR task.

### The ACLEW DiViMe: An Easy-to-use Diarization Tool

*Adrien Le Franc[1,2], Eric Riebling[3], Julien Karadayi[1,2], Yun Wang[3], Camila Scaff[1], Florian Metze[3] and Alejandrina Cristia[1]*

[1]LSCP, Département d'études cognitives, ENS, EHESS, CNRS, PSL University, Paris, France

[2]Institut national de recherche en informatique et en automatique, Paris, France

[3]Language Technologies Institute; Carnegie Mellon University; Pittsburgh, PA, U.S.A.

**Tue-O-2-3-4, Time: 15:30-15:50**

We present "DiViMe", an open-source virtual machine aimed at packaging speech technology for real-life data and developed in the context of the "Analyzing Children's Language Environments across the World" Project. This first release focuses on Speech Activity Detection, Speaker Diarization and their evaluation. The present paper introduces the set of included tools and the current workflow, which is focused on making minimal assumptions regarding users' technical skills. Additionally, we show how the current DiViMe tools fare against three sets of challenging data. In a first experiment, we look at performance with samples extracted from daylong recordings gathered using the LENA{TM}, system from English-learning children. We find that the performance of the tools currently in DiViMe is not far from that achieved by the lena proprietary software. In a second experiment, we generalize to other samples of child-centered daylong files, gathered with non-LENA{TM}, hardware from non-English-learning children, showing that performance does not degrade in this condition. Finally, we report on performance in the DiHARD 2018 Challenge Test Data. Originally conceived in the "Speech Recognition Virtual Kitchen", DiViMe is a promising platform for packaging speech technology tools for widespread re-use, with potential impact on both fundamental and applied speech and language research.

### Automatic Detection of Multi-speaker Fragments with High Time Resolution

*Evdokia Kazimirova[1] and Andrey Belyaev[1,2]*

[1]Neurodatalab, USA

[2]Lomonosov MSU, Russia

**Tue-O-2-3-5, Time: 15:50-16:10**

Interruptions and simultaneous talking represent important patterns of speech behavior. However, there is a lack of approaches to their automatic detection in continuous audio data. We have developed a solution for automatic labeling of multi-speaker fragments using harmonic traces analysis. Since harmonic traces in multi-speaker intervals form an irregular pattern as opposed to the structured pattern typical for a single speaker, we resorted to computer vision methods to detect multi-speaker fragments.

A convolutional neural network was trained on synthetic material to differentiate between single-speaker and multi-speaker fragments. For evaluation of the proposed method the SSPNet Conflict Corpus with provided manual diarization was used. We also examined factors affecting algorithm performance.

The main advantages of the proposed method are calculation simplicity and high time resolution. With our approach it is possible to detect segments with minimum duration of 0.5 seconds. The proposed method demonstrates highly accurate results and may be used for speech segmentation, speaker tracking, content analysis such as conflict detection and other practical purposes.

### Neural Speech Turn Segmentation and Affinity Propagation for Speaker Diarization

*Ruiqing Yin, Hervé Bredin and Claude Barras*

LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, F-91405 Orsay, France

**Tue-O-2-3-6, Time: 16:10-16:30**

Speaker diarization is the task of determining who speaks when in an audio stream. Most diarization systems rely on statistical models to address four sub-tasks: speech activity detection (SAD), speaker change detection (SCD), speech turn clustering and re-segmentation. First, following the recent success of recurrent neural networks (RNN) for SAD and SCD, we propose to address re-segmentation with Long-Short Term Memory (LSTM) networks. Then, we propose to use affinity propagation on top of neural speaker embeddings for speech turn clustering, outperforming regular Hierarchical Agglomerative Clustering (HAC). Finally, all these modules are combined and jointly optimized to form a speaker diarization pipeline in which all but the clustering step are based on RNNs. We provide experimental results on the French Broadcast dataset ETAPE where we reach state-of-the-art performance.

## Tue-O-2-4: Phonation

MR G.01-G.02; 14:30-16:30; Tuesday, 4 September, 2018

Chairs: Christer Gobl and Volker Dellwo

### Pitch or Phonation: on the Glottalization in Tone

**Notes**

## Productions in the Ruokeng Hui Chinese Dialect

*Minghui Zhang[1] and Fang Hu[2]*

[1]Si-Mian Institute for Advanced Studies in Humanities, East China Normal University

[2]Institute of Linguistics, Chinese Academy of Social Sciences

**Tue-O-2-4-1, Time: 14:30-14:50**

This paper examines the interplay of glottalization and tones in tonal phonology of the Ruokeng Hui Chinese. Acoustic data from 10 native speakers were analyzed in terms of pitch (F0), duration, H1-H2, H1-A1/2/3, CPP, HNR, SHR, etc. Fine-grained phonetic details reveal the interactions between phonation and tones and shed light on the ongoing tonal change from a glottalized tone to a plain high falling tone in the Ruokeng dialect.

## Speaker-specific Structure in German Voiceless Stop Voice Onset Times

*Marc Antony Hullebus[1], Stephen Tobin[1,2] and Adamantios Gafos[1,2]*

[1]Department für Linguistik, Universität Potsdam, Germany

[2]Haskins Laboratories, New Haven, USA

**Tue-O-2-4-2, Time: 14:50-15:10**

Voice onset time (VOT), a primary cue for voicing in many languages including English and German, is known to vary greatly between speakers, but also displays robust within-speaker consistencies, at least in English. The current analysis extends these findings to German. VOT measures were investigated from voiceless alveolar and velar stops in CV syllables cued by a visual prompt in a cue-distractor task. Comparably to English, a considerable portion of German VOT variability can be attributed to the syllable's vowel length and the stop's place of articulation. Individual differences in VOT still remain irrespective of speech rate. However, significant correlations across places of articulation and between speaker-specific mean VOTs and standard deviations indicate that talkers employ a relatively unified VOT profile across places of articulation. This could allow listeners to more efficiently adapt to speaker-specific realisations.

## Creak in the Respiratory Cycle

*Kätlin Aare[1,2], Pärtel Lippus[1], Marcin Wlodarczak[2] and Mattias Heldner[2]*

[1]University of Tartu, Estonia

[2]Stockholm University, Sweden

**Tue-O-2-4-3, Time: 15:10-15:30**

Creakiness is a well-known turn-taking cue and has been observed to systematically accompany phrase and turn ends in several languages. In Estonian, creaky voice is frequently used by all speakers without any obvious evidence for its systematic use as a turn-taking cue. Rather, it signals a lack of prominence and is favored by lengthening and later timing in phrases. In this paper, we analyze the occurrence of creak with respect to properties of the respiratory cycle. We show that creak is more likely to accompany longer exhalations. Furthermore, the results suggest there is little difference in lung volume values regardless of the presence of creak, indicating that creaky voice might be employed to preserve air over the course of longer utterances. We discuss the results in connection to processes of speech planning in spontaneous speech.

## Acoustic Analysis of Whispery Voice Disguise in Mandarin Chinese

*Cuiling Zhang[1,2], Bin Li[3], Si Chen[4] and Yike Yang[4]*

[1]Southwest University of Political Science & Law, Chongqing, China

[2]Chongqing Institutes of Higher Education Key Forensic Science Laboratory, Chongqing, China

[3]City University of Hong Kong, Hong Kong S.A.R., China

[4]The Hong Kong Polytechnic University, Hong Kong S.A.R., China

**Tue-O-2-4-4, Time: 15:30-15:50**

This paper investigates the auditory and acoustical characteristics of whispery disguised voice and compares the patterns with those of normal (non-disguised) voices. It also evaluates effects of whispery disguise on forensic voice comparison. Recordings of eleven male college students' normal voices and whispery disguised voices were collected. All their normal and whisper speech was acoustically analyzed and compared. The parameters including average syllable duration, intensity, vowel formant frequencies and long term average spectrum (LTAS) were measured and statistically analyzed. The effect of whispery voice disguise on speaker recognition by auditory perception and an automatic system were evaluated. Correlation and regression analyses were made on the parameters of whispery voice and normal voice. These simple regression models can be used for parameter compensation in forensic casework.

## The Zurich Corpus of Vowel and Voice Quality, Version 1.0

*Dieter Maurer[1], Christian d'Heureuse[1], Heidy Suter[1], Volker Dellwo[2], Daniel Friedrichs[3] and Thayabaran Kathiresan[2]*

[1]Institute for the Performing Arts and Film, Zurich University of the Arts, Switzerland

[2]Department of Computational Linguistics, University of Zurich, Switzerland

[3]Department of Speech, Hearing and Phonetic Sciences, University College London, United Kingdom

**Tue-O-2-4-5, Time: 15:50-16:10**

Existing databases of isolated vowel sounds or vowel sounds embedded in consonantal context generally document only limited variation of basic production parameters. Thus, concerning the possible variation range of vowel and voice quality-related sound characteristics, there is a lack of broad phenomenological and descriptive references that allow for a comprehensive understanding of vowel acoustics and for an evaluation of the extent to which corresponding existing approaches and models can be generalised. In order to contribute to the building up of such references, a novel database of vowel sounds that exceeds any existing collection by size and diversity of vocalic characteristics is presented here, comprised of c. 34 600 utterances of 70 speakers (46 non-professional speakers, children, women and men and 24 professional actors/actresses and singers of straight theatre, contemporary singing and European classical singing). The database focuses on sounds of the long Standard German vowels /i–y–e–ø–ɛ–a–o–u/ produced with varying basic production parameters such as phonation type, vocal effort, fundamental frequency, vowel context and speaking or singing style. In addition, a read text and, for professionals, songs are also included. The database is accessible for scientific use and further extensions are in progress.

## Weighting of Coda Voicing Cues: Glottalisation and Vowel Duration

*Joshua Penney[1,2], Felicity Cox[1,2] and Anita Szakay[1,2]*

[1]Centre for Language Sciences, Department of Linguistics, Macquarie University, Australia

[2]ARC Centre of Excellence in Cognition and its Disorders, Macquarie University, Australia

**Tue-O-2-4-6, Time: 16:10-16:30**

Recent research suggests that a trading relationship may exist in speech production between vowel duration and glottalisation as cues to coda stop voicing in Australian English. Younger speakers have been shown to use glottalisation to signal voicelessness more than older speakers who instead make greater use of vowel duration. This suggests a sound change in progress for the voicing cues. In addition, the vowel duration cue to voicing is greater in inherently long vowel contexts compared to inherently short vowel contexts. We report on a perceptual study designed to examine whether the weighting of these two cues found in production is replicated in perception.

Older and younger listeners were presented with audio stimuli co-varying in vowel duration and glottalisation. In accord with findings from production, the vowel duration cue was weaker for contexts containing inherently short vowels than for those containing inherently long vowels. Complementarily, glottalisation had a stronger effect on the perception of coda voicelessness in inherently short vowel contexts. Older and younger listeners did not differ in their use of glottalisation as a perceptual cue to voicelessness despite previously identified age differences in production. This finding raises

**Notes**

questions about the link between perception and production in sound change.

## Tue-O-2-5: Cognition and Brain Studies

MR G.03-G.04; 14:30-16:30; Tuesday, 4 September, 2018

Chairs: Jessie Nixon and Mark Liberman

### Revealing Spatiotemporal Brain Dynamics of Speech Production Based on EEG and Eye Movement

*Bin Zhao[1], Jinfeng Huang[2], Gaoyan Zhang[1], Jianwu Dang[1,2], Minbo Chen[1], YingjianFu[1] and Longbiao Wang[1]*

[1]Tianjin key Laboratory of Cognitive Computing and Application, Tianjin University, China

[2]Japan Advanced Institute of Science and Technology, Japan

**Tue-O-2-5-1, Time: 14:30-14:50**

To understand the neural circuitry associated with speech production in oral reading, it is essential to describe the whole-range spatiotemporal brain dynamics in the processes including visual word recognition, orthography-phonology mapping, semantic accessing, speech planning, articulation, self-monitoring, etc. This has turned out to be extremely difficult because of demanding resolution in both spatial and temporal domains and advanced algorithms to eliminate severe contamination by articulatory movements. To tackle this hard target, we recruited 16 subjects in a sentence reading task and measured multimodal signals of electroencephalography (EEG), eye movement and speech simultaneously. The onset/offset of gazing and utterance were used for segmenting brain activation stages. Cortical modeling of causal interactions among anatomical regions was conducted on EEG signals through (i) independent component analysis to identify cortical regions of interest (ROIs); (ii) multivariate autoregressive modeling of representative cortical activity from each ROI; and (iii) quantification of the dynamic causal interactions among ROIs using the Short-time direct Directed Transfer function. The resulting brain dynamic model reveals a widely connected bilateral organization with left-lateralized semantic, orthographic and phonological sub-networks, right-lateralized prosody and motor sequencing sub-networks and bi-lateralized auditory and multisensory integration sub-networks that cooperate along interlaced and paralleled temporal stages for speech processing.

### Neural Response Development During Distributional Learning

*Natalie Boll-Avetisyan[1], Jessie S. Nixon[2], Tomas O. Lentz[3], Liquan Liu[4,5], Sandrien van Ommen[6], Çağrı Çöltekin[2] and Jacolien van Rij[7]*

[1]University of Potsdam, Germany

[2]University of Tübingen, Germany

[3]University of Amsterdam, The Netherlands

[4]Western Sydney University, Australia

[5]Centre of Excellence for the Dynamics of Language, Australian Research Council, Australia

[6]Laboratoire Psychologie de la Perception, Université Paris Descartes, France

[7]University of Groningen, The Netherlands

**Tue-O-2-5-2, Time: 14:50-15:10**

We investigated online electrophysiological components of distributional learning, specifically of tones by listeners of a non-tonal language. German listeners were presented with a bimodal distribution of syllables with lexical tones from a synthesized continuum based on Cantonese level tones. Tones were presented in sets of four standards (within-category tokens) followed by a deviant (across-category token). Mismatch negativity (MMN) was measured. Earlier behavioral data showed that exposure to this bimodal distribution improved both categorical perception and perceptual acuity for level tones [1]. In the present study we present analyses of the electrophysiological response recorded during this exposure, i.e. the development of the MMN response during distributional learning. This development over time is analyzed using Generalized Additive Mixed Models and results showed that the MMN amplitude increased for both within- and across-category tokens, reflecting higher perceptual acuity accompanying category formation. This is evidence that learners zooming in on phonological categories undergo neural changes associated with more accurate phonetic perception.

### Learning Two Tone Languages Enhances the Brainstem Encoding of Lexical Tones

*Akshay Raj Maggu[1], Wenqing Zong[1], Vina Law[2] and Patrick C. M. Wong[1]*

[1]The Chinese University of Hong Kong

[2]University of Toronto

**Tue-O-2-5-3, Time: 15:10-15:30**

Auditory brainstem encoding is influenced by experience-dependent factors such as language and music. Tone language speakers exhibit more robust brainstem encoding of lexical tones than non-tone language speakers. Studies suggest that the effects of experience with a tone language generalize to the brainstem encoding of lexical tones from other tone languages. However, the effects of learning two tone languages, with different tonal systems, on brainstem encoding of lexical pitch are unknown. In the current study, we investigated whether or not the experience with two tone languages (Mandarin and Cantonese) enhances the brainstem encoding of lexical pitch, using frequency following response (FFR). Mandarin has four lexical tones- high level, rising, dipping and falling while Cantonese has a richer tone system with three level tones (high, mid, low), two rising tones (high and low) and one falling tone. We compared speakers fluent in Cantonese vs. those fluent in both Cantonese and Mandarin on their brainstem encoding of Cantonese and Mandarin lexical tones. We found that the Cantonese-Mandarin speakers exhibited more robust brainstem encoding of the lexical tones as compared to Cantonese speakers. From the current findings, we conclude that learning two tone languages may enhance lexical pitch encoding at the brainstem.

### Perceptual Sensitivity to Spectral Change in Australian English Close Front Vowels: An Electroencephalographic Investigation

*Daniel Williams[1,2,3], Paola Escudero[2,3] and Adamantios Gafos[1]*

[1]Linguistics Department, University of Potsdam, Germany

[2]MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Australia

[3]ARC Centre of Excellence for the Dynamics of Language, Australia

**Tue-O-2-5-4, Time: 15:30-15:50**

Speech scientists have long noted that the qualities of naturally-produced vowels do not remain constant over their durations – regardless of being nominally "monophthongs" or "diphthongs". Recent acoustic corpora show that there are consistent patterns of first (F1) and second (F2) formant frequency change across different vowel categories. The three Australian English (AusE) close front vowels /**iː, ɪ, ɪə**/ provide a striking example: while their midpoint or mean F1 and F2 frequencies are virtually identical, their spectral change patterns distinctly differ. The present study utilizes a pre-attentive discrimination paradigm with electroencephalography to assess AusE listeners' perceptual sensitivity to close front vowels with different F1 × F2 trajectory lengths (TLs) and directions (TDs). When TLs are modest, there is an asymmetry in perceptual sensitivity: closing vowels, e.g., /iː/ whose trajectory terminates high in the F1 × F2 vowel space, are perceptually prominent, whereas centering vowels, e.g., /**ɪ, ɪə**/ whose trajectories end more centrally, are not. However, when TLs are exaggerated, the asymmetry in the perceptual sensitivity to the two TDs is substantially reduced. The results indicate that, despite the distinct patterns of spectral change of AusE /**iː, ɪ, ɪə**/ in production, its perceptual relevance is not uniform, but rather vowel-category dependent.

**Notes**

### Effective Acoustic Cue Learning Is Not Just Statistical, It Is Discriminative

*Jessie S. Nixon*

Quantitative Linguistics, Eberhard Karls Universität Tübingen, Germany

**Tue-O-2-5-5, Time: 15:50-16:10**

A growing statistical learning literature suggests that listeners extract statistical information from the linguistic environment. However, distributional frequency may be insufficient for important but relatively low-frequency cues. Acquisition of linguistic knowledge may rely not merely on co-occurrences but on predictive relationships between cues and their outcomes. The present study investigates effects of predictive temporal cue structure on acquisition of a non-native acoustic cue dimension.

During training, native English speakers saw coloured shape objects and heard spoken Min Chinese words with six different lexical tones. Tones were the only reliable cue to identifying the associated object. Words also contained a salient cue that did not discriminate between objects. Three tones occurred with high-frequency and three with low-frequency in training. The critical manipulation was the presentation order: either words, containing complex cue structure, preceded object outcomes (discriminative order) or objects preceded words (non-discriminative order).

Generalised linear mixed models showed accuracy was significantly higher in the discriminative order than the non-discriminative order. These results demonstrate that predictive cue structure can facilitate acquisition of a non-native cue dimension. Feedback from prediction error drives learners to ignore salient non-discriminative cues and effectively learn to use the target cue dimension.

### Analyzing EEG Signals in Auditory Speech Comprehension Using Temporal Response Functions and Generalized Additive Models

*Kimberley Mulder, Louis ten Bosch and Lou Boves*

Center for Language Studies, Radboud University, Netherlands

**Tue-O-2-5-6, Time: 16:10-16:30**

Analyzing EEG signals recorded while participants are listening to continuous speech with the purpose of testing linguistic hypotheses is complicated by the fact that the signals simultaneously reflect exogenous acoustic excitation and endogenous linguistic processing. This makes it difficult to trace subtle differences that occur in mid-sentence position. We apply an analysis based on multivariate temporal response functions to uncover subtle mid-sentence effects. This approach is based on a per-stimulus estimate of the response of the neural system to speech input. Analyzing EEG signals predicted on the basis of the response functions might then bring to light condition specific differences in the filtered signals. We validate this approach by means of an analysis of EEG signals recorded with isolated word stimuli. Then, we apply the validated method to the analysis of the responses to the same words in the middle of meaningful sentences.

## Tue-SS-2-1: Special Session: Deep Neural Networks: How Can We Interpret What They Learned?

MR 1.01-1.02; 14:30-16:30; Tuesday, 4 September, 2018

Chairs: Louis ten Bosch, Hugo van Hamme and Lou Boves

### Information Encoding by Deep Neural Networks: What Can We Learn?

*Louis ten Bosch[1,2] and Lou Boves[1]*

[1]Radboud University Nijmegen, NL

[2]Max Planck Institute for Psycholinguistics

**Tue-SS-2-1-1, Time: 14:30-14:50**

The recent advent of deep learning techniques in speech technology and in particular in automatic speech recognition has yielded substantial performance improvements. This suggests that deep neural networks (DNNs) are able to capture structure in speech data that older methods for acoustic modeling, such as Gaussian Mixture Models and shallow neural networks fail to uncover. In image recognition it is possible to link representations on the first couple of layers in DNNs to structural properties of images and to representations on early layers in the visual cortex. This raises the question whether it is possible to accomplish a similar feat with representations on DNN layers when processing speech input. In this paper we present three different experiments in which we attempt to untangle how DNNs encode speech signals and to relate these representations to phonetic knowledge, with the aim to advance conventional phonetic concepts and to choose the topology of a DNNs more efficiently. Two experiments investigate representations formed by auto-encoders. A third experiment investigates representations on convolutional layers that treat speech spectrograms as if they were images. The results lay the basis for future experiments with recursive networks.

### Scalable Factorized Hierarchical Variational Autoencoder Training

*Wei-Ning Hsu and James Glass*

Computer Science and Artificial Intelligence Laboratory

Massachusetts Institute of Technology, Cambridge, MA 02139, USA

**Tue-SS-2-1-2, Time: 14:50-15:10**

Deep generative models have achieved great success in unsupervised learning with the ability to capture complex nonlinear relationships between latent generating factors and observations. Among them, a factorized hierarchical variational autoencoder (FHVAE) is a variational inference-based model that formulates a hierarchical generative process for sequential data. Specifically, an FHVAE model can learn disentangled and interpretable representations, which have been proven useful for numerous speech applications, such as speaker verification, robust speech recognition and voice conversion. However, as we will elaborate in this paper, the training algorithm proposed in the original paper is not scalable to datasets of thousands of hours, which makes this model less applicable on a larger scale. After identifying limitations in terms of runtime, memory and hyperparameter optimization, we propose a hierarchical sampling training algorithm to address all three issues. Our proposed method is evaluated comprehensively on a wide variety of datasets, ranging from 3 to 1,000 hours and involving different types of generating factors, such as recording conditions and noise types. In addition, we also present a new visualization method for qualitatively evaluating the performance with respect to the interpretability and disentanglement. Models trained with our proposed algorithm demonstrate the desired characteristics on all the datasets.

### State Gradients for RNN Memory Analysis

*Lyan Verwimp, Hugo van Hamme, Vincent Renkens and Patrick Wambacq*

ESAT – PSI, KU Leuven, Belgium

**Tue-SS-2-1-3, Time: 15:10-15:30**

We present a framework for analyzing what the state in RNNs remembers from its input embeddings. Our approach is inspired by backpropagation, in the sense that we compute the gradients of the states with respect to the input embeddings. The gradient matrix is decomposed with Singular Value Decomposition to analyze which directions in the embedding space are best transferred to the hidden state space, characterized by the largest singular values. We apply our approach to LSTM language models and investigate to what extent and for how long certain classes of words are remembered on average for a certain corpus. Additionally, the extent to which a specific property or relationship is remembered by the RNN can be tracked by comparing a vector characterizing that property with the direction(s) in embedding space that are best preserved in hidden state space.

### Exploring How Phone Classification Neural Networks Learn Phonetic Information by Visualising and Interpreting Bottleneck Features

**Notes**

*Linxue Bai, Philip Weber, Peter Jančovič and Martin Russell*
School of Engineering, The University of Birmingham,
Birmingham B15 2TT, UK

**Tue-SS-2-1-4, Time: 15:30-15:50**

Neural networks have a reputation for being "black boxes", which it has been suggested that techniques from user interface development and visualisation in particular, could help lift.

In this paper, we explore 9-dimensional bottleneck features (BNFs) that have been shown in our earlier work to well represent speech in the context of speech recognition and 2-dimensional BNFs directly extracted from bottleneck neural networks. The 9-dimensional BNFs obtained from a phone classification neural network are visualised in 2-dimensional spaces using linear discriminant analysis (LDA) and t-distributed stochastic neighbour embedding (t-SNE). The 2-dimensional BNF space is analysed with regard to phonetic features. A back-propagation method is used to create "cardinal" features for each phone under a particular neural network. Both the visualisations of 9-dimensional and 2-dimensional BNFs show distinctions between most phone categories. Particularly, the 2-dimensional BNF space seems to be a union of phonetic category related subspaces that preserve local structures within each subspace where the organisations of phones appear to correspond to phone production mechanisms. By applying LDA to the features of higher dimensional non-bottleneck layers, we observe a triangular pattern which may indicate that silence, friction and voicing are the three main properties learned by the neural networks.

### Memory Time Span in LSTMs for Multi-Speaker Source Separation

*Jeroen Zegers and Hugo van Hamme*
KU Leuven, Dept. ESAT, Belgium

**Tue-SS-2-1-5, Time: 15:50-16:10**

With deep learning approaches becoming state-of-the-art in many speech (as well as non-speech) related machine learning tasks, efforts are being taken to delve into the neural networks which are often considered as a black box. In this paper it is analyzed how recurrent neural network (RNNs) cope with temporal dependencies by determining the relevant memory time span in a long short-term memory (LSTM) cell. This is done by leaking the state variable with a controlled lifetime and evaluating the task performance. This technique can be used for any task to estimate the time span the LSTM exploits in that specific scenario. The focus in this paper is on the task of separating speakers from overlapping speech. We discern two effects: A long term effect, probably due to speaker characterization and a short term effect, probably exploiting phone-size formant tracks.

### Visualizing Phoneme Category Adaptation in Deep Neural Networks

*Odette Scharenborg[1,2], Sebastian Tiesmeyer[1], Mark Hasegawa-Johnson[3] and Najim Dehak[4]*
[1]Centre for Language Studies, Radboud University, Nijmegen, the Netherlands
[2]Multimedia Computing Group, Delft University of Technology, Delft, the Netherlands
[3]ECE Department & Beckman Institute, University of Illinois, Urbana-Champaign, IL, USA
[4]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA

**Tue-SS-2-1-6, Time: 16:10-16:30**

Both human listeners and machines need to adapt their sound categories whenever a new speaker is encountered. This perceptual learning is driven by lexical information. The aim of this paper is two-fold: investigate whether a deep neural network-based (DNN) ASR system can adapt to only a few examples of ambiguous speech as humans have been found to do; investigate a DNN's ability to serve as a model of human perceptual learning. Crucially, we do so by looking at intermediate levels of phoneme category adaptation rather than at the output level. We visualize the activations in the hidden layers of the DNN during perceptual learning. The results show that, similar to humans, DNN systems learn speaker-adapted phone category boundaries from a few labeled examples. The DNN adapts its category boundaries not only by adapting the weights of the output layer, but also by adapting the implicit feature maps computed by the hidden layers, suggesting the possibility that human perceptual learning might involve a similar nonlinear distortion of a perceptual space that is intermediate between the acoustic input and the phonological categories. Comparisons between DNNs and humans can thus provide valuable insights into the way humans process speech and improve ASR technology.

**Notes**

## Tue-S&T-2-1: Show and Tell 4

MR G.05-G.06; 14:30-16:30; Tuesday, 4 September, 2018
Chairs: Rajagopal Jayaraman and Ajay Srinivasamurthy

### Early Vocabulary Development Through Picture-based Software Solutions

*Kasthuri G[1,2], Prabha Ramanathan[1,2], Hema Murthy[2], Namita Jacob[1] and Anil Prabhakar[1,2]*
[1]Enability Foundation for Rehabilitation, Chennai
[2]IIT Madras, Chennai 600036

**Tue-S&T-2-1-1, Time: 14:30-16:30**

Assistive technology enables children with disabilities to gain access, function independently and take advantage of schooling and social opportunities[1]. The need for alternative and augmentative communication (AAC) in all children is not the same[2] but AAC can aid in expressive language and also support intelligible speakers in developing and using communication skills in varied situations. While the range and flexibility of AAC has grown over the years, making devices accessible to children at varied economic and regional backgrounds, is still a challenge.

KAVI-PTS is designed as a picture-to-speech Android application and has been made available in several Asian languages. This application is conceived of as an inexpensive software alternative to communication charts. It can be easily configured to adjust contrast levels and customize selection modes, enabling children to have access to a tailor-made communication solution.

### Automatic Detection of Expressiveness in Oral Reading

*Kamini Sabu, Kanhaiya Kumar and Preeti Rao*
Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, India

**Tue-S&T-2-1-2, Time: 14:30-16:30**

We present a Computer-Aided Language Learning (CALL) system that assesses a child's oral reading skill including the prosodic aspects. With children who have otherwise achieved word decoding automaticity, prosodic fluency is a reliable predictor of comprehension. Prosody includes attributes such as pace, phrasing and expression. Based on the acoustic correlates of prosodic events, we propose and test features that discriminate expressive speech from monotonous speech and further detect whether the expression is meaningful or simply a rhythmic cadence with no relation to the underlying syntax or semantics of the text. Finally the system based on processing short samples of recorded oral reading and providing feedback on the goodness of both lexical and prosodic aspects is described.

### PannoMulloKathan: Voice Enabled Mobile App for Agricultural Commodity Price Dissemination in Bengali Language

*Madhab Pal, Rajib Roy, Soma Khan, Milton S. Bepari and Joyanta Basu*

Centre for Development of Advanced Computing, Kolkata, India

**Tue-S&T-2-1-3, Time: 14:30-16:30**

In this work we present a voice based mobile application for dissemination of agricultural commodity procurement and consumer prices. Disbursed information is crawled at daily basis from government authorized websites of agricultural marketing departments. The app incorporates mix media multiple access means in form of touch-type-see, touch-type-listen, speak-see and speak-listen modalities and also includes a robust Automatic Speech Recognition (ASR) engine in Bengali language to support real time voice queries. Colorful interactive app based user interface and ASR incorporated core client-server architecture altogether provides an efficient framework for serving registered users of different educational and economical background including people having little or no computer knowledge, semi-literate or illiterate rural people.

### Visualizing Punctuation Restoration in Speech Transcripts with Prosograph

*Alp Öktem[1], Mireia Farrús[1] and Antonio Bonafonte[2]*

[1]Universitat Pompeu Fabra, Spain

[2]Universitat Politècnica de Catalunya, Spain

**Tue-S&T-2-1-4, Time: 14:30-16:30**

We have developed a neural architecture that tests the effect of lexical, morphosyntactic and prosodic features in restoring punctuation in speech transcriptions. Having outperformed a baseline model in terms of precision and recall, we further extend our performance tests by attaching it in a speech recognition pipeline. The visual and interactive testing environment that we prepared helps us observe how our models generalizes in unseen data and also plan our next steps for improvement.

### CACTAS - Collaborative Audio Categorization and Transcription for ASR Systems

*Mithul Mathivanan, Kinnera Saranu, Abhishek Pandey and Jithendra Vepa*

Samsung R&D Institute, India - Bangalore

**Tue-S&T-2-1-5, Time: 14:30-16:30**

We present a web based tool that allows collaborative analysis and/or transcription of audios with respect to Automatic Speech Recognition (ASR) systems. The tool presents a webpage consisting of audios and their corresponding references and hypotheses obtained offline. Several other information and features are provided that allow the audios to be categorized and references to be corrected efficiently in a collaborative way almost 10 times faster, without the need for prior knowledge on speech or ASR systems. The analysis can later be summarized and acted upon to improve or triage the ASR system.

## Tue-P-2-1: Speech and Singing Production

Hall 4-6: Poster-1; 14:30-16:30; Tuesday, 4 September, 2018

Chair: T V Sreenivas

### FACTS: A Hierarchical Task-based Control Model of Speech Incorporating Sensory Feedback

*Benjamin Parrell[3,4], Vikram Ramanarayanan[1,2,4], Srikantan Nagarajan[2] and John Houde[2]*

[1]Educational Testing Service R&D, San Francisco, CA

[2]University of California, San Francisco, CA

[3]University of Wisconsin, Madison, WI

[4]Joint first authors, contributed equally to this work.

**Tue-P-2-1-1, Time: 14:30-16:30**

We present a computational model of speech motor control that integrates vocal tract state prediction with sensory feedback. This hierarchical model, called FACTS, incorporates both a high-level and low-level controller. The high-level controller orchestrates linguistically-relevant speech tasks, which are represented as desired constrictions along the vocal tract (e.g., closure of the lips). The output of the high-level controller is passed to a low-level controller that can issue motor commands at the level of the speech articulators in order to accomplish the desired constrictions. In order to generate these articulatory motor commands, this low-level articulatory controller relies on an estimate of the current state of the vocal tract. This estimate combines internal predictions about the consequences of issued motor commands with auditory and somatosensory feedback from the vocal tract using an Unscented Kalman Filter based state estimation method. FACTS is able to replicate important aspects of human speech behavior, in that it reproduces: (i) stable speech behavior in the presence of noisy motor and sensory systems, (ii) partial acoustic compensation to auditory feedback perturbations, (iii) complete compensations to mechanical perturbations only when they interfere with current production goals and (iv) the observed relationship between sensory acuity and response to sensory perturbations.

### Sensorimotor Response to Tongue Displacement Imagery by Talkers with Parkinson's Disease

*William Katz[1], Patrick Reidy[1] and Divya Prabhakaran[2]*

[1]University of Texas at Dallas

[2]Vanderbilt University, Tennessee

**Tue-P-2-1-2, Time: 14:30-16:30**

In a previous study, we asked healthy adult speakers to produce the word head under noise-masked (visual only) conditions and while watching videos of a 3D tongue avatar that gradually morphed from producing head to had. Results indicated that during the visual mismatch phases all participants entrained to the visually presented word, head, without being aware that their vowel quality had changed. Here, we explore whether similar effects occur for individuals with presumed sensor neural processing disorders, patients with Parkinson's disease (PD). We also examine the effects of PD treatment on this entrainment behavior. Participants were 14 individuals with PD, with eight in ongoing speech/language therapy and six reporting no recent therapy. Participants heard pink noise over headphones and produced the word head under four viewing conditions: First, while viewing repetitions of head (baseline); next, during "morphed" videos shifting gradually from head to had (ramp); then videos of had (maximum hold); and finally videos of head (after effects). Analysis with a linear mixed effects model indicated a significant F1 difference between baseline and maximum hold phases for the productions of the treated PD group, but not for the untreated group. Implications for the causes and treatment of PD speech disorders are discussed.

### Automatic Pronunciation Evaluation of Singing

*Chitralekha Gupta[1], Haizhou Li[2] and Ye Wang[1]*

[1]School of Computing and NUS Graduate School for Integrated Science and Engineering

[2]Department of Electrical and Computer Engineering

National University of Singapore

**Tue-P-2-1-3, Time: 14:30-16:30**

In this work, we develop a strategy to automatically evaluate pronunciation of singing. We apply singing-adapted automatic speech recognizer (ASR) in a two-stage approach for evaluating pronunciation of singing. First, we force-align the lyrics with the sung utterances to obtain the word boundaries. We improve the word boundaries by a novel lexical modification technique. Second, we investigate the performance of the phonetic posteriorgram (PPG) based template independent and dependent methods for scoring the aligned words. To validate the evaluation scheme, we obtain reliable human pronunciation evaluation scores using a crowd-sourcing platform. We show that the automatic evaluation scheme offers quality scores that are close to human judgments.

### Classification of Nonverbal Human Produced Audio Events: A Pilot Study

*Rachel E. Bouserhal[1,4], Philippe Chabot[1], Milton Sarria-Paja[3], Patrick Cardinal[2] and Jérémie Voix[1,4]*

**Notes**

École de technologie supérieure, Departments of
[1]Mechanical Engineering and
[2]Software and Information Technology Engineering, 1100 Notre-Dame St W, Montréal, Québec, Canada
[3]Universidad Santiago de Cali Calle 5 No 62-00 Pampalinda, Cali, Colombia
[4]Centre for Interdisciplinary Research in Music Media and Technology, 527 Rue Sherbrooke O, Montréal, Québec, Canada

**Tue-P-2-1-4, Time: 14:30-16:30**

The accurate classification of nonverbal human produced audio events opens the door to numerous applications beyond health monitoring. Voluntary events, such as tongue clicking and teeth chattering, may lead to a novel way of silent interface command. Involuntary events, such as coughing and clearing the throat, may advance the current state-of-the-art in hearing health research. The challenge of such applications is the balance between the processing capabilities of a small intra-aural device and the accuracy of classification. In this pilot study, 10 nonverbal audio events are captured inside the ear canal blocked by an intra-aural device. The performance of three classifiers is investigated: Gaussian Mixture Model (GMM), Support Vector Machine and Multi-Layer Perceptron. Each classifier is trained using three different feature vector structures constructed using the mel-frequency cepstral (MFCC) coefficients and their derivatives. Fusion of the MFCCs with the auditory-inspired amplitude modulation features (AAMF) is also investigated. Classification is compared between binaural and monaural training sets as well as for noisy and clean conditions. The highest accuracy is achieved at 75.45% using the GMM classifier with the binaural MFCC+AAMF clean training set. Accuracy of 73.47% is achieved by training and testing the classifier with the binaural clean and noisy dataset.

## UltraFit: A Speaker-friendly Headset for Ultrasound Recordings in Speech Science

*Lorenzo Spreafico[1], Michael Pucher[2] and Anna Matosova[1]*
[1]Free University of Bozen-Bolzano, Italy
[2]Acoustics Research Institute, Vienna, Austria

**Tue-P-2-1-5, Time: 14:30-16:30**

UltraFit is a headset for Ultrasound Tongue Imaging (UTI) printed in Nylon; altogether, it weighs about 350 g. It was developed through an iterative process of rapid prototyping a proof of concept, asking for feedback from researchers and subjects of the experiments and instantly incorporating changes based on their feedback into the design. We evaluated the UltraFit headset by recording a speaker using an optical marker tracking system that provides sub-millimeter tracking accuracy. We show that the overall error range of the headset movement for this speaker lies within 3mm with most errors lying in a 1-2mm range. This makes the headset potentially suitable for speech science applications. Furthermore, we analyze the superior usability of the headset compared to other existing designs and describe the headsets development process.

## Articulatory Consequences of Vocal Effort Elicitation Method

*Elisabet Eir Cortes[1], Marcin Wlodarczak[1] and Juraj Šimko[2]*
[1]Department of Linguistics, Stockholm University, Stockholm, Sweden
[2]Department of Digital Humanities, University of Helsinki, Helsinki, Finland

**Tue-P-2-1-6, Time: 14:30-16:30**

Articulatory features from two datasets, Slovak and Swedish, were compared to see whether different methods of eliciting loud speech (ambient noise vs. visually presented loudness target) result in different articulatory behavior. The features studied were temporal and kinematic characteristics of lip separation within the closing and opening gestures of bilabial consonants and of the tongue body movement from /i/ to /a/ through a bilabial consonant. The results indicate larger hyper-articulation in the speech elicited with visually presented target. While individual articulatory strategies are evident, the speaker groups agree on increasing the kinematic features consistently within each gesture in response to the increased vocal effort. Another concerted strategy is keeping the tongue response considerably smaller than that of the lips, presumably to preserve acoustic prerequisites necessary for the adequate vowel identity. While the method of visually presented loudness target elicits larger span of vocal effort, the two elicitation methods achieve comparable consistency per loudness conditions.

## Age-related Effects on Sensorimotor Control of Speech Production

*Anne Hermes, Jane Mertens and Doris Mücke*
IfL – Phonetics, University of Cologne, Germany

**Tue-P-2-1-7, Time: 14:30-16:30**

The current study investigates the effect of aging on the speech motor control, more specifically the labial and lingual system. We provide an acoustic and articulatory analysis comparing younger (20-30 years old) and older speakers (70-80 years old) of German, all of them recorded with electromagnetic articulography. We analyzed target words in contrastive focus condition.

In the acoustic domain, target syllables were not prolonged in the productions of the older speakers. However, when looking at the articulatory domain, we found systematic modifications: Especially vocalic gestures, requiring movements of the lingual system, showed slower peak velocities for older subjects. Furthermore, we found age-related effects on the symmetry of articulatory gestures. Older subjects produce longer deceleration and shorter acceleration phases leading to a strong asymmetry of the movement components. Variability between and across speakers were considerably higher in the group of older speakers compared to younger ones.

Our results on age-related effects on speech motor control are comparable with those from general motor control, where e.g. prolonged deceleration phases are an indicator for a decrease in sensory feedback control.

## An Ultrasound Study of Gemination in Coronal Stops in Eastern Oromo

*Maida Percival, Alexei Kochetov and Yoonjung Kang*
University of Toronto

**Tue-P-2-1-8, Time: 14:30-16:30**

This study extends the use of ultrasound methodology to stops in Eastern Oromo (Cushitic; Ethiopia) to examine the link between gemination, laryngeal features and tongue shape.

Ultrasound data were collected from 5 native speakers of Eastern Oromo. Tokens consisted of 12 repetitions per speaker of [t$^{h}$, t, d, ɗ] and six of [tt$^{h}$, tt, dd, ɗɗ] in the environment of a_a. Tongue images at the point of maximum constriction during the stop closure were traced following Kochetov et al. (2014) and their coordinates submitted to linear mixed effects models.

Results indicated differences in tongue shape between singletons and geminates, especially for ejectives and implosives. Singleton ejectives displayed raised tongue bodies not found in geminate ejectives. Singleton implosives resembled voiceless stops, but geminate implosives were variably produced with tongue body raising.

I suggest that the results can be attributed to fortition in geminates. Tongue body raising in singleton ejectives may be an enhancement strategy to the ejective contrast that is not necessary in longer geminates. The singleton implosive resembling a voiceless aspirated stop is predicted by Lloret (1994) while the geminate tongue body raising may be retraction, c.f. Payne (2006). The results support a link between tongue, larynx and gemination.

## Processing Transition Regions of Glottal Stop Substituted /S/ for Intelligibility Enhancement of Cleft Palate Speech

*Protima Nomo Sudro[1], Sishir Kalita[1] and S R Mahadeva Prasanna[1,2]*
[1]Indian Institute of Technology Guwahati, Guwahati, India
[2]Indian Institute of Technology Dharwad, Dharwad, India

**Notes**

**Tue-P-2-1-9, Time: 14:30-16:30**
The speech intelligibility of cleft palate (CP) individuals is degraded primarily due to compensatory articulation errors and hyper nasality. The present work proposes a method to enhance the CP speech intelligibility, where fricatives are substituted by compensatory articulation errors. Apart from the distortion present in the sustained fricative region, the fricative-vowel and vowel-fricative regions are also deviated due to co-articulation effect. Since important perceptual cues are embedded in the transition regions. Therefore, it is necessary to enhance the transition regions for more intelligibility. Motivated by the perceptual significance of the transition regions, 2D-DCT based joint spectro-temporal features are exploited for the modification. The 2D DCT coefficients of CP speech are modified by projecting them onto the singular vectors derived from the SVD analysis of normal speech. Further, for the evaluation of speech intelligibility, objective and subjective assessment is conducted. The results show significant improvement in the speech intelligibility of the modified speech.

### Reconstructing Neutral Speech from Tracheoesophageal Speech

*Abinay Reddy N, Achuth Rao MV, G. Nisha Meenakshi and Prasanta Kumar Ghosh*

Electrical Engineering, Indian Institute of Science (IISc), Bangalore-560012, India

**Tue-P-2-1-10, Time: 14:30-16:30**
In this work, we propose a tracheoesophageal (TE) speech to neutral speech conversion system using data collected from a laryngectomee. In laryngectomees, in the absence of vocal folds, it is the vibration of the esophagus that gives rise to a low-frequency pitch during speech production. This pitch is manifested as impulse-like noise in the recorded speech. We propose a method to first 'whisperize' the TE speech prior to the linear predictive coding (LPC) based synthesis which uses pitch derived from the energy contour. In order to perform 'whisperization', we model the LPC residual signal as the sum of white noise and impulses introduced by the esophageal vibrations. We model these impulses and white noise using Bernoulli-Gaussian distribution and Gaussian distribution, respectively. The strength and location of the impulses are estimated using Gibbs sampling in order to remove the impulse-like noise from speech to obtain whispered speech. Subjective evaluation via listening test reveals that the 'whisperization' step in the proposed method aids in synthesizing a more natural sounding neutral speech. A different listening test shows that the listeners prefer the synthesized speech from the proposed method ~ 93% (absolute) times more than the best baseline scheme.

### Automatic Evaluation of Soft Articulatory Contact for Stuttering Treatment

*Keiko Ochi[1], Koichi Mori[2] and Naomi Sakai[2]*

[1]Tokyo University of Technology, Hachioji, Japan

[2]National Rehabilitation Center for Persons with Disabilities, Tokorozawa, Japan

**Tue-P-2-1-11, Time: 14:30-16:30**
We describe a new method for the automatic discrimination and evaluation of phonation beginning with a consonant with soft articulatory contact, which is used in the treatment of stuttering and normal phonation. Soft articulatory contact is trained to relax articulators and remove hard contacts that occur during stuttering. We use features related to the changes in acoustic characteristics and the voice quality under the hypothesis that the slowing down of articulatory movement of the initial consonant and the relaxing of phonatory muscles co-occur with soft articulatory contact. The results of an experimental evaluation showed that high accuracy was obtained when acoustic features were related to the peaks of the first derivative of the mel frequency cepstral coefficients (MFCCs) corresponded to the slowing down of the movement of the articulators. The features of vocal quality only slightly contributed to the classification.

### Korean Singing Voice Synthesis Based on an LSTM Recurrent Neural Network

*Juntae Kim[1], Heejin Choi[1], Jinuk Park[1], Minsoo Hahn[1], Sangjin Kim[2] and Jong-Jin Kim[2]*

[1]School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea

[2]SK Telecom, Seoul, Korea

**Tue-P-2-1-12, Time: 14:30-16:30**
Singing voice synthesis (SVS) systems generate the singing voice from a musical score. Similar to the text-to-speech synthesis (TTS) field, SVS systems have also been greatly improved since the deep neural network (DNN) framework was introduced. Although they share many parts of the framework, the main difference between TTS and SVS systems is that the feature composing method, between linguistic and musical features, is important for SVS systems. In this paper, we propose a Korean SVS system based on a long-short term memory recurrent neural network (LSTM-RNN). At the feature composing stage, we propose a novel composing method, based on Korean syllable structure. At the synthesis stage, we adopt LSTM-RNN for the SVS. According to our experiments, our composed feature improved the naturalness of the voice, specifically in any part that has to be pronounced for a long time. Furthermore, LSTM-RNN outperformed the DNN based SVS system in both quantitative and qualitative evaluations.

### The Trajectory of Voice Onset Time with Vocal Aging

*Chen Xuanda[1,2], Xiong Ziyu[2] and Hu Jian[3]*

[1]Department of Linguistics, Graduate School of Chinese Academy of Social Sciences, China

[2]Institute of Linguistics, Chinese Academy of Social Sciences, China

[3]School of Foreign Studies, Anhui University, China

**Tue-P-2-1-13, Time: 14:30-16:30**
Vocal aging, a universal process of human aging, can largely affect one's language use, possibly including some subtle acoustic features of one's utterances like Voice Onset Time. To figure out the time effects, Queen Elizabeth's Christmas speeches are documented and analyzed in the long-term trend. We build statistical models of time dependence in Voice Onset Time, controlling a wide range of other fixed factors, to present annual variations and the simulated trajectory. It is revealed that the variation range of Voice Onset Time has been narrowing over fifty years with a slight reduction in the mean value, which, possibly, is an effect of diminishing exertion, resulting from subdued muscle contraction, transcending other non-linguistic factors in forming Voice Onset Time patterns over a long time.

.....

## Tue-P-2-2: Robust Speech Recognition

Hall 4-6: Poster-2; 14:30-16:30; Tuesday, 4 September, 2018
Chair: Rohit Prabhavalkar

.....

### The Fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, Task and Baselines

*Jon Barker[1], Shinji Watanabe[2], Emmanuel Vincent[3] and Jan Trmal[2]*

[1]University of Sheffield, UK

[2]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, USA

[3]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

**Tue-P-2-2-1, Time: 14:30-16:30**
The CHiME challenge series aims to advance robust automatic speech recognition (ASR) technology by promoting research at the interface of speech and language processing, signal processing and machine learning. This paper introduces the 5th CHiME Challenge, which considers the task of distant multi-microphone conversational ASR in real home environments. Speech material was elicited using a dinner party scenario with efforts taken to capture data that is representative of natural conversational speech and recorded by 6 Kinect microphone arrays and 4 binaural microphone pairs. The challenge features a single-array track and a multiple-array track and, for each track, distinct rankings will be produced for systems focusing on

**Notes**

robustness with respect to distant-microphone capture versus systems attempting to address all aspects of the task including conversational language modeling. We discuss the rationale for the challenge and provide a detailed description of the data collection procedure, the task and the baseline systems for array synchronization, speech enhancement and conventional and end-to-end ASR.

## Voices Obscured in Complex Environmental Settings (VOiCES) Corpus

*Colleen Richey[2] and Maria A. Barrios[1], Zeb Armstrong[2], Chris Bartels[2], Horacio Franco[2], Martin Graciarena[2], Aaron Lawson[2], Mahesh Kumar Nandwana[2], Allen Stauffer[2], Julien van Hout[2], Paul Gamble[1], Jeffrey Hetherly[1], Cory Stephenson[1] and Karl Ni[1]*

[1]Lab41, In-Q-Tel Laboratories, Menlo Park, CA 94025

[2]SRI International, Menlo Park, CA 94025

**Tue-P-2-2-2, Time: 14:30-16:30**

This paper introduces the Voices Obscured In Complex Environmental Settings (VOiCES) corpus, a freely available dataset under Creative Commons BY 4.0. This dataset will promote speech and signal processing research of speech recorded by far-field microphones in noisy room conditions. Publicly available speech corpora are mostly composed of isolated speech at close-range microphony. A typical approach to better represent realistic scenarios, is to convolve clean speech with noise and simulated room response for model training. Despite these efforts, model performance degrades when tested against uncurated speech in natural conditions. For this corpus, audio was recorded in furnished rooms with background noise played in conjunction with foreground speech selected from the LibriSpeech corpus. Multiple sessions were recorded in each room to accommodate for all foreground speech-background noise combinations. Audio was recorded using twelve microphones placed throughout the room, resulting in 120 hours of audio per microphone. This work is a multi-organizational effort led by SRI International and Lab41 with the intent to push forward state-of-the-art distant microphone approaches in signal processing and speech recognition.

## Building State-of-the-art Distant Speech Recognition Using the CHiME-4 Challenge with a Setup of Speech Enhancement Baseline

*Szu-Jui Chen, Aswin Shanmugam Subramanian, Hainan Xu and Shinji Watanabe*

Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD 21218, USA

**Tue-P-2-2-3, Time: 14:30-16:30**

This paper describes a new baseline system for automatic speech recognition (ASR) in the CHiME-4 challenge to promote the development of noisy ASR in speech processing communities by providing 1) state-of-the-art system with a simplified single system comparable to the complicated top systems in the challenge, 2) publicly available and reproducible recipe through the main repository in the Kaldi speech recognition toolkit. The proposed system adopts generalized eigenvalue beamforming with bidirectional long short-term memory (LSTM) mask estimation. We also propose to use a time delay neural network (TDNN) based on the lattice-free version of the maximum mutual information (LF-MMI) trained with augmented all six microphones plus the enhanced data after beamforming. Finally, we use a LSTM language model for lattice and n-best re-scoring. The final system achieved 2.74% WER for the real test set in the 6-channel track, which corresponds to the 2nd place in the challenge. In addition, the proposed baseline recipe includes four different speech enhancement measures, short-time objective intelligibility measure (STOI), extended STOI (eSTOI), perceptual evaluation of speech quality (PESQ) and speech distortion ratio (SDR) for the simulation test set. Thus, the recipe also provides an experimental platform for speech enhancement studies with these performance measures.

## Unsupervised Adaptation with Interpretable Disentangled Representations for Distant Conversational Speech Recognition

*Wei-Ning Hsu, Hao Tang and James Glass*

Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

**Tue-P-2-2-4, Time: 14:30-16:30**

The current trend in automatic speech recognition is to leverage large amounts of labeled data to train supervised neural network models. Unfortunately, obtaining data for a wide range of domains to train robust models can be costly. However, it is relatively inexpensive to collect large amounts of unlabeled data from domains that we want the models to generalize to. In this paper, we propose a novel unsupervised adaptation method that learns to synthesize labeled data for the target domain from unlabeled in-domain data and labeled out-of-domain data. We first learn without supervision an interpretable latent representation of speech that encodes linguistic and nuisance factors (e.g., speaker and channel) using different latent variables. To transform a labeled out-of-domain utterance without altering its transcript, we transform the latent nuisance variables while maintaining the linguistic variables. To demonstrate our approach, we focus on a channel mismatch setting, where the domain of interest is distant conversational speech and labels are only available for close-talking speech. Our proposed method is evaluated on the AMI dataset, outperforming all baselines and bridging the gap between unadapted and in-domain models by over 77% without using any parallel data.

## Investigating Generative Adversarial Networks Based Speech Dereverberation for Robust Speech Recognition

*Ke Wang[1,2], Junbo Zhang[2], Sining Sun[1], Yujun Wang[2], Fei Xiang[2] and Lei Xie[1]*

[1]Shaanxi Provincial Key Laboratory of Speech and Image Information Processing, School of Computer Science, Northwestern Polytechnical University, Xi'an, China

[2]Xiaomi, Beijing, China

**Tue-P-2-2-5, Time: 14:30-16:30**

We investigate the use of generative adversarial networks (GANs) in speech dereverberation for robust speech recognition. GANs have been recently studied for speech enhancement to remove additive noises, but there still lacks of a work to examine their ability in speech dereverberation and the advantages of using GANs have not been fully established. In this paper, we provide deep investigations in the use of GAN-based dereverberation front-end in ASR. First, we study the effectiveness of different dereverberation networks (the generator in GAN) and find that LSTM leads a significant improvement as compared with feed-forward DNN and CNN in our dataset. Second, further adding residual connections in the deep LSTMs can boost the performance as well. Finally, we find that, for the success of GAN, it is important to update the generator and the discriminator using the same mini-batch data during training. Moreover, using reverberant spectrogram as a condition to discriminator, as suggested in previous studies, may degrade the performance. In summary, our GAN-based dereverberation front-end achieves 14%~19% relative CER reduction as compared to the baseline DNN dereverberation network when tested on a strong multi-condition training acoustic model.

## Monaural Multi-Talker Speech Recognition with Attention Mechanism and Gated Convolutional Networks

*Xuankai Chang[1], Yanmin Qian[1] and Dong Yu[2]*

[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China

[2]Tencent AI Lab, Tencent, Bellevue, WA, USA

**Tue-P-2-2-6, Time: 14:30-16:30**

To improve the speech recognition accuracy under the multi-talker

**Notes**

scenario, we propose a novel model architecture that incorporates the attention mechanism and gated convolutional network (GCN) into our previously developed permutation invariant training based multi-talker speech recognition system (PIT-ASR). The new architecture has three components: an encoding transformer, an attention module and a frame-level senone predictor. The encoding transformer first transforms a mixed speech sequence into a sequence of embedding vectors. Then the attention mechanism extracts individual context vectors from this embedding sequence for different speaker sources. Finally the predictor generates the senone posteriors for all speaker sources independently with the knowledge from the context vectors. To get better embedding representations we explore gated convolutional networks in the encoding transformer. The experimental results on the artificially mixed two-talker WSJ0 corpus show that our proposed model can reduce the word error rate (WER) by more than 15% relatively compared to our previous PIT-ASR system.

## Weighting Time-Frequency Representation of Speech Using Auditory Saliency for Automatic Speech Recognition

*Cong-Thanh Do and Yannis Stylianou*

Toshiba Cambridge Research Laboratory, Cambridge, United Kingdom

**Tue-P-2-2-7, Time: 14:30-16:30**

This paper proposes a new method for weighting two-dimensional (2D) time-frequency (T-F) representation of speech using auditory saliency for noise-robust automatic speech recognition (ASR). Auditory saliency is estimated via 2D auditory saliency maps which model the mechanism for allocating human auditory attention. These maps are used to weight T-F representation of speech, namely the 2D magnitude spectrum or spectrogram, prior to features extraction for ASR. Experiments on Aurora-4 corpus demonstrate the effectiveness of the proposed method for noise-robust ASR. In multi-stream ASR, relative word error rate (WER) reduction of up to 5.3% and 4.0% are observed when comparing the multi-stream system using the proposed method with the baseline single-stream system not using T-F representation weighting and that using conventional spectral masking noise-robust technique, respectively. Combining the multi-stream system using the proposed method and the single-stream system using the conventional spectral masking technique reduces further the WER.

## Acoustic Modeling from Frequency Domain Representations of Speech

*Pegah Ghahremani[1], Hossein Hadian[1,3], Hang Lv[1,4], Daniel Povey[1,2] and Sanjeev Khudanpur[1,2]*

[1]Center of Language and Speech Processing, Johns Hopkins University, Baltimore, MD

[2]Human Language Technology Center Of Excellence, Johns Hopkins University, Baltimore, MD

[3]Department of Computer Engineering, Sharif University of Technology, Iran

[4]School of Computer Science, Northwestern Polytechnical University, Xian, China

**Tue-P-2-2-8, Time: 14:30-16:30**

In recent years, different studies have proposed new methods for DNN-based feature extraction and joint acoustic model training and feature learning from raw waveform for large vocabulary speech recognition. However, conventional pre-processed methods such as MFCC and PLP are still preferred in the state-of-the-art speech recognition systems as they are perceived to be more robust. Besides, the raw waveform methods - most of which are based on the time-domain signal - do not significantly outperform the conventional methods.

In this paper, we propose a frequency-domain feature-learning layer which can allow acoustic model training directly from the waveform. The main distinctions from previous works are a new normalization block and a short-range constraint on the filter weights.

The proposed setup achieves consistent performance improvements compared to the baseline MFCC and log-Mel features as well as other proposed time and frequency domain setups on different LVCSR tasks. Finally, based on the learned filters in our feature-learning layer, we propose a new set of analytic filters using polynomial approximation, which outperforms log-Mel filters significantly while being equally fast.

## Non-Uniform Spectral Smoothing for Robust Children's Speech Recognition

*Ishwar Chandra Yadav, Avinash Kumar, Syed Shahnawazuddin and Gayadhar Pradhan*

Department of Electronics and Communication Engineering

National Institute of Technology Patna, India

**Tue-P-2-2-9, Time: 14:30-16:30**

Insufficient spectral smoothing during front-end speech parametrization results in pitch-induced distortions in the short-time magnitude spectra. This, in turn, degrades the performance of an automatic speech recognition (ASR) system for high-pitched speakers. Motivated by this fact, a non-uniform spectral smoothing algorithm is proposed in this paper in order to mitigate the acoustic mismatch resulting from pitch differences. In the proposed technique, the speech utterance is first segmented into vowel and non-vowel regions. The short-time magnitude spectrum obtained by discrete Fourier transform is then processed through a single-pole low-pass filter with different pole values for vowel and non-vowel regions. Sufficiently smoothed spectra is obtained by keeping higher values for the pole in the case of vowels while lower values are chosen for non-vowel regions. The Mel-frequency cepstral coefficients computed using the derived smoothed spectra are observed to be less affected by pitch variations. In order to validate this claim, an ASR system is developed on speech from adult speakers and evaluated on a test set which consists of children's speech to simulate large pitch differences. The experimental evaluations as well as signal domain analyses presented in this paper support the claim.

## Bidirectional Long-Short Term Memory Network-based Estimation of Reliable Spectral Component Locations

*Aaron Nicolson and Kuldip K. Paliwal*

Signal Processing Laboratory, Griffith University, Brisbane, Australia

**Tue-P-2-2-10, Time: 14:30-16:30**

An accurate Ideal Binary Mask (IBM) estimate is essential for Missing Feature Theory (MFT)-based speaker identification, as incorrectly labelled spectral components (where a component is either reliable or unreliable) will degrade the performance of an Automatic Speaker Identification (ASI) system adversely in the presence of noise. In this work a Bidirectional Recurrent Neural Network (BRNN) with Long-Short Term Memory (LSTM) cells is proposed for improved IBM estimation. The proposed system had an average IBM estimate accuracy improvement of 4.5% and an average MFT-based speaker identification accuracy improvement of 3.1% over all tested SNR dB levels, when compared to the previously proposed Multilayer Perceptron (MLP)-IBM estimator. When used for speech enhancement the proposed system had an average MOS-LQO (objective quality measure) improvement of 0.32 and an average QSTI (objective intelligibility measure) improvement of 0.01 over all tested SNR dB levels, when compared to the MLP-IBM estimator. The results presented in this work highlight the effectiveness of the proposed BRNN-IBM estimator for MFT-based speaker identification and IBM-based speech enhancement.

## Speech Emotion Recognition by Combining Amplitude and Phase Information Using Convolutional Neural Network

*Lili Guo[1], Longbiao Wang[1], Jianwu Dang[1,2], Linjuan Zhang[1], Haotian Guan[3] and Xiangang Li[4]*

[1]Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin University, Tianjin, China

[2]Japan Advanced Institute of Science and Technology, Ishikawa,

**Notes**

Japan
[3]Intelligent Spoken Language Technology (Tianjin) Co., Ltd., Tianjin, China
[4]AI Labs, Didi Chuxing, Beijing, China

**Tue-P-2-2-11, Time: 14:30-16:30**

Previous studies of speech emotion recognition utilize convolutional neural network (CNN) directly on amplitude spectrogram to extract features. CNN combines with bidirectional long short term memory (BLSTM) has become the state-of-the-art model. However, phase information has been ignored in this model. The importance of phase information in speech processing field is gathering attention. In this paper, we propose feature extraction of amplitude spectrogram and phase information using CNN for speech emotion recognition. The modified group delay cepstral coefficient (MGDCC) and relative phase are used as phase information. Firstly, we analyze the influence of phase information on speech emotion recognition. Then we design a CNN-based feature representation using amplitude and phase information. Finally, experiments were conducted on EmoDB to validate the effectiveness of phase information. Integrating amplitude spectrogram with phase information, the relative emotion error recognition rates are reduced by over 33% in comparison with using only amplitude-based feature.

### Bubble Cooperative Networks for Identifying Important Speech Cues

*Viet Anh Trinh[1], Brian McFee[2,3] and Michael I Mandel[1,4]*
[1]The Graduate Center, CUNY, New York, USA
[2]Center for Data Science, New York University, USA
[3]Music and Audio Research Laboratory, New York University, USA
[4]Brooklyn College, CUNY, New York, USA

**Tue-P-2-2-12, Time: 14:30-16:30**

Predicting the intelligibility of noisy recordings is difficult and most current algorithms treat all speech energy as equally important to intelligibility. Our previous work on human perception used a listening test paradigm and correlational analysis to show that some energy is more important to intelligibility than other energy. In this paper, we propose a system called the Bubble Cooperative Network (BCN), which aims to predict important areas of individual utterances directly from clean speech. Given such a prediction, noise is added to the utterance in unimportant regions and then presented to a recognizer. The BCN is trained with a loss that encourages it to add as much noise as possible while preserving recognition performance, encouraging it to identify important regions precisely and place the noise everywhere else. Empirical evaluation shows that the BCN can obscure 97.7% of the spectrogram with noise while maintaining recognition accuracy for a simple speech recognizer that compares a noisy test utterance with a clean reference utterance. The masks predicted by a single BCN on several utterances show patterns that are similar to analyses derived from human listening tests that analyze each utterance separately, while exhibiting better generalization and less context-dependence than previous approaches.

## Tue-P-2-3: Applications in Education and Learning

Hall 4-6: Poster-3; 14:30-16:30; Tuesday, 4 September, 2018
Chair: Helmer Strik

### Real-Time Scoring of an Oral Reading Assessment on Mobile Devices

*Jian Cheng*
Analytic Measures Inc., Palo Alto, California, U.S.A.

**Tue-P-2-3-1, Time: 14:30-16:30**

We discuss the real-time scoring logic for a self-administered oral reading assessment on mobile devices (Moby.Read) to measure the three components of children's oral reading fluency skills: words correct per minute, expression and comprehension. Critical techniques that make the assessment real-time on-device are discussed in detail. We propose the idea of producing comprehension scores by measuring the semantic similarity between the prompt and the retelling response utilizing the recent advance of document embeddings in natural language processing. By combining features derived from word embedding with the normalized number of common types, we achieved a human-machine correlation coefficient of 0.90 at the participant level for comprehension scores, which was better than the human inter-rater correlation 0.88. We achieved a better human-machine correlation coefficient than that of the human inter-rater in expression scores too. Experimental results demonstrate that Moby.Read can provide highly accurate words correct per minute, expression and comprehension scores in real-time and validate the use of machine scoring methods to automatically measure oral reading fluency skills.

### A Deep Learning Approach to Assessing Non-native Pronunciation of English Using Phone Distances

*Konstantinos Kyriakopoulos, Kate Knill and Mark Gales*
ALTA Institute / Engineering Department
Cambridge University, Trumpington St, Cambridge CB2 1PZ, UK

**Tue-P-2-3-2, Time: 14:30-16:30**

The way a non-native speaker pronounces the phones of a language is an important predictor of their proficiency. In grading spontaneous speech, the pairwise distances between generative statistical models trained on each phone have been shown to be powerful features. This paper presents a deep learning alternative to model-based phone distances in the form of a tunable Siamese network feature extractor to extract distance metrics directly from the audio frame sequence. Features are extracted at the phone instance level and combined to phone-level representations using an attention mechanism. Pair-wise distances between phone features are then projected through a feed-forward layer to predict score. The extraction stage is initialised on either a binary phone instance-pair classification task, or to mimic the model-based features, then the whole system is fine-tuned end-to-end, optimising the learning of the distance metric to the score prediction task. This method is therefore more adaptable and more sensitive to phone instance level phenomena. Its performance is compared against a DNN trained on Gaussian phone model distance features.

### Paired Phone-Posteriors Approach to ESL Pronunciation Quality Assessment

*Yujia Xiao[1,2], Frank Soong[2] and Wenping Hu[2]*
[1]South China University of Technology, Guangzhou, China
[2]Microsoft Research Asia, Beijing, China

**Tue-P-2-3-3, Time: 14:30-16:30**

This work proposes to incorporate paired phone-posteriors as input features into a neural net (NN) model for assessing ESL learner's pronunciation quality. In this work, posteriors of forty phones, instead of several thousand sub-phonemic senones, are used to circumvent the sparsity issues in NN training. Phone posteriors are assembled with their corresponding senone posteriors estimated via a speaker-independent, DNN-based acoustic model, trained with standard American English speech data (i.e., Wall Street Journal database). Phone posteriors of both reference(standard American English speaker) and test speaker are paired together as augmented input feature vectors to train an NN based, 2-class, i.e., native vs nonnative speaker, classifier. The Goodness of Pronunciation (GOP), a proven effective measure, is used as the baseline for comparison. The binary NN classifier trained with such features achieves a high classification accuracy of 89.6% on native and non-native speakers' data. The classifier also shows a better equal error rate (EER) than the GOP-based baseline classifier in either phone or word level pronunciation, i.e., at phone level from 18.3% to 6.2% and at word level from 12.98% to 2.54%.

### Investigating the Role of L1 in Automatic Pronunciation Evaluation of L2 Speech

*Ming Tu[1], Anna Grabek[2], Julie Liss[1] and Visar Berisha[1,2]*

**Notes**

[1]Speech and Hearing Science Department, Arizona State University
[2]School of Electrical, Computer and Energy Engineering, Arizona State University

**Tue-P-2-3-4, Time: 14:30-16:30**

Automatic pronunciation evaluation plays an important role in pronunciation training and second language education. This field draws heavily on concepts from automatic speech recognition (ASR) to quantify how close the pronunciation of non-native speech is to native-like pronunciation. However, it is known that the formation of accent is related to pronunciation patterns of both the target language (L2) and the speaker's first language (L1). In this paper, we propose to use two native speech acoustic models, one trained on L2 speech and the other trained on L1 speech. We develop two sets of measurements that can be extracted from two acoustic models given accented speech. A new utterance-level feature extraction scheme is used to convert these measurements into a fixed-dimension vector which is used as an input to a statistical model to predict the accentedness of a speaker. On a data set consisting of speakers from 4 different L1 backgrounds, we show that the proposed system yields improved correlation with human evaluators compared to systems only using the L2 acoustic model.

## Impact of ASR Performance on Free Speaking Language Assessment

*Kate Knill[1], Mark Gales[1], Konstantinos Kyriakopoulos[1], Andrey Malinin[1], Anton Ragni[1], Yu Wang[1] and Andrew Caines[2]*
[1]ALTA Institute / Engineering Department, Cambridge University, UK
[2]ALTA Institute / Computer Lab, Cambridge University, UK

**Tue-P-2-3-5, Time: 14:30-16:30**

In free speaking tests candidates respond in spontaneous speech to prompts. This form of test allows the spoken language proficiency of a non-native speaker of English to be assessed more fully than read aloud tests. As the candidate's responses are unscripted, transcription by automatic speech recognition (ASR) is essential for automated assessment. ASR will never be 100% accurate so any assessment system must seek to minimise and mitigate ASR errors. This paper considers the impact of ASR errors on the performance of free speaking test auto-marking systems. Firstly rich linguistically related features, based on part-of-speech tags from statistical parse trees, are investigated for assessment. Then, the impact of ASR errors on how well the system can detect whether a learner's answer is relevant to the question asked is evaluated. Finally, the impact that these errors may have on the ability of the system to provide detailed feedback to the learner is analysed. In particular, pronunciation and grammatical errors are considered as these are important in helping a learner to make progress. As feedback resulting from an ASR error would be highly confusing, an approach to mitigate this problem using confidence scores is also analysed.

## Automatic Miscue Detection Using RNN Based Models with Data Augmentation

*Yoon Seok Hong, Kyung Seo Ki and Gahgene Gweon*
Graduate School of Convergence Science and Technology, Seoul National University, South Korea

**Tue-P-2-3-6, Time: 14:30-16:30**

This study proposes a method of using data augmentation to address the problem of data shortages in miscue detection tasks. Three main steps were taken. First, a phoneme classifier was developed to acquire force-aligned data, which would be used for miscue classification and data augmentation. In order to create the phoneme classifier, phonetic features of "Seoul Reading Speech" (SRS) corpus were extracted by using grapheme-to-phoneme (G2P) to train CNN-based models. Second, to obtain miscue labeled corpus, we performed data augmentation using the phoneme classifier output, which is artificially generated miscue corpus of SRS (modified-SRS). This miscue corpus was created by randomly deleting or modifying sound sections according to three miscue categories; extension (EXT), pause (PAU) and pre-correction (PRE). Third, the performance of the miscue classifier was tested after training three types of RNN based models (LSTM, BiLSTM, BiGRU) with the modified-SRS corpus. The results show that the BiGRU model performed best at 0.819 in F1-score on augmented data, while BiLSTM model performed best at 0.512 on real data.

## A Study of Objective Measurement of Comprehensibility through Native Speakers' Shadowing of Learners' Utterances

*Yusuke Inoue[1], Suguru Kabashima[1], Daisuke Saito[1], Nobuaki Minematsu[1], Kumi Kanamura[2] and Yutaka Yamauchi[3]*
[1]The University of Tokyo
[2]Nagoya University of Economics
[3]Tokyo International University

**Tue-P-2-3-7, Time: 14:30-16:30**

While learners desire to acquire so comprehensible pronunciations as to make themselves understood smoothly, acquisition often becomes difficult because, outside of classrooms, it is not rare that learners can hardly find chances to talk in the target language. Even when they talk to native speakers, they may receive only lenient or superficial suggestions from native speakers. How can learners know native speakers' honest perception on their utterances? In this paper, shadowing is introduced not to learners but to native listeners, who are asked to shadow learners' utterances. Since shadowing is as simultaneous repetition as possible, it is expected that native listeners' perceived comprehensibility can be measured objectively as smoothness of natives' shadowings. Experiments show that 1) shadowers' subjective assessment of learners' speech and that of their shadowings are highly correlated and that 2) the former is more correlated with the GOP scores of natives' shadowings than that of learners' speech. These results indicate that it is valid to regard comprehensible pronunciation as shadowable pronunciation.

## Factorized Deep Neural Network Adaptation for Automatic Scoring of L2 Speech in English Speaking Tests

*Dean Luo[1], Chunxiao Zhang[1], Linzhong Xia[1] and Lixin Wang[2]*
[1]School of Electronic Communication Technology, Shenzhen Institute of Information Technology, China,
[2]Shenzhen Seaskyland Technologies, China

**Tue-P-2-3-8, Time: 14:30-16:30**

Speaker adaptation has been shown to be effective on speech recognition and evaluation of L2 speech. However, other factors, such as environments and foreign accents, can affect the speech signal in addition to speakers. Factorizing the speaker, environment and other acoustic factors is crucial in evaluating L2 speech to effectively reduce acoustic mismatch between train and test conditions. In this study, we investigate the effects of deep neural network factorized adaptation techniques on L2 speech assessment in real speaking tests. Through recognition and automatic scoring experiments on L2 speech, we demonstrate that factorized fMLLR and iVector based DNN adaptation can better utilize adaptation data to efficiently adapt to complex speaker and environment conditions. Combining the factored components of iVectors and fMLLR transforms can further improve robustness of DNN models in speech recognition and automatic scoring of L2 speech in dynamic environments.

## On the Difficulties of Automatic Speech Recognition for Kindergarten-Aged Children

*Gary Yeung and Abeer Alwan*
Dept. of Electrical and Computer Engineering, University of California, Los Angeles, USA

**Tue-P-2-3-9, Time: 14:30-16:30**

Automatic speech recognition (ASR) systems for children have lagged behind in performance when compared to adult ASR. The exact problems and evaluation methods for child ASR have not yet been fully investigated. Recent work from the robotics community suggests that ASR for kindergarten speech is especially difficult, even though this age group may benefit most from voice-based educational and diagnostic tools. Our study focused on ASR performance for specific grade levels (K-10) using a word identification task. Grade-specific ASR systems were evaluated, with particular attention

**Notes**

placed on the evaluation of kindergarten-aged children (5-6 years old). Experiments included investigation of grade-specific interactions with triphone models using feature space maximum likelihood linear regression (fMLLR), vocal tract length normalization (VTLN) and subglottal resonance (SGR) normalization. Our results indicate that kindergarten ASR performs dramatically worse than even 1st grade ASR, likely due to large speech variability at that age. As such, ASR systems may require targeted evaluations on kindergarten speech rather than being evaluated under the guise of "child ASR." Additionally, results show that systems trained in matched conditions on kindergarten speech may be less suitable than mismatched-grade training with 1st grade speech. Finally, we analyzed the phonetic errors made by the kindergarten ASR.

### Improved Acoustic Modelling for Automatic Literacy Assessment of Children

*Mauro Nicolao[1], Michiel Sanders[2] and Thomas Hain[1]*

[1]Speech and Hearing Research Group, The University of Sheffield, UK

[2]ITSLanguage BV, The Netherlands

**Tue-P-2-3-10, Time: 14:30-16:30**

Automatic literacy assessment of children is a complex task that normally requires carefully annotated data. This paper focuses on a system for the assessment of reading skills, aiming to detection of a range of fluency and pronunciation errors. Naturally, reading is a prompted task and thereby the acquisition of training data for acoustic modelling should be straightforward. However, given the prominence of errors in the training set and the importance of labelling them in the transcription, a lightly supervised approach to acoustic modelling has better chances of success. A method based on weighted finite state transducers is proposed, to model specific prompt corrections, such as repetitions, substitutions and deletions, as observed in real recordings. Iterative cycles of lightly-supervised training are performed in which decoding improves the transcriptions and the derived models. Improvements are due to increasing accuracy in phone-to-sound alignment and in the training data selection. The effectiveness of the proposed methods for relabelling and acoustic modelling is assessed through experiemnts on the CHOREC corpus, in terms of sequence error rate and alignment accuracy. Improvements over the baseline of up to 60% and 23.3% respectively are observed.

## Tue-SS-2-2: Special Session: Integrating Speech Science and Technology for Clinical Applications

Hall 4-6: Poster-4; 14:30-16:30; Tuesday, 4 September, 2018

Chairs: Christina Hagedorn, Shrikanth Narayanan and Uttam Sinha

### Anomaly Detection Approach for Pronunciation Verification of Disordered Speech Using Speech Attribute Features

*Mostafa Shahin[1], Beena Ahmed[1], Jim X. Ji[1] and Kirrie Ballard[2]*

[1]Dept. of Electrical and Computer Engineering, Texas A&M University, Doha, Qatar

[2]Faculty of Health Sciences, The University of Sydney, Sydney, Australia

**Tue-SS-2-2-1, Time: 14:30-16:30**

The automatic assessment of speech is a powerful tool in computer aided speech therapy for disorders such as Childhood Apraxia of Speech (CAS). However, the lack of sufficient annotated disordered speech data seriously impedes the accurate detection of pronunciation errors. To handle this deficiency, in this paper, we used the novel approach of tackling pronunciation verification as an anomaly detection problem. We achieved this by modeling only the correct pronunciation of each individual phoneme with a one-class Support Vector Machine (SVM) trained using a set of speech attributes features, namely the manner and place of articulation. These features are extracted from a bank of pre-trained Deep Neural Network (DNN) speech attributes classifiers. The one-class SVM model classifies each phoneme production as normal (correct) or an anomaly (incorrect). We evaluated the system using both native speech with artificial errors and disordered speech collected from children with apraxia of speech and compared it with the DNN Goodness of Pronunciation (GOP) algorithm. The results show that our approach reduces the false-rejection rates by around 35% when applied to disordered speech.

### Effectiveness of Voice Quality Features in Detecting Depression

*Amber Afshan[1], Jinxi Guo[1], Soo Jin Park[1], Vijay Ravi[1], Jonathan Flint[2] and Abeer Alwan[1]*

[1]Dept. of Electrical Engineering, University of California Los Angeles, USA

[2]Dept. of Psychiatry and Biobehavioral Sciences, UCLA School of Medicine, Los Angeles, USA

**Tue-SS-2-2-2, Time: 14:30-16:30**

Automatic assessment of depression from speech signals is affected by variabilities in acoustic content and speakers. In this study, we focused on addressing these variabilities. We used a database comprised of recordings of interviews from a large number of female speakers: 735 individuals suffering from depressive (dysthymia and major depression) and anxiety disorders (generalized anxiety disorder, panic disorder with or without agoraphobia) and 953 healthy individuals. Leveraging this unique and extensive database, we built an i-vector framework. In order to capture various aspects of speech signals, we used voice quality features in addition to conventional cepstral features. The features (F0, F1, F2, F3, H1-H2, H2-H4, H4-H2k, A1, A2, A3 and CPP) were inspired by a psychoacoustic model of voice quality [1]. An i-vector-based system using Mel Frequency Cepstral Coefficients (MFCCs) and another using voice quality features was developed. Voice quality features performed as well as MFCCs. A score-level fusion was then used to combine these two systems, resulting in a 6% relative improvement in accuracy in comparison with the i-vector system based on MFCCs alone. The system was robust even when the duration of the utterances was shortened to 10 seconds.

### Fusing Text-dependent Word-level i-Vector Models to Screen 'at Risk' Child Speech

*Prasanna Kothalkar[1], Johanna Rudolph[2], Christine Dollaghan[2], Jennifer McGlothlin[2], Thomas Campbell[2] and John H.L. Hansen[1]*

[1]Center for Robust Speech Systems, University of Texas at Dallas

[2]Callier Center for Communication Disorders, University of Texas at Dallas.

**Tue-SS-2-2-3, Time: 14:30-16:30**

Speech sound disorders (SSDs) are the most prevalent type of communication disorder among preschoolers. The earlier an SSD is identified, the earlier an intervention can be provided to potentially reduce the social/academic impact of the disorder. The challenge, lies in early identification of such disorders. In this study 29 carefully selected words were produced by 165 children from 3-6 years of age. The audio recordings, were collected by parents using a mobile application /platform. "Ground truth" child status as 'typically developing' vs 'at risk' was based on a percentage of consonants correct-revised growth curve model. State-of-the-art speech processing/speaker recognition models were employed along with our clinical group verification framework. Results showed that text-dependent i-Vector models were superior to both text dependent and text-independent Gaussian Mixture Models (GMMs) for correct classification of children. Fusing individual word, i-Vector models provides insight into word and consonant groupings that are more indicative of 'at risk' child speech.

**Notes**

## Testing Paradigms for Assistive Hearing Devices in Diverse Acoustic Environments

*Ram Charan Chandra Shekar, Hussnain Ali and John H.L. Hansen*

Cochlear Implant Laboratory – Center for Robust Speech Systems, The University of Texas at Dallas, Richardson TX, USA

**Tue-SS-2-2-4, Time: 14:30-16:30**

Many individuals worldwide are at risk of hearing loss due to unsafe acoustical exposure and chronic listening experience using personal audio devices. Assistive hearing devices(AHD), such as hearing-aids(HAs) and cochlear-implants(CIs) are a common choice for the restoration and rehabilitation of the auditory function. Audio sound processors in CIs and HAs operate within limits, prescribed by audiologists, not only for acceptable sound perception but also for safety reasons. Signal processing(SP) engineers follow best design practices to ensure reliable performance and incorporate necessary safety checks within the design of SP strategies to ensure safety limits are never exceeded irrespective of acoustic environments. This paper proposes a comprehensive testing and evaluation paradigm to investigate the behavior of audio devices that addresses the safety concerns in diverse acoustic conditions. This is achieved by characterizing the performance of devices with large amounts of acoustic inputs and monitoring the output behavior. The CCi-MOBILE Research-Interface(RI) (used for CI/HA research) is used in this study as the testing paradigm. Factors such as pulse-width(PW), inter-phase gap(IPG) and a number of other parameters are estimated to evaluate the impact of AHDs on hearing comfort, subjective sound quality and characterize audio devices in terms of listening perception and biological safety.

## Detection of Dementia from Responses to Atypical Questions Asked by Embodied Conversational Agents

*Tsuyoki Ujiro[1], Hiroki Tanaka[1], Hiroyoshi Adachi[2], Hiroaki Kazui[3], Manabu Ikeda[4], Takashi Kudo[2] and Satoshi Nakamura[1]*

[1]Graduate School of Information Science, Nara Institute of Science and Technology, Japan

[2]Health and Counseling Center, Osaka University, Japan

[3]Department of Neuropsychiatry, Kochi Medical School, Japan

[4]Graduate School of Medicine, Osaka University, Japan

**Tue-SS-2-2-5, Time: 14:30-16:30**

Detection of dementia requires examinations, such as blood tests and functional magnetic resonance imaging (fMRI), that can be very stressful for the patient. Previous studies proposed screenings for easy detection of dementia that utilized acoustic and language information derived from conversations between patients and medical staff. Although these studies demonstrated effectiveness in automatically detecting dementia, the tasks used were created based on neuropsychological tests. The effect of habituation on this limited variety of tasks might have a negative impact on routine dementia screening. We propose a method to detect dementia using responses to more atypical questions asked by embodied conversational agents. Through consultations with neuropsychologists, we created a total of 13 questions. The embodied conversational agent obtained answers to these questions from 24 participants (12 dementia and 12 non-dementia). We recorded their responses and extracted speech and language features. We classified the two groups (dementia/non-dementia) by a machine learning algorithm (support vector machines and logistic regression) using the extracted features. The results showed a 0.95 detection performance in the area under the curve of the receiver operating characteristic (AUROC). This result demonstrates that our system using atypical questions can detect dementia.

## Acoustic Features Associated with Sustained Vowel and Continuous Speech Productions by Chinese Children with Functional Articulation Disorders

*Wang Zhang[1,2], Xiangquan Gui[1], Tianqi Wang[2], Manwa Ng[3], Feng Yang[4], Lan Wang[2] and Nan Yan[2,5]*

[1]School of Computer and Communication, Lanzhou University Of Technology, Lanzhou, China

[2]CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

[3]Speech Science Laboratory, University of Hong Kong, Hong Kong, China

[4]Department of Speech Therapy, Shenzhen Children's Hospital, Shenzhen, China

[5]Guangdong Provincial Key Laboratory of Robotics and Intelligent System, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

**Tue-SS-2-2-6, Time: 14:30-16:30**

Functional articulation disorder (FAD) is a speech disorder commonly found in preschoolers, negatively affecting their day-to-day communication and in the long run their psychological development. Current FAD research mainly focused on the perceptual aspects, but not other means such as acoustic and physiological analyses. The present study aimed to evaluate the different acoustic features associated with sustained vowels and continuous speech produced by children with FAD and their age-matched controls. Speech samples produced by 67 children with FAD and 30 typically developing children. Articulatory-acoustic vowel space features, including formant centralization ratio (FCR3), F1 range ratio (F1RR), F2 range ratio (F2RR) and triangular vowel space area (TVSA), were calculated using the first two formant frequencies from vowels /a/, /i/, /u/. Voice onset time (VOT) values associated with the stop consonants were also obtained. Results indicated that children with FAD exhibited articulatory undershooting with reduced range of articulatory movements, as well as poorer control over the release of oral occlusion when producing aspirated or unaspirated stops, when compared with normal counterparts. The findings support the notion that these acoustic features can be used to differentiate misarticulated speech from healthy speech and could be used to objectively classify and evaluate FAD speech.

## Estimation of Hypernasality Scores from Cleft Lip and Palate Speech

*Vikram C M[1], Ayush Tripathi[2], Sishir Kalita[1] and S R Mahadeva Prasanna[1,3]*

[1]Indian Institute of Technology Guwahati, Guwahati, India

[2]Visvesvaraya National Institute of Technology, Nagpur, India

[3]Indian Institute of Technology Dharwad, Dharwad, India

**Tue-SS-2-2-7, Time: 14:30-16:30**

Hypernasality refers to the perception of excessive nasal resonances in vowels and voiced consonants. Existing speech processing based approaches concentrate only on the classification of speech into normal or hypernasal, which do not give the degree of hypernasality in terms of continuous values like nasometer. Motivated by the functionality of nasometer, in this work, a method is proposed for the evaluation of hypernasality. Speech signals representing two extremely opposite cases of nasality are used to develop the acoustic models, where oral sentences (rich in vowels, stops and fricatives) of normal speakers and nasal sentences (rich in nasals and nasalized vowels) of moderate-severe hypernasal speakers represent the groups with minimum and maximum attainable degrees of nasality, respectively. The acoustic features derived from glottal activity regions are used to model the maximum and minimum nasality classes using Gaussian mixture model and deep neural network approaches. The posterior probabilities obtained for nasal sentence class are referred to as hypernasality scores. The scores show a significant correlation ($p<0.01$) with respect to perceptual ratings of hypernasality, provided by expert speech-language pathologists. Further, hypernasality scores are used for the detection of hypernasality and the results are compared with the nasometer based approach.

## Detecting Alzheimer's Disease Using Gated Convolutional Neural Network from Audio Data

*Tifani Warnita, Nakamasa Inoue and Koichi Shinoda*

Tokyo Institute of Technology, Tokyo, Japan

**Tue-SS-2-2-8, Time: 14:30-16:30**

We propose an automatic detection method of Alzheimer's diseases using a gated convolutional neural network (GCNN) from speech data. This GCNN

**Notes**

can be trained with a relatively small amount of data and can capture the temporal information in audio paralinguistic features. Since it does not utilize any linguistic features, it can be easily applied to any languages. We evaluated our method using Pitt Corpus. The proposed method achieved the accuracy of 73.6%, which is better than the conventional sequential minimal optimization (SMO) by 7.6 points.

### Automatic Detection of Orofacial Impairment in Stroke

*Andrea Bandini[1], Jordan Green[2], Brian Richburg[2] and Yana Yunusova[1,3,4]*

[1]University Health Network: Toronto Rehabilitation Institute, Toronto, Canada

[2]MGH Institute of Health Professions, Boston, USA

[3]Brain Sciences, Sunnybrook Research Institute, Toronto, Canada

[4]Department of Speech-Language Pathology, University of Toronto, Canada

**Tue-SS-2-2-9, Time: 14:30-16:30**

Stroke is a devastating condition that affects the ability of people to communicate through speech, leading to social isolation and poor quality of life. The quantitative evaluation of speech and orofacial movements is essential for assessing the impairment and identifying treatment targets. However, to our knowledge, a tool for the automatic orofacial assessment, which considers multiple aspects of orofacial impairment (e.g., range of motion in addition to asymmetry), has not been developed for this clinical population. In this work, we tested a video-based approach for the automatic orofacial assessment in stroke survivors, combining low-cost depth sensor and face alignment algorithms for extracting facial features. Twelve patients post-stroke and 11 control subjects were evaluated during speech and non-speech tasks. By using a small feature-set representing range of motion and asymmetry of face movements, it was possible to discriminate patients post-stroke from control subjects with high accuracy (87%). Further insights on the choice of the task and face alignment algorithm are provided, demonstrating that a non-parametric approach such as SDM can provide better results. Through this work we demonstrated the feasibility of an objective tool to support clinicians in the assessment of speech and orofacial impairment post-stroke.

### Detecting Depression with Audio/Text Sequence Modeling of Interviews

*Tuka Al Hanai[1], Mohammad Ghassemi[2] and James Glass[1]*

[1]Computer Science and Artificial Intelligence Laboratory

[2]Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

**Tue-SS-2-2-10, Time: 14:30-16:30**

Medical professionals diagnose depression by interpreting the responses of individuals to a variety of questions, probing lifestyle changes and ongoing thoughts. Like professionals, an effective automated agent must understand that responses to queries have varying prognostic value. In this study we demonstrate an automated depression-detection algorithm that models interviews between an individual and agent and learns from sequences of questions and answers without the need to perform explicit topic modeling of the content. We utilized data of 142 individuals undergoing depression screening and modeled the interactions with audio and text features in a Long-Short Term Memory (LSTM) neural network model to detect depression. Our results were comparable to methods that explicitly modeled the topics of the questions and answers which suggests that depression can be detected through sequential modeling of an interaction, with minimal information on the structure of the interview.

## Tue-P-2-5: Speaker Characterization and Analysis

Hall 4-6: Poster-5; 14:30-16:30; Tuesday, 4 September, 2018

Chair: Chung-Hsien Wu

### Discourse Marker Detection for Hesitation Events on Mandarin Conversation

*Yu-Wun Wang[1], Hen-Hsen Huang[1], Kuan-Yu Chen[2] and Hsin-Hsi Chen[1,3]*

[1]Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan

[2]Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan

[3]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan

**Tue-P-2-5-1, Time: 14:30-16:30**

The occurrence of hesitation events in spontaneous conversations can be associated with the difficulties in memory recall. One indicator of hesitation in speech in Taiwanese Mandarin is the usage of discourse markers. This paper introduces an approach to the detection of discourse markers that denote hesitation events. We propose a sequential labeling model to detect discourse markers in conversations by taking information on both acoustic level and word level into account. Experimental results show the integration of word-level acoustic feature extraction network significantly enhances the detection performance. Our approach for further applications is also discussed.

### Acoustic and Perceptual Characteristics of Mandarin Speech in Homosexual and Heterosexual Male Speakers

*Puyang Geng[1], Wentao Gu[1] and Hiroya Fujisaki[2]*

[1]School of Chinese Language and Culture, Nanjing Normal University, Nanjing, China

[2]The University of Tokyo, Tokyo, Japan

**Tue-P-2-5-2, Time: 14:30-16:30**

The present study investigated both acoustic and perceptual characteristics of Mandarin speech in homosexual and heterosexual male speakers. Acoustic analyses of monosyllabic words showed significant differences between the two groups in F0 features (including the mean, the max and the range), F1 and F2 of vowels, aspiration/frication duration of consonants and center of gravity as well as skewness for /s/. Especially, the patterns were found to be opposite between Mandarin and American English speakers, which might be due to social psychological differences between the two societies. The perceptual experiment showed that the perceived score of gayness differed significantly between the speeches of the two groups. Among those acoustic parameters showing significant differences, fricative duration may be the most salient cue for sexual orientation of Mandarin male speakers.

### Automatic Question Detection from Acoustic and Phonetic Features Using Feature-wise Pre-training

*Atsushi Ando[1], Reine Asakawa[1,2], Ryo Masumura[1], Hosana Kamiyama[1], Satoshi Kobashikawa[1] and Yushi Aono[1]*

[1]NTT Media Intelligence Laboratories, NTT Corporation, Japan

[2]Toyohashi University of Technology, Japan

**Tue-P-2-5-3, Time: 14:30-16:30**

This paper presents a novel question detection method from natural speech using acoustic and phonetic features. The conventional methods based on Recurrent Neural Networks (RNNs) use only acoustic features.

**Notes**

However, lexical cues are essential to identify some questions such as declarative questions. To this end we propose a new RNN-based question detection model which utilizes both acoustic and lexical information. Phonetic features which are suitable to describe interrogative cues are used as lexical information. Furthermore, we also propose a new training framework named feature-wise pre-training (FP) to combine the acoustic and phonetic features effectively. FP attempts to acquire interrogative cues in individual features instead of the combination of the features, which makes the model training more stable. The estimation models of the interrogatives are then integrated and fine-tuning is applied to obtain the unified comprehensive model. Experiments show that the proposed method offers better performance than the conventional benchmarks.

## Improving Response Time of Active Speaker Detection Using Visual Prosody Information Prior to Articulation

*Fasih Haider[1,2], Saturnino Luz[1], Carl Vogel[2] and Nick Campbell[2]*

[1]IPHSI, University of Edinburgh, UK

[2]ADAPT Centre, Trinity College Dublin, Ireland

**Tue-P-2-5-4, Time: 14:30-16:30**

Natural multi-party interaction commonly involves turning one's gaze towards the speaker who has the floor. Implementing virtual agents or robots who are able to engage in natural conversations with humans therefore requires enabling machines to exhibit this form of communicative behaviour. This task is called active speaker detection. In this paper, we propose a method for active speaker detection using visual prosody (lip and head movements) information before and after speech articulation to decrease the machine response time; and also demonstrate the discriminating power of visual prosody before and after speech articulation for active speaker detection. The results show that the visual prosody information one second before articulation is helpful in detecting the active speaker. Lip movements provide better results than head movements and fusion of both improves accuracy. We have also used visual prosody information of the first second of the speech utterance and found that it provides more accurate results than one second before articulation. We conclude that the fusion of lip movements from both regions (the first one second of speech and the one second before articulation) improves the accuracy of active speaker detection.

## Audio-Visual Prediction of Head-Nod and Turn-Taking Events in Dyadic Interactions

*Bekir Berker Türker, Engin Erzin, Yücel Yemez and Metin Sezgin*

Koç University, Turkey

**Tue-P-2-5-5, Time: 14:30-16:30**

Head-nods and turn-taking both significantly contribute conversational dynamics in dyadic interactions. Timely prediction and use of these events is quite valuable for dialog management systems in human-robot interaction. In this study, we present an audio-visual prediction framework for the head-nod and turn-taking events that can also be utilized in real-time systems. Prediction systems based on Support Vector Machines (SVM) and Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) are trained on human-human conversational data. Unimodal and multimodal classification performances of head-nod and turn-taking events are reported over the IEMOCAP dataset.

## Analyzing Effect of Physical Expression on English Proficiency for Multimodal Computer-Assisted Language Learning

*Haoran Wu, Yuya Chiba, Takashi Nose and Akinori Ito*

Graduate School of Engineering, Tohoku University, Japan

**Tue-P-2-5-6, Time: 14:30-16:30**

English proficiency is important for communication in English. Computer-Assisted Language Learning (CALL) systems are introduced to provide a convenient and low-cost language learning environment. Most of the conventional speech-based CALL systems concentrate on developing verbal fluency of the learners. However, actual English communication involves not only verbal expressions but also facial expressions and gestures, which could affect the perceived proficiency. The objective of our research is to develop a CALL system that can evaluate fluency of physical expressions as well as the verbal fluency of English.

However, it is not clear how physical expressions affect the overall proficiency of English. Therefore, this study investigates the relationship between the proficiency of English and the fluency of the physical expression by analyzing the dialog data of the multimodal CALL system.

## Analysis of the Effect of Speech-Laugh on Speaker Recognition System

*Sri Harsha Dumpala, Ashish Panda and Sunil Kumar Kopparapu*

TCS Research and Innovation Labs

**Tue-P-2-5-7, Time: 14:30-16:30**

A robust speaker recognition system should be able to recognize a speaker despite all the possible variations in speaker's speech. A common variation of the neutral speech is speech-laugh, which occurs when a person is speaking and laughing, simultaneously. In this paper, we show that speech-laugh significantly degrades the performance of an i-vector based speaker recognition system. Further, we show that laughter and neutral speech contain complementary speaker information, which can be combined to improve the performance of the speaker recognition system for speech-laugh scenarios. Using AMI meeting corpus database, we show that by including neutral speech and laughter in enrollment phase, the performance of the system in the speech-laugh scenarios can be relatively improved by 36% in EER.

## Vocal Biomarkers for Cognitive Performance Estimation in a Working Memory Task

*Jennifer Sloboda[1], Adam Lammert[1], James Williamson[1], Christopher Smalt[1], Daryush D. Mehta[1,2], COL Ian Curry[3], Kristin Heaton[4], Jeffrey Palmer[1] and Thomas Quatieri[1]*

[1]MIT Lincoln Laboratory, Lexington, MA, USA

[2]Massachusetts General Hospital, Center for Laryngeal Surgery and Voice Rehabilitation

[3]US Army Aeromedical Research Laboratory

[4]US Army Research Institute of Environmental Medicine

**Tue-P-2-5-8, Time: 14:30-16:30**

The ability to non-invasively estimate cognitive fatigue and workload as contributing factors to cognitive performance has value for planning and decision making surrounding human participation in cognitively demanding situations and environments. Growing evidence supports the use of speech as an effective modality for assessing cognitive fatigue and workload, while also being operationally appropriate in a wide variety of environments. To assess ability to discriminate changes in cognitive fatigue and load from speech, features that measure speech onset time, speaking rate, voice quality and vocal tract coordination from the delta-mel-cepstrum are evaluated on two independent data sets that employ the same auditory working memory task. Feature effect sizes due to fatigue were generally larger than those due to load. Speech onset time, speaking rate and vocal tract coordination features show strong potential for speech-based fatigue estimation.

## Lexical and Acoustic Deep Learning Model for Personality Recognition

*Guozhen An[1,2] and Rivka Levitan[1,3]*

[1]Department of Computer Science, CUNY Graduate Center, USA

[2]Department of Mathematics and Computer Science, York College (CUNY), USA

[3]Department of Computer and Information Science, Brooklyn College (CUNY), USA

**Tue-P-2-5-9, Time: 14:30-16:30**

Deep learning has been very successful on labeling tasks such as image classification and neural network modeling, but there has not yet been much work on using deep learning for automatic personality recognition. In this study, we propose two deep learning structures for the task of personality recognition using acoustic-prosodic, psycholinguistic and lexical features and present empirical results of several experimental configurations, including

**Notes**

a cross-corpus condition to evaluate robustness. Our best models match or outperform state-of-the-art on the well-known myPersonality corpus and also set a new state-of-the-art performance on the more difficult CXD corpus.

## Perspective Talk-2

Hall 3; 17:00-17:30; Tuesday, 4 September, 2018

Chair: Shrikanth Narayanan

### Open Problems in Speech Recognition

*Bhuvana Ramabhadran*

Google, USA

**Tue-Perspective-2, Time: 17:00-17:30**

In this talk, I will focus on the evolution of ideas in speech recognition over the last couple of decades, with emphasis on the key breakthroughs over the last ten years, its impact across spoken language processing in several languages, recent trends and open challenges that remain to be addressed. One such breakthrough is the use of several neural network model variants, which has had an enormous impact on the performance of state-of-the-art large vocabulary speech recognition systems. They have also had impact on keyword search which is the task of localizing an orthographic query in a speech corpus, and is typically performed through analysis of automatic speech recognition (ASR). Using the recently concluded IARPA funded Babel program as an example of a well-benchmarked task that focussed on the rapid development of speech recognition capability for keyword search in a previously unstudied language, I will present the successes and challenges that persist with limited amounts of transcription. Interpreting and understanding the hidden representations of various models remains a challenge today. I will also discuss current research taking advantage of such interpretations to improve robustness to noisy environments, speaker/domain adaptation algorithms, and dialects/accents. I will conclude with relevant metrics to measure speech recognition performance today that include and ignore the bigger picture of end to end user experience.

## Plenary Talk-2

Hall 3; 8:30-9:30; Wednesday, 5 September, 2018

Chair: Paavo Alku

### Evolution of Neural Network Architectures for Speech Recognition

*Hervé Bourlard*

Idiap Research Institute and EPFL, Switzerland

**Wed-Plenary-2, Time: 08:30-09:30**

Over these last few years, the use of Artificial Neural Networks (ANNs), now often referred to as deep learning or Deep Neural Networks (DNNs), has significantly reshaped research and development in a variety of signal and information processing tasks. While further boosting the state-of-the-art in Automatic Speech Recognition (ASR), recent progresses in the field have also allowed for more flexible and faster developments in emerging markets and multilingual societies (e.g., under-resourced languages).

In this talk, we will provide a historical account of ANN architectures used for ASR since the mid-1980's, and now used in most ASR and spoken language understanding applications. We will start by recalling/revisiting key links between ANNs and statistical inference, discriminant analysis, and linear/nonlinear algebra. Finally, we will briefly discuss more recent trends towards novel DNN-based ASR approaches, including complex hierarchical systems, sparse recovery modeling, and "end-to-end systems."

However, and in spite of the recent progress in the area, we still lack basic understanding of the problems in hands. Although more and more tools are now available, in association with basically "unlimited" processing and data resources, we still fail in building principled ASR models and theories. Alternatively, we are still relying on "ignorance-based" models, often exposing limitations of our understanding, rather than enriching the field of ASR. Discussion of these limitations will underpin all of our overview.

## Wed-O-1-1: Novel Neural Network Architectures for Acoustic Modelling

Hall 3; 10:00-12:00; Wednesday, 5 September, 2018

Chairs: Ralf Schlüter and Tara Sainath

### Layer Trajectory LSTM

*Jinyu Li, Changliang Liu and Yifan Gong*

Microsoft AI and Research

**Wed-O-1-1-1, Time: 10:00-10:20**

It is popular to stack LSTM layers to get better modeling power, especially when large amount of training data is available. However, an LSTM-RNN with too many vanilla LSTM layers is very hard to train and there still exists the gradient vanishing issue if the network goes too deep. This issue can be partially solved by adding skip connections between layers, such as residual LSTM. In this paper, we propose a layer trajectory LSTM (ltLSTM) which builds a layer-LSTM using all the layer outputs from a standard multi-layer time-LSTM. This layer-LSTM scans the outputs from time-LSTMs and uses the summarized layer trajectory information for final senone classification. The forward-propagation of time-LSTM and layer-LSTM can be handled in two separate threads in parallel so that the network computation time is the same as the standard time-LSTM. With a layer-LSTM running through layers, a gated path is provided from the output layer to the bottom layer, alleviating the gradient vanishing issue. Trained with 30 thousand hours of EN-US Microsoft internal data, the proposed ltLSTM performed significantly better than the standard multi-layer LSTM and residual LSTM, with up to 9.0% relative word error rate reduction across different tasks.

### Semi-tied Units for Efficient Gating in LSTM and Highway Networks

*Chao Zhang and Philip Woodland*

Cambridge University Engineering Dept., Trumpington St., Cambridge, CB2 1PZ U.K.

**Wed-O-1-1-2, Time: 10:20-10:40**

Gating is a key technique used for integrating information from multiple sources by long short-term memory (LSTM) models and has recently also been applied to other models such as the highway network. Although gating is powerful, it is rather expensive in terms of both computation and storage as each gating unit uses a separate full weight matrix. This issue can be severe since several gates can be used together in e.g. an LSTM cell. This paper proposes a semi-tied unit (STU) approach to solve this efficiency issue, which uses one shared weight matrix to replace those in all the units in the same layer. The approach is termed "semi-tied" since extra parameters are used to separately scale each of the shared output values. These extra scaling factors are associated with the network activation functions and result in the use of parameterised sigmoid, hyperbolic tangent and rectified linear unit functions. Speech recognition experiments using British English multi-genre broadcast data showed that using STUs can reduce the calculation and storage cost by a factor of three for highway networks and four for LSTMs, while giving similar word error rates to the original models.

### Gaussian Process Neural Networks for Speech Recognition

*Max W. Y. Lam[1], Shoukang Hu[1], Xurong Xie[2], Shansong Liu[1], Jianwei Yu[1], Rongfeng Su[3], Xunying Liu[1] and Helen Meng[1]*

**Notes**

[1]Department of Systems Engineering and Engineering Management, [2]Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China
[3]Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

**Wed-O-1-1-3, Time: 10:40-11:00**

Deep neural networks (DNNs) play an important role in state-of-the-art speech recognition systems. One important issue associated with DNNs and artificial neural networks in general is the selection of suitable model structures, for example, the form of hidden node activation functions to use. Due to lack of automatic model selection techniques, the choice of activation functions has been largely empirically based. In addition, the use of deterministic, fixed-point parameter estimates is prone to over-fitting when given limited training data. In order to model both models' structural and parametric uncertainty, a novel form of DNN architecture using non-parametric activation functions based on Gaussian process (GP), Gaussian process neural networks (GPNN), is proposed in this paper. Initial experiments conducted on the ARPA Resource Management task suggest that the proposed GPNN acoustic models outperformed the baseline sigmoid activation based DNN by 3.40% to 24.25% relatively in terms of word error rate. Consistent performance improvements over the DNN baseline were also obtained by varying the number of hidden nodes and the number of spectral basis functions.

### Acoustic Modeling with Densely Connected Residual Network for Multichannel Speech Recognition

*Jian Tang[1], Yan Song[1], Lirong Dai[1] and Ian McLoughlin[2]*
[1]National Engineering Laboratory for Speech and Language Information Processing University of Science and Technology of China, Hefei, Anhui, P.R.China
[2]School of Computing, University of Kent, Medway, UK

**Wed-O-1-1-4, Time: 11:00-11:20**

Motivated by recent advances in computer vision research, this paper proposes a novel acoustic model called Densely Connected Residual Network (DenseRNet) for multichannel speech recognition. This combines the strength of both DenseNet and ResNet. It adopts the basic "building blocks" of ResNet with different convolutional layers, receptive field sizes and growth rates as basic components that are densely connected to form socalled denseR blocks. By concatenating the feature maps of all preceding layers as inputs, DenseRNet can not only strengthen gradient back-propagation for the vanishing-gradient problem, but also exploit multi-resolution feature maps. Preliminary experimental results on CHiME 3 have shown that DenseRNet achieves a word error rate (WER) of 7.58% on beamforming-enhanced speech with six channel real test data by cross entropy criteria training while WER is 10.23% for the official baseline. Besides, additional experimental results are also presented to demonstrate that DenseRNet exhibits the robustness to beamforming-enhanced speech as well as near and far-field speech.

### Gated Recurrent Unit Based Acoustic Modeling with Future Context

*Jie Li[1], Xiaorui Wang[1], Yuanyuan Zhao[2] and Yan Li[1]*
[1]Kwai, Beijing, P.R. China
[2]Institute of Automation, Chinese Academy of Sciences, Beijing, P.R. China

**Wed-O-1-1-5, Time: 11:20-11:40**

The use of future contextual information is typically shown to be helpful for acoustic modeling. However, for the recurrent neural network (RNN), it's not so easy to model the future temporal context effectively, meanwhile keep lower model latency. In this paper, we attempt to design a RNN acoustic model that being capable of utilizing the future context effectively and directly, with the model latency and computation cost as low as possible. The proposed model is based on the minimal gated recurrent unit (mGRU) with an input projection layer inserted in it. Two context modules, temporal encoding and temporal convolution, are specifically designed for this architecture to model the future context. Experimental results on the Switchboard task and an internal Mandarin ASR task show that, the proposed model performs much better than long short-term memory (LSTM) and mGRU models, whereas enables online decoding with a maximum latency of 170 ms. This model even outperforms a very strong baseline, TDNN-LSTM, with smaller model latency and almost half less parameters.

### Output-Gate Projected Gated Recurrent Unit for Speech Recognition

*Gaofeng Cheng[1,2], Daniel Povey[4,5], Lu Huang[3], Ji Xu[1], Sanjeev Khudanpur[4,5] and Yonghong Yan[1,2,6]*
[1]Key Laboratory of Speech Acoustics and Content Understanding, IOA, CAS, China
[2]University of Chinese Academy of Sciences, China
[3]Tsinghua University, China
[4]Center of Language and Speech Processing, The Johns Hopkins University, Baltimore, MD, USA
[5]Human Language Technology Center Of Excellence, The Johns Hopkins University, Baltimore, MD, USA
[6]Xinjiang Laboratory of Minority Speech and Language Information Processing, Xinjiang Technical Institute of Physics and Chemistry, CAS, China

**Wed-O-1-1-6, Time: 11:40-12:00**

In this paper, we describe the work on accelerating decoding speed while improving the decoding accuracy. Firstly, we propose an architecture which we call Projected Gated Recurrent Unit (PGRU) for automatic speech recognition (ASR) tasks and show that the PGRU could outperform the standard GRU consistently. Secondly, in order to improve the PGRU's generalization, especially for large-scale ASR task, the Output-gate PGRU (OPGRU) is proposed. Finally, time delay neural network (TDNN) and normalization skills are found to be beneficial to the proposed projected-based GRU. The finally proposed unidirectional TDNN-OPGRU acoustic model achieves 3.3% / 4.5% relative reduction in word error rate (WER) compared with bidirectional projected LSTM (BLSTMP) on Eval2000 / RT03 test sets. Meanwhile, TDNN-OPGRU acoustic model speeds up the decoding speed by around 2.6 times compared with BLSTMP.

## Wed-O-1-2: Language Identification

Hall 1; 10:00-12:00; Wednesday, 5 September, 2018
Chairs: Hynek Boril and Margarita Kotti

### Performance Analysis of the 2017 NIST Language Recognition Evaluation

*Seyed Omid Sadjadi[1], Timothee Kheyrkhah[1], Craig Greenberg[1], Elliot Singer[2], Douglas Reynolds[2], Lisa Mason[3] and Jaime Hernandez-Cordero[3]*
[1]NIST/ITL/IAD/Multimodal Information Group, MD, USA
[2]MIT Lincoln Laboratory, Lexington, MA, USA
[3]U.S. Department of Defense, MD, US

**Wed-O-1-2-1, Time: 10:00-10:20**

The 2017 NIST Language Recognition Evaluation (LRE) was held in the autumn of 2017. Similar to past LREs, the basic task in LRE17 was language detection, with an emphasis on discriminating closely related languages (14 in total) selected from 5 language clusters. LRE17 featured several new aspects including: audio data extracted from online videos; a development set for system training and development use; log-likelihood system output submissions; a normalized cross-entropy performance measure as an alternative metric; and, the release of a baseline system developed using the NIST Speaker and Language Recognition Evaluation (SLRE) toolkit for participant use. A total of 18 teams from 25 academic and industrial organizations participated in the evaluation and submitted 79 valid systems

**Notes**

under fixed and open training conditions first introduced in LRE15. In this paper, we report an in-depth analysis of system performance broken down by multiple factors such as data source and gender, as well as a cross-year performance comparison of leading systems from LRE15 and LRE17 to measure progress over the 2-year period. In addition, we present a comparison of primary versus "single best" submissions to understand the effect of fusion on overall performance.

## Using Deep Neural Networks for Identification of Slavic Languages from Acoustic Signal

*Lukas Mateju, Petr Cerva, Jindrich Zdansky and Radek Safarik*

Faculty of Mechatronics, Informatics and Interdisciplinary Studies, Technical University of Liberec, Studentska 2, 461 17 Liberec, Czech Republic

**Wed-O-1-2-2, Time: 10:20-10:40**

This paper investigates the use of deep neural networks (DNNs) for the task of spoken language identification. Various feed-forward fully connected, convolutional and recurrent DNN architectures are adopted and compared against a baseline i-vector based system. Moreover, DNNs are also utilized for extraction of bottleneck features from the input signal. The dataset used for experimental evaluation contains utterances belonging to languages that are all related to each other and sometimes hard to distinguish even for human listeners: it is compiled from recordings of the 11 most widespread Slavic languages. We also released this Slavic dataset to the general public, because a similar collection is not publicly available through any other source. The best results were yielded by a bidirectional recurrent DNN with gated recurrent units that was fed by bottleneck features. In this case, the baseline ER was reduced from 4.2% to 1.2% and Cavg from 2.3% to 0.6%.

## Adding New Classes without Access to the Original Training Data with Applications to Language Identification

*Hagai Taitelbaum, Ehud Ben-Reuven and Jacob Goldberger*

Engineering Faculty, Bar-Ilan University, Israel

**Wed-O-1-2-3, Time: 10:40-11:00**

In this study we address the problem of adding new classes to an existing neural network classifier. We assume that new training data with the new classes is available.

In many applications, dataset used to train machine learning algorithms contain confidential information that cannot be accessed during the process of extending the class set. We propose a method for training an extended class-set classifier using only examples with labels from the new classes while avoiding the problem of forgetting the original classes. This incremental training method is applied to the problem of language identification. We report results on the 50 languages NIST 2015 dataset where we were able to classify all the languages even though only part of the classes was available during the first training phase and the other languages were only available during the second phase.

## Feature Representation of Short Utterances Based on Knowledge Distillation for Spoken Language Identification

*Peng Shen, Xugang Lu, Sheng Li and Hisashi Kawai*

Institute of Information and Communications Technology, Japan

**Wed-O-1-2-4, Time: 11:00-11:20**

The performance of spoken language identification (LID) on short utterances is drastically degraded even though model is completely trained on short utterance data set. The degradation is because of the large pattern confusion caused by the large variation of feature representation on short utterances. In this paper, we propose a teacher-student network learning algorithm to explore discriminative features for short utterances. With the teacher-student network learning, the feature representation for short utterances (explored by the student network) are normalized to their representations corresponding to long utterances (provided by the teacher network). With this learning algorithm, the feature representation on short utterances is supposed to reduce pattern confusion. Experiments on a 10-language LID task were carried out to test the algorithm. Our results showed the proposed algorithm significantly improved the performance.

## Sub-band Envelope Features Using Frequency Domain Linear Prediction for Short Duration Language Identification

*Sarith Fernando[1,2], Vidhyasaharan Sethu[1] and Eliathamby Ambikairajah[1,2]*

[1]School of Electrical Engineering and Telecommunications, UNSW Sydney

[2]DATA61, CSIRO, Sydney, Australia

**Wed-O-1-2-5, Time: 11:20-11:40**

Mismatch between training and testing utterances can significantly degrade the performance of language identification (LID) systems, especially in the case of short duration utterances. This work explores the hypothesis that long-term trends are less affected by this mismatch compared to short-term features. In particular, it proposes the use of features based on temporal envelopes within sub-bands. In this work, the temporal envelopes are obtained using linear prediction in the frequency domain. These envelopes are then transformed into cepstral features. The proposed features are then used as a front-end to a bidirectional long short term memory recurrent neural network to identify languages. Experimental evaluations on the AP17-OLR dataset under different conditions indicate that the proposed features exhibit substantially greater robustness under different noise and mismatch conditions, compared to baseline features. Specifically, the proposed features outperform state-of-the-art bottleneck features and show a relative improvement of 38.4% averaged across the test set.

## Effectiveness of Single-Channel BLSTM Enhancement for Language Identification

*Peter Sibbern Frederiksen[1], Jesús Villalba[2], Shinji Watanabe[2], Zheng-Hua Tan[1] and Najim Dehak[2]*

[1]Department of Electronic Systems, Aalborg University, Denmark

[2]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD

**Wed-O-1-2-6, Time: 11:40-12:00**

This paper proposes to apply deep neural network (DNN)-based single-channel speech enhancement (SE) to language identification. The 2017 language recognition evaluation (LRE17) introduced noisy audios from videos, in addition to the telephone conversation from past challenges. Because of that, adapting models from telephone speech to noisy speech from the video domain was required to obtain optimum performance. However, such adaptation requires knowledge of the audio domain and availability of in-domain data. Instead of adaptation, we propose to use a speech enhancement step to clean up the noisy audio as preprocessing for language identification. We used a bi-directional long short-term memory (BLSTM) neural network, which given log-Mel noisy features predicts a spectral mask indicating how clean each time-frequency bin is. The noisy spectrogram is multiplied by this predicted mask to obtain the enhanced magnitude spectrogram and it is transformed back into the time domain by using the unaltered noisy speech phase. The experiments show significant improvement to language identification of noisy speech, for systems with and without domain adaptation, while preserving the identification performance in the telephone audio domain. In the best adapted state-of-the-art bottleneck i-vector system the relative improvement is 11.3% for noisy speech.

**Notes**

# Wed-O-1-3: Production of Prosody

Hall 2; 10:00-12:00; Wednesday, 5 September, 2018

Chairs: Juraj Simko and Hansjörg Mixdorff

## Articulation Rate as a Speaker Discriminant in British English

*Erica Gold*

Department of Linguistics and Modern Languages, University of Huddersfield, United Kingdom

**Wed-O-1-3-1, Time: 10:00-10:20**

Identifying speech parameters that have both a low level of intra-speaker variability and a high level of inter-speaker variability is key when discriminating between individuals in forensic speaker comparison cases. A substantial amount of research in the field of forensic phonetics has been devoted to identifying highly discriminant speaker parameters. To this end, the vast majority of the existing literature has focused solely on vowels and constants. However, the discriminant power of speaking tempo has yet to be examined, despite its broad use in practice and it having been recognized.

This paper examines, for the first time, the discriminant power of articulation rate (AR) in British English. Approximately 3000 local ARs were measured in this study for 100 Southern Standard British English male speakers. In order to assess the evidential value of AR, likelihood ratios were calculated. The results suggest that AR performs well for same speaker comparisons. However, for different speaker comparisons, the system is performing just worse than chance. Overall, it appears that AR may not be the best speaker discriminant, although it is important to still consider AR in forensic speaker comparisons as there may be some individuals for which AR is highly idiosyncratic.

## Truncation and Compression in Southern German and Australian English

*Jenny Yu[1,2] and Katharina Zahner[3]*

[1]The MARCS Institute, Western Sydney University, Australia

[2]ARC Centre of Excellence for the Dynamics of Language, Australia

[3]University of Konstanz, Germany

**Wed-O-1-3-2, Time: 10:20-10:40**

Nuclear pitch accents are realized differently when there is little sonorant material (as in monosyllabic compared to disyllabic words): Southern British English speakers compress rises and falls, while Northern German speakers truncate falls and compress rises [1] (Grabe 1998). This leads to different phonetic surface patterns for final falls. Within these languages, dialectal variation affects alignment and the frequency of occurrence of nuclear tunes. We test whether the differences in compression and truncation use are a stable cross-linguistic phenomenon (and occur in other varieties of English and German) or whether they are limited to the varieties tested in [1]. Here, we investigated productions of rises and falls in Australian English and Southern German in words with different proportions of sonorant material. Australian English speakers compressed rises and falls, while Southern German speakers only compressed rises but truncated falls, consistent with Grabe's findings for Southern British English and Northern German. This indicates consistent use of strategies within a language, even though the varieties under investigation display other phonetic differences from previous varieties tested. We discuss implications of these findings for automatic labelling.

## Prominence-based Evaluation of L2 Prosody

*Heini Kallio, Antti Suni, Päivi Virkkunen and Juraj Šimko*

University of Helsinki, Finland

**Wed-O-1-3-3, Time: 10:40-11:00**

Prosody in terms of word and sentence stress is one of the most difficult features for many second language (L2) speakers to learn and it can be hypothesized that assessing the learner's prosodic abilities could provide a good measure for assessing the learners' spoken language skills in general. Automatic assessment is, however, dependent on reliable automatic analyses of prosodic features for comparing the productions between native (L1) and L2 speech. Here we investigate, whether estimated prosodic prominence levels of syllables can be used to predict the prosodic competence of Finnish learners of Swedish. Syllable level prominence was estimated for 99 L2 and 25 native Swedish utterances using continuous wavelet transform analysis with combinations of F0, energy and duration features. The L2 utterances were assessed by four expert raters using the revised CEFR scale for prosodic features. Correlations of prominence estimates for L2 utterances with estimates for L1 utterances and linguistic stress patterns were used as a measure of prosodic proficiency of the L2 speakers. The results show that these estimates correlate significantly with the assessments of expert raters. Overall, the results provide strong support for the use of the wavelet-based prominence estimation techniques in automatic assessment of L2 proficiency.

## Length Contrast and Covarying Features: Whistled Speech as a Case Study

*Rachid Ridouane[1], Giuseppina Turco[2] and Julien Meyer[3]*

[1]LPP, CNRS-Sorbonne Nouvelle, Paris

[2]LLF, CNRS-Université Paris-Diderot, Paris

[3]Univ. Grenoble Alpes, CNRS, GIPSA-lab, Grenoble 38000, France

**Wed-O-1-3-4, Time: 11:00-11:20**

The status of covarying features to sound contrasts is a long-standing issue in speech: are they deliberately controlled by the speakers, or are they contingent automatic effects required by the defining features? We address this question by drawing parallels between the way gemination is implemented in spoken language and the way it is rendered in whistled speech. Audio materials were collected with five Berber whistlers in Morocco. The spoken and whistled data were composed of pairs of words contrasting singletons to geminates in different word positions.

Compared to spoken forms, whistling, while adapting to the specific constraints imposed by the medium, transposes the basic strategies used in normal speech. As in normal speech, the primary and most salient acoustic attribute differentiating whistled singletons and geminates is closure duration. But duration is not used alone. Covarying secondary attributes are conveyed which may serve to enhance the primary correlate by contributing additional properties increasing the distance between the two lexical categories. These enhancing correlates may take on distinctive function in cases where the primary correlate is not implemented. This is, for instance, the case of higher frequency values in word-initial position where duration differences cannot be acoustically implemented using whistled speech.

## Information Structure, Affect and Prenuclear Prominence in American English

*Eleanor Chodroff and Jennifer Cole*

Northwestern University, Department of Linguistics, Evanston, IL, USA

**Wed-O-1-3-5, Time: 11:20-11:40**

The influence of information structure (IS: givenness, accessibility, newness and focus) on pitch accent assignment and acoustic prominence measures of prenuclear words was investigated for American English speech elicited through read production of mini-stories. Results showed a consistent pattern of accenting the initial content word in the sentence, supporting an analysis of prenuclear accent as structural, or 'rhythmic'. While no association was observed between IS and accent type (e.g., H*, L*, L+H*, L*+H), the acoustic-phonetic realization of prominence was modulated by information structure. In particular, words that carry contrastive focus generally showed more extreme f0 excursions relative to the average. In addition, there was a strong influence of speaking style or 'affect' on both pitch accent type and the acoustic-phonetic realization of prominence. Speakers were more likely to produce L+H* accents in a lively than a neutral speaking style. Differences in affect were also strongly reflected in f0 excursion, duration and amplitude within the target word. Overall, this study indicates both linguistic (information structure) and paralinguistic (affect) influences on the phonetic

**Notes**

implementation of prenuclear prominence, with varying influence of these two factors on the phonological assignment of prenuclear pitch accents.

### Effects of User Controlled Speech Rate on Intelligibility in Noisy Environments

*John S. Novak, III and Robert V. Kenyon*

University of Illinois at Chicago, United States of America

**Wed-O-1-3-6, Time: 11:40-12:00**

Talkers intentionally producing high-intelligibility speech for listeners in challenging situations often reduce their speech rate. This study affords listeners fine-grained control over the playback rate of a desired speech signal in varying levels of background noise and tests listener intelligibility with their preferred and unmodified rates of speech. We find clear listener preference for decreased rates of speech as background noise increased. However, we also found degraded performance on a speech-in-noise intelligibility test relative to unmodified speech in these same conditions.

## Wed-O-1-4: Speech Intelligibility and Quality

MR G.01-G.02; 10:00-12:00; Wednesday, 5 September, 2018

Chairs: Yannis Stylianou and Sebastian Möller

### Binaural Speech Intelligibility Estimation Using Deep Neural Networks

*Kazuhiro Kondo[1], Kazuya Taira[1,2] and Yosuke Kobayashi[3]*

[1]Yamagata University, Japan

[2]Yamamoto Electric Corp., Japan

[3]Muroran Institute of Technology, Japan

**Wed-O-1-4-1, Time: 10:00-10:20**

We attempted to estimate the speech intelligibility of binaural speech signal with additive noise. The assumption here was that both the target speech signal and the noise source are directional sources. In this case, when the speech and noise sources are located away from each other, the intelligibility generally improves since the human auditory system can potentially segregate these two sources. However since intelligibility tests are commonly conducted using monaurally recorded signals, the intelligibility is often underestimated compared to live human listeners since this segregation capability is neglected. We have previously proposed to use binaurally recorded signals to estimate the speech intelligibility and compared the estimation accuracy of several machine learning methods on this signal. We showed that random forests (RF) combined with the better ear model and Mel filter banks gives the highest accuracy compared to other methods, such as the support vector machines or logistic regression. In this paper, we attempt to introduce deep neural networks (DNN) to this task. Initial evaluation results show that the use of DNN can provide a modest improvement over RF.

### Multi-resolution Gammachirp Envelope Distortion Index for Intelligibility Prediction of Noisy Speech

*Katsuhiko Yamamoto[1], Toshio Irino[1], Narumi Ohashi[1], Shoko Araki[2], Keisuke Kinoshita[2] and Tomohiro Nakatani[2]*

[1]Graduate School of Systems Engineering, Wakayama University, Japan

[2]NTT Communication Science Laboratories, Japan

**Wed-O-1-4-2, Time: 10:20-10:40**

A multi-resolution version of the gammachirp envelope distortion index (mr-GEDI) is proposed for the intelligibility prediction of noisy speech processed using speech enhancement algorithms. The proposed model calculates the short-time signal-to-distortion ratio in the temporal envelope modulation extracted from the output of the gammachirp auditory filterbank. The predictions were compared with human subjective results for various signal-to-noise ratio conditions with pink and babble noise. The mr-GEDI predicts the intelligibility curves better than the hearing-aid speech perception index (HASPI).

### Speech Intelligibility Enhancement Based on a Non-causal Wavenet-like Model

*Muhammed Shifas PV, Vassilis Tsiaras and Yannis Stylianou*

Speech Signal Processing Laboratory (SSPL), University Of Crete, Greece

**Wed-O-1-4-3, Time: 10:40-11:00**

Low speech intelligibility in noisy listening conditions makes more difficult our communication with others. Various strategies have been suggested to modify a speech signal before it is presented in a noisy listening environment with the goal to increase its intelligibility. A state-of-the art approach, referred to as Spectral Shaping and Dynamic Range Compression (SSDRC), relies on modifying spectral and temporal structure of the clean speech and has been shown to considerably improve the intelligibility of speech in noisy listening conditions. In this paper, we present a non-causal Wavenet-like model for mapping clean speech samples to samples generated by SSDRC. A successful non-linear mapping function has the potential to be used a) in improving the intelligibility of noisy speech and b) in the Wavenet-based speech synthesizers as a model based intelligibility improvement layer. Objective and subjective results show that the Wavenet-based mapping function is able to reproduce the intelligibility gains of SSDRC, while by far it improves the quality of the modified signal compared to the quality obtained by SSDRC.

### Quality-Net: An End-to-End Non-intrusive Speech Quality Assessment Model Based on BLSTM

*Szu-wei Fu[1,2], Yu Tsao[1], Hsin-Te Hwang[3] and Hsin-Min Wang[3]*

[1]Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan

[2]Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan

[3]Institute of Information Science, Academia Sinica, Taipei, Taiwan

**Wed-O-1-4-4, Time: 11:00-11:20**

Nowadays, most of the objective speech quality assessment tools (e.g., perceptual evaluation of speech quality (PESQ)) are based on the comparison of the degraded/processed speech with its clean counterpart. The need of a "golden" reference considerably restricts the practicality of such assessment tools in real-world scenarios since the clean reference usually cannot be accessed. On the other hand, human beings can readily evaluate the speech quality without any reference (e.g., mean opinion score (MOS) tests), implying the existence of an objective and non-intrusive (no clean reference needed) quality assessment mechanism. In this study, we propose a novel end-to-end, non-intrusive speech quality evaluation model, termed Quality-Net, based on bidirectional long short-term memory. The evaluation of utterance-level quality in Quality-Net is based on the frame-level assessment. Frame constraints and sensible initializations of forget gate biases are applied to learn meaningful frame-level quality assessment from the utterance-level quality label. Experimental results show that Quality-Net can yield high correlation to PESQ (0.9 for the noisy speech and 0.84 for the speech processed by speech enhancement). We believe that Quality-Net has potential to be used in a wide variety of applications of speech signal processing.

### Global SNR Estimation of Speech Signals Using Entropy and Uncertainty Estimates from Dropout Networks

*Rohith Aralikatti, Dilip Kumar Margam, Tanay Sharma, Abhinav Thanda and Shankar Venkatesan*

Samsung R&D Institute Bangalore, India

**Wed-O-1-4-5, Time: 11:20-11:40**

This paper demonstrates two novel methods to estimate the global SNR of speech signals. In both methods, Deep Neural Network-Hidden Markov

**Notes**

Model (DNN-HMM) acoustic model used in speech recognition systems is leveraged for the additional task of SNR estimation. In the first method, SNR is estimated using the entropy of the posterior distribution obtained from DNN of an ASR system. Recent work on bayesian deep learning has shown that a DNN-HMM trained with dropout can be used to estimate model uncertainty by approximating it as a deep Gaussian process. In the second method, this approximation is used to obtain model uncertainty estimates. Noise specific regressors are used to predict the SNR from the entropy and model uncertainty. The DNN-HMM is trained on GRID corpus and tested on different noise profiles from the DEMAND noise database at SNR levels ranging from -10 dB to 30 dB.

### Detecting Packet-Loss Concealment Using Formant Features and Decision Tree Learning

*Gabriel Mittag[1] and Sebastian Möller[1,2]*

[1]Quality and Usability Lab, Technische Universität Berlin, Berlin, Germany

[2]Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Berlin, Germany

**Wed-O-1-4-6, Time: 11:40-12:00**

One of the main quality impairments in today's packet-based voice services are interruptions caused by transmission errors. Therefore, most codecs comprise concealment algorithms that attempt to reduce the perceived quality degradation of missing speech packets. In case the algorithm fails to properly synthesize the lost speech, interruptions or unnatural sounds are usually perceivable by the user. When measuring the quality of a voice network, there are excellent tools available, which can predict the perceived speech quality. However, they offer only little insight into the technical cause of a quality degradation. A packet-loss detection model could explain the influence of transmission errors on the speech quality and state a packet-loss rate. Thus, making it easier to identify technical problems in the network. In this paper, we examine a new approach for detecting (perceived) packet-loss of transmitted speech by audio analysis. After finding a lost packet, the model classifies in a second stage if the loss was perceivable as a quality degradation. In the model, we use meaningful features that are easy to interpret and obtained promising results in a simulated environment. Therefore, this detector could also be used to evaluate new packet-loss concealment algorithms and help in optimizing the same.

## Wed-SS-1-1: Special Session: Integrating Speech Science and Technology for Clinical Applications

MR G.03-G.04; 10:00-12:00; Wednesday, 5 September, 2018

Chairs: Christina Hagedorn, Shrikanth Narayanan and Uttam Sinha

### UltraSuite: A Repository of Ultrasound and Acoustic Data from Child Speech Therapy Sessions

*Aciel Eshky[1], Manuel Sam Ribeiro[1], Joanne Cleland[2], Korin Richmond[1], Zoe Roxburgh[3], James M Scobbie[3] and Alan Wrench[3,4]*

[1]Centre for Speech Technology Research, University of Edinburgh, UK

[2]Psychological Sciences and Health, University of Strathclyde, UK

[3]Clinical Audiology, Speech and Language Research Centre, Queen Margaret University, UK

[4]Articulate Instruments Ltd., UK

**Wed-SS-1-1-1, Time: 10:00-10:20**

We introduce UltraSuite, a curated repository of ultrasound and acoustic data, collected from recordings of child speech therapy sessions. This release includes three data collections, one from typically developing children and two from children with speech sound disorders. In addition, it includes a set of annotations, some manual and some automatically produced and software tools to process, transform and visualise the data.

### Detecting Signs of Dementia Using Word Vector Representations

*Bahman Mirheidari[1], Daniel Blackburn[2], Traci Walker[3], Annalena Venneri[2], Markus Reuber[4] and Heidi Christensen[1,5]*

[1]Department of Computer Science, University of Sheffield

[2]Sheffield Institute for Translational Neuroscience (SITraN), University of Sheffield

[3]Department of Human Communication Sciences, University of Sheffield

[4]Academic Neurology Unit, University of Sheffield, Royal Hallamshire Hospital

[5]Centre for Assistive Technology and Connected Healthcare (CATCH), University of Sheffield, Sheffield, UK

**Wed-SS-1-1-2, Time: 10:20-10:40**

Recent approaches to word vector representations, e.g., 'w2vec' and 'GloVe', have been shown to be powerful methods for capturing the semantics and syntax of words in a text. The approaches model the co-occurrences of words and recent successful applications on written text have shown how the vector representations and their interrelations represent the meaning or sentiment in the text. Most applications have targeted written language, however, in this paper, we investigate how these models port to the spoken language domain where the text is the result of (erroneous) automatic speech transcription. In particular, we are interested in the task of detecting signs of dementia in a person's spoken language. This is motivated by the fact that early signs of dementia are known to affect a person's ability to express meaning articulately for example when they engage in a conversation – something which is known to be cognitively very demanding. We analyse conversations designed to probe people's short and long-term memory and propose three different methods for how word vectors may be used in a classification setup. We show that it is possible to identify dementia from the output of a speech recogniser despite a high occurrence of recognition errors.

### Classification of Huntington Disease Using Acoustic and Lexical Features

*Matthew Perez[1], Wenyu Jin[1], Duc Le[1], Noelle Carlozzi[2], Praveen Dayalu[3], Angela Roberts[4] and Emily Mower Provost[1]*

[1]Computer Science and Engineering, University of Michigan, Ann Arbor, MI

[2]Physical Medicine & Rehabilitation, University of Michigan, Ann Arbor, MI

[3]Michigan Medicine, University of Michigan, Ann Arbor, MI

[4]Department of Communication Sciences and Disorders, Northwestern University, Evanston, IL

**Wed-SS-1-1-3, Time: 10:40-11:00**

Speech is a critical biomarker for Huntington Disease (HD), with changes in speech increasing in severity as the disease progresses. Speech analyses are currently conducted using either transcriptions created manually by trained professionals or using global rating scales. Manual transcription is both expensive and time-consuming and global rating scales may lack sufficient sensitivity and fidelity. Ultimately, what is needed is an unobtrusive measure that can cheaply and continuously track disease progression. We present first steps towards the development of such a system, demonstrating the ability to automatically differentiate between healthy controls and individuals with HD using speech cues. The results provide evidence that objective analyses can be used to support clinical diagnoses, moving towards the tracking of symptomatology outside of laboratory and clinical environments.

**Notes**

### The PRIORI Emotion Dataset: Linking Mood to Emotion Detected In-the-Wild

*Soheil Khorram[1,2], Mimansa Jaiswal[1], John Gideon[1], Melvin McInnis[2] and Emily Mower Provost[1]*

[1]Department of Computer Science and Engineering, University of Michigan

[2]Department of Psychiatry, University of Michigan

**Wed-SS-1-1-4, Time: 11:00-11:20**

Bipolar Disorder is a chronic psychiatric illness characterized by pathological mood swings associated with severe disruptions in emotion regulation. Clinical monitoring of mood is key to the care of these dynamic and incapacitating mood states. Frequent and detailed monitoring improves clinical sensitivity to detect mood state changes, but typically requires costly and limited resources. Speech characteristics change during both depressed and manic states, suggesting automatic methods applied to the speech signal can be effectively used to monitor mood state changes. However, speech is modulated by many factors, which renders mood state prediction challenging. We hypothesize that emotion can be used as an intermediary step to improve mood state prediction. This paper presents critical steps in developing this pipeline, including (1) a new in the wild emotion dataset, the PRIORI Emotion Dataset, collected from everyday smartphone conversational speech recordings, (2) activation/valence emotion recognition baselines on this dataset (PCC of 0.71 and 0.41, respectively) and (3) significant correlation between predicted emotion and mood state for individuals with bipolar disorder. This provides evidence and a working baseline for the use of emotion as a meta-feature for mood state monitoring.

### Language Features for Automated Evaluation of Cognitive Behavior Psychotherapy Sessions

*Nikolaos Flemotomos[1], Victor Martinez[2], James Gibson[1], David Atkins[3], Torrey Creed[4] and Shrikanth Narayanan[1,2]*

[1]Department of Electrical Engineering

[2]Department of Computer Science,
University of Southern California, Los Angeles, CA, USA

[3]Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA

[4]Department of Psychiatry, University of Pennsylvania, Philadelphia, PA, USA

**Wed-SS-1-1-5, Time: 11:20-11:40**

Cognitive Behavior Therapy (CBT) is a psychotherapy treatment that uses cognitive change strategies to address mental health problems. Quality assessment of a CBT session is traditionally addressed by human raters who evaluate recorded sessions along specific behavioral codes, a cost prohibitive and time consuming method. In this work we examine how linguistic features can be effectively used to develop an automatic competency rating tool for CBT. We explore both standard, widely-used lexical features and domain-specific ones, adapting methods which have been successfully used in similar psychotherapy session coding tasks. Experiments are conducted on manual transcripts of CBT sessions and on automatically derived ones, thus introducing an end-to-end approach. Our results suggest that a real-world system could be developed to automatically evaluate CBT sessions to assist training, supervision, or quality assurance of services.

### Automatic Early Detection of Amyotrophic Lateral Sclerosis from Intelligible Speech Using Convolutional Neural Networks

*Kwanghoon An[1], Myungjong Kim[1], Kristin Teplansky[1,2], Jordan Green[3], Thomas Campbell[2], Yana Yunusova[4], Daragh Heitzman[5] and Jun Wang[1,2]*

[1]Speech Disorders & Technology Lab, Department of Bioengineering

[2]Callier Center for Communication Disorders, University of Texas at Dallas, United States

[3]Department of Communication Sciences and Disorders, MGH Institute of Health Professions, United States

[4]Department of Speech-Language Pathology, University of Toronto, Canada

[5]MDA/ALS Center, Texas Neurology, United States

**Wed-SS-1-1-6, Time: 11:40-12:00**

Amyotrophic lateral sclerosis (ALS) is a rapidly progressive neurodegenerative disease of the motor system that leads to the impairment of speech and swallowing functions. The lack of a biomarker typically causes a diagnostic delay. To advance the current diagnostic process, we explored the feasibility of automatic detection of patients with ALS at an early stage from highly intelligible speech. A speech dataset was collected from thirteen newly diagnosed patients with ALS and thirteen age- and gender-matched healthy controls. Convolutional Neural Networks (CNNs), including time-domain CNN and frequency-domain CNN, were used to classify the intelligible speech produced by patients with ALS and those by healthy individuals. Experimental results indicated both time- and frequency-CNN outperformed standard neural network. The best sample-level sensitivity and specificity were obtained by time-CNN (71.6% and 80.9%, respectively). When multiple samples were used to vote to estimate a person-level performance, the best result was obtained by frequency-CNN (76.9% sensitivity and 92.3% specificity). Results demonstrated the possibility of early detection of ALS from intelligible speech signals.

## Wed-SS-1-2: Special Session: Speech Technologies for Code-Switching in Multilingual Communities

MR 1.01-1.02; 10:00-12:00; Wednesday, 5 September, 2018
Chairs: Kalika Bali, Alan W Black, Mona Diab, Julia Hirschberg, Sunayana Sitaram and Thamar Solorio

### A Study of Lexical and Prosodic Cues to Segmentation in a Hindi-English Code-switched Discourse

*Preeti Rao[1], Mugdha Pandya[2], Kamini Sabu[1], Kanhaiya Kumar[1] and Nandini Bondale[3]*

[1]Department of Electrical Engineering, Indian Institute of Technology Bombay, India

[2]Vellore Institute of Technology, India

[3]School of Technology and Comp. Science, Tata Institute of Fundamental Research, India

**Wed-SS-1-2-1, Time: 10:00-10:20**

Bilingualism, almost universal in India, routinely appears in communication in many forms. Code-switching with English is common among city dwellers with the matrix language typically being the speaker's native tongue. While a number of English words have made their way into the lexicon of Indian languages, also prevalent is insertional code-switching, i.e. switching at sentence or clause level. We consider an interesting and widely encountered variety of code-switched speech in the form of public discourses by a popular motivational speaker who uses English, probably for effect, in her Hindi language speeches. We effectively observe three categories of segments in the discourse: Hindi, Hindi with embedded English words and English. In this work, we present the characteristics of our data and investigate the discrimination potential of lexical and prosodic cues on manually segmented fragments. Lexical cues are obtained via Google Speech API for Indian English recognition. Prosodic cues computed from pitch, intensity and syllable duration estimates are found to demonstrate significant differences between Hindi and English segments, indicating more careful articulation of the embedded language.

**Notes**

## Building a Unified Code-Switching ASR System for South African Languages

*Emre Yılmaz[1,2], Astik Biswas[3], Ewald van der Westhuizen[3], Febe de Wet[3] and Thomas Niesler[3]*

[1]CLS/CLST, Radboud University, Nijmegen, Netherlands

[2]Dept. of Electrical and Computer Engineering, National University of Singapore, Singapore

[3]Dept. of Electrical and Electronic Engineering, Stellenbosch University, South Africa

**Wed-SS-1-2-2, Time: 10:20-10:40**

We present our first efforts towards building a single multilingual automatic speech recognition (ASR) system that can process code-switching (CS) speech in five languages spoken within the same population. This contrasts with related prior work which focuses on the recognition of CS speech in bilingual scenarios. Recently, we have compiled a small five-language corpus of South African soap opera speech which contains examples of CS between 5 languages occurring in various contexts such as using English as the matrix language and switching to other indigenous languages. The ASR system presented in this work is trained on 4 corpora containing English-isiZulu, English-isiXhosa, English-Setswana and English-Sesotho CS speech. The interpolation of multiple language models trained on these language pairs enables the ASR system to hypothesize mixed word sequences from these 5 languages. We evaluate various state-of-the-art acoustic models trained on this 5-lingual training data and report ASR accuracy and language recognition performance on the development and test sets of the South African multilingual soap opera corpus.

## Study of Semi-supervised Approaches to Improving English-Mandarin Code-Switching Speech Recognition

*Pengcheng Guo[1,2], Haihua Xu[2], Lei Xie[1] and Eng Siong Chng[2,3]*

[1]School of Computer Science, Northwestern Polytechnical University, Xi'an, China

[2]Temasek Laboratories, Nanyang Technological University, Singapore

[3]School of Computer Science and Engineering, Nanyang Technological University, Singapore

**Wed-SS-1-2-3, Time: 10:40-11:00**

In this paper, we present our overall efforts to improve the performance of a code-switching speech recognition system using semi-supervised training methods from lexicon learning to acoustic modeling, on the South East Asian Mandarin-English (SEAME) data. We first investigate semi-supervised lexicon learning approach to adapt the canonical lexicon, which is meant to alleviate the heavily accented pronunciation issue within the code-switching conversation of the local area. As a result, the learned lexicon yields improved performance. Furthermore, we attempt to use semi-supervised training to deal with those transcriptions that are highly mismatched between human transcribers and ASR system. Specifically, we conduct semi-supervised training assuming those poorly transcribed data as unsupervised data. We found the semi-supervised acoustic modeling can lead to improved results. Finally, to make up for the limitation of the conventional n-gram language models due to data sparsity issue, we perform lattice rescoring using neural network language models and significant WER reduction is obtained.

## Acoustic and Textual Data Augmentation for Improved ASR of Code-Switching Speech

*Emre Yılmaz[1,2], Henk van den Heuvel[1] and David van Leeuwen[1]*

[1]CLS/CLST, Radboud University, Nijmegen, Netherlands

[2]Dept. of Electrical and Computer Engineering, National University of Singapore, Singapore

**Wed-SS-1-2-4, Time: 11:00-12:00**

In this paper, we describe several techniques for improving the acoustic and language model of an automatic speech recognition (ASR) system operating on code-switching (CS) speech. We focus on the recognition of Frisian-Dutch radio broadcasts where one of the mixed languages, namely Frisian, is an under-resourced language. In previous work, we have proposed several automatic transcription strategies for CS speech to increase the amount of available training speech data. In this work, we explore how the acoustic modeling (AM) can benefit from monolingual speech data belonging to the high-resourced mixed language. For this purpose, we train state-of-the-art AMs, which were ineffective due to lack of training data, on a significantly increased amount of CS speech and monolingual Dutch speech. Moreover, we improve the language model (LM) by creating code-switching text, which is in practice almost non-existent, by (1) generating text using recurrent LMs trained on the transcriptions of the training CS speech data, (2) adding the transcriptions of the automatically transcribed CS speech data and (3) translating Dutch text extracted from the transcriptions of a large Dutch speech corpora. We report significantly improved CS ASR performance due to the increase in the acoustic and textual training data.

## The Role of Cognate Words, POS Tags and Entrainment in Code-Switching

*Victor Soto, Nishmar Cestero and Julia Hirschberg*

Columbia University, New York, USA

**Wed-SS-1-2-5, Time: 11:00-12:00**

The linguistic or contextual stimuli that elicit code-switching are largely unknown, despite the fact that these are of key importance to understanding mixed language and building tools that can handle it. In this paper, we test the following hypotheses proposed in linguistics literature: first, that cognate stimuli are directly correlated to code-switching; second, that syntactic information facilitates or inhibits code switching; and third that speakers entrain to one another in code-switching in conversation between bilinguals. In order to test these hypotheses, we built a lexical database of cognate pairs for English Spanish. Using statistical significance tests on a corpus of conversational code-switched English Spanish, we found that a) there is strong statistical evidence that cognates and switches occur simultaneously in the same utterance and that cognates facilitate switching when they precede a code-switch, b) there is strong statistical evidence of the relationship between part-of-speech tags and code-switching and c) speakers tend to show converging entrainment behavior with respect to their rate of code-switching in conversation.

## Homophone Identification and Merging for Code-switched Speech Recognition

*Brij Mohan Lal Srivastava and Sunayana Sitaram*

Microsoft Research India

**Wed-SS-1-2-6, Time: 11:00-12:00**

Code-switching or mixing is the use of multiple languages in a single utterance or conversation. Borrowing occurs when a word from a foreign language becomes part of the vocabulary of a language. In multilingual societies, switching/mixing and borrowing are not always clearly distinguishable. Due to this, transcription of code-switched and borrowed words is often not standardized and leads to the presence of homophones in the training data. In this work, we automatically identify and disambiguate homophones in code-switched data to improve recognition of code-switched speech. We use a WX-based common pronunciation scheme for both languages being mixed and unify the homophones during training, which results in a lower word error rate for systems built using this data. We also extend this framework to propose a metric for code-switched speech recognition that takes into account homophones in both languages while calculating WER, which can help provide a more accurate picture of errors the ASR system makes on code-switched speech.

## Code-switching in Indic Speech Synthesisers

*Anju Leela Thomas, Anusha Prakash, Arun Baby and Hema Murthy*

Indian Institute of Technology Madras, India

**Wed-SS-1-2-7, Time: 11:00-12:00**

Most Indians are inherently bilingual or multilingual owing to the diverse linguistic culture in India. As a result, code-switching is quite common in conversational speech. The objective of this work is to train good quality text-to-speech (TTS) synthesisers that can seamlessly handle code-switching. To achieve this, bilingual TTSes that are capable of handling phonotactic variations across languages are trained using combinations of monolingual data in a unified framework. In addition to segmenting Indic speech data using signal processing cues in tandem with hidden Markov model-deep

**Notes**

neural network (HMM-DNN), we propose to segment Indian English data using the same approach after NIST syllabification. Then, bilingual HTS-STRAIGHT based systems are trained by randomizing the order of data so that the systematic interactions between the two languages are captured better. Experiments are conducted by considering three language pairs: Hindi+English, Tamil+English and Hindi+Tamil. The code-switched systems are evaluated on monolingual, code-mixed and code-switched texts. Degradation mean opinion score (DMOS) for monolingual sentences shows marginal degradation over that of an equivalent monolingual TTS system, while the DMOS for bilingual sentences is significantly better than that of the corresponding monolingual TTS systems.

### A Novel Approach for Effective Recognition of the Code-Switched Data on Monolingual Language Model

*Sreeram Ganji and Rohit Sinha*

Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati - 781039, India

**Wed-SS-1-2-8, Time: 11:00-12:00**

Code-switching refers to the phenomena of mixing of words or phrases from foreign languages while communicating in a native language by the multilingual speakers. Code-switching is a global phenomenon and is widely accepted in multilingual communities. However, for training the language model (LM) for such tasks, a very limited code-switched textual resources are available as yet. In this work, we present an approach to reduce the perplexity (PPL) of Hindi-English code-switched data when tested over the LM trained on purely native Hindi data. For this purpose, we propose a novel textual feature which allows the LM to predict the code-switching instances. The proposed feature is referred to as code-switching factor (CS-factor). Also, we developed a tagger that facilitates the automatic tagging of the code-switching instances. This tagger is trained on a development data and assigns an equivalent class of foreign (English) words to each of the potential native (Hindi) words. For this study, the textual resource has been created by crawling the blogs from a couple of websites educating about the usage of the Internet. In the context of recognition of the code-switching data, the proposed technique is found to yield a substantial improvement in terms of PPL.

## Wed-S&T-1-1: Show and Tell 5

MR G.05-G.06; 10:00-12:00; Wednesday, 5 September, 2018

Chairs: Venkata Subramanian Viraraghavan and K. Sreenivasa Rao

### Hierarchical Accent Determination and Application in a Large Scale ASR System

*Ramya Viswanathan, Periyasamy Paramasivam and Jithendra Vepa*

Samsung R&D Institute, India - Bangalore

**Wed-S&T-1-1-1, Time: 10:00-12:00**

In deploying Automatic Speech Recognition Systems (ASR) on a global scale, several challenges arise for supporting a widely used language such as English. The primary one among them is to deal with a wide variety of accents. We propose a Hierarchical Accent Determination system that deals with accent variations across large geographical regions at macro level and then the variations at the sub-regions within a selected large geographical region at micro level along with taking context cues. Eight accents [GB, US, Australian, Canadian, Spanish, Korean, Indian & Chinese] are identified at macro level and accent-specific models corresponding to the identified accents are used. The accuracy of the accent identification system is around 80% with ASR as well as using context cues such as phone language and keyboard language. The deployment of the accent identification system has improved the overall accuracy of Speech Recognition system by 10% for accented speech. It is planned to expand the approach to identify accents with significant variations found at sub-regional level in India such as Hindi, Tamil, Telugu, Malayalam and Bengali.

### Toward Scalable Dialog Technology for Conversational Language Learning: Case Study of the TOEFL® MOOC

*Vikram Ramanarayanan[1], David Pautler[1], Patrick Lange[1], Eugene Tsuprun[2], Rutuja Ubale[1], Keelan Evanini[2] and David Suendermann-Oeft[1]*

[1]Educational Testing Service R&D, 90 New Montgomery Street, Suite #1500, San Francisco, CA, USA

[2]Educational Testing Service R&D, 660 Rosedale Rd., Princeton, NJ, USA

**Wed-S&T-1-1-2, Time: 10:00-12:00**

We present a scalable, dialog-based conversational practice tool for English language learners that is operationally deployed on the TOEFL® MOOC. The tool consists of three applications of varying duration that recognize the learner's speech input and responds appropriately. Learners are also provided with basic feedback regarding task performance once complete. We envision this as the first milestone towards the proliferation of many such scalable dialog applications that can help language learners practice, assess and improve their spoken conversation skills.

### Machine Learning Powered Data Platform for High-Quality Speech and NLP Workflows

*João Freitas, Jorge Ribeiro, Daan Baldwijns, Sara Oliveira and Daniela Braga*

DefinedCrowd Corporation

**Wed-S&T-1-1-3, Time: 10:00-12:00**

Machine learning (ML) models - like deep neural networks - require substantial amounts of training data. Also, the training dataset should be properly annotated to obtain satisfactory results. This paper describes a platform designed to create high-quality datasets. By using data workflows adapted for speech technologies and natural language processing systems, the user can collect and enrich speech and text data. Depending on the end goal, the data is passed through multiple processing steps based on human input and ML services. To guarantee data quality, the platform combines several mechanisms like language tests, real-time audits and user behavior into several ML models that act as quality gateways.

### Fully Automatic Speaker Separation System, with Automatic Enrolling of Recurrent Speakers

*Raphael Cohen, Orgad Keller, Jason Levy, Russell Levy, Micha Breakstone and Amit Ashkenazi*

Chorus.ai

**Wed-S&T-1-1-4, Time: 10:00-12:00**

We present a system to enable speaker separation and identification, designed to operate without requiring any effort from the end-user.

In the system, single channel conversations are transformed into i-vectors, clustered into speakers and matched to a database of known speakers. Enrollment is automatic and a voice print is constructed for the recording user, taking advantage of the meta-data identifying that user's conversations. Further information is used when available from other information sources such as video and the ASR transcribed content to identify speakers.

We describe the system architecture, novel unsupervised enrollment algorithm and describe the difficulties encountered in solving this problem.

### Online Speech Translation System for Tamil

*Madhavaraj Ayyavu, Shiva Kumar H R and Ramakrishnan A G*

MILE Lab, Electrical Engineering, Indian Institute of Science, Bangalore 560012, India

**Notes**

**Wed-S&T-1-1-5, Time: 10:00-12:00**
In this paper, we present an application, which recognizes spoken Tamil utterances and speaks out the recognized text in Tamil through our Tamil text-to-speech (TTS) system. Further, we translate the recognized Tamil text to English using google translate and play it through our English TTS. Our Tamil speech recognition system, which can recognize about 75,000 words, has been trained on a 150-hour transcribed speech corpus. We have trained a deep neural network for the acoustic model and employed tri-gram language models to build our recognition system. Our Thirukkural TTS system performs unit-selection based, concatenative speech synthesis, using 2.5 hours of Tamil spoken utterances transcribed at the phone-level. Our English TTS uses 2.7 hours of phone-transcribed utterances. This is a technology demonstration of a complete web application, which, when perfected, could be used to assist Tamil users in learning English, by speaking in Tamil into the system. The playback of the recognized text from Tamil TTS serves to demonstrate the effectiveness of the Tamil ASR to the majority of the conference registrants (who cannot read the recognized Tamil text).

# Wed-P-1-1: Voice Conversion and Speech Synthesis

Hall 4-6: Poster-1; 10:00-12:00; Wednesday, 5 September, 2018
Chair: Paavo Alku

## Unsupervised Vocal Tract Length Warped Posterior Features for Non-Parallel Voice Conversion

*Nirmesh Shah[1], Maulik C. Madhavi[2] and Hemant Patil[1]*

[1]Speech Research Lab, DA-IICT, Gandhinagar-382007, India
[2]Electrical and Computer Engineering Department, National University of Singapore, Singapore

**Wed-P-1-1-1, Time: 10:00-12:00**
In the non-parallel Voice Conversion (VC) with the Iterative combination of Nearest Neighbor search step and Conversion step Alignment (INCA) algorithm, the occurrence of one-to-many and many-to-one pairs in the training data will deteriorate the performance of the stand-alone VC system. The work on handling these pairs during the training is less explored. In this paper, we establish the relationship via intermediate speaker-independent posteriorgram representation, instead of directly mapping the source spectrum to the target spectrum. To that effect, a Deep Neural Network (DNN) is used to map the source spectrum to posteriorgram representation and another DNN is used to map this posteriorgram representation to the target speaker's spectrum. In this paper, we propose to use unsupervised Vocal Tract Length Normalization (VTLN)-based warped Gaussian posteriorgram features as the speaker-independent representations. We performed experiments on a small subset of publicly available Voice Conversion Challenge (VCC) 2016 database. We obtain the lower Mel Cepstral Distortion (MCD) values with the proposed approach compared to the baseline as well as the supervised phonetic posteriorgram feature-based speaker-independent representations. Furthermore, subjective evaluation gave relative improvement of 13.3% with the proposed approach in terms of Speaker Similarity (SS).

## Voice Conversion with Conditional SampleRNN

*Cong Zhou, Michael Horgan, Vivek Kumar, Cristina Vasco and Dan Darcy*

Dolby Laboratories, San Francisco, CA, USA

**Wed-P-1-1-2, Time: 10:00-12:00**
Here we present a novel approach to conditioning the SampleRNN [1] generative model for voice conversion (VC). Conventional methods for VC modify the perceived speaker identity by converting between source and target acoustic features. Our approach focuses on preserving voice content and depends on the generative network to learn voice style. We first train a multi-speaker SampleRNN model conditioned on linguistic features, pitch contour and speaker identity using a multi-speaker speech corpus. Voice-converted speech is generated using linguistic features and pitch contour extracted from the source speaker and the target speaker identity. We demonstrate that our system is capable of many-to-many voice conversion without requiring parallel data, enabling broad applications. Subjective evaluation demonstrates that our approach outperforms conventional VC methods.

## A Voice Conversion Framework with Tandem Feature Sparse Representation and Speaker-Adapted WaveNet Vocoder

*Berrak Sisman, Mingyang Zhang and Haizhou Li*

National University of Singapore, Singapore

**Wed-P-1-1-3, Time: 10:00-12:00**
A voice conversion system typically consists of two modules, the feature conversion module that is followed by a vocoder. The exemplar-based sparse representation marks a success in feature conversion when we only have a very limited amount of training data. While parametric vocoder is generally designed to simulate the mechanics of the human speech generation process under certain simplification assumptions, it doesn't work consistently well for all target applications. In this paper, we study two effective ways to make use of the limited amount of training data for voice conversion. Firstly, we study a novel technique for sparse representation that augments the spectral features with phonetic information, or Tandem Feature. Secondly, we study the use of WaveNet vocoder that can be trained on multi-speaker and target speaker data to improve the vocoding quality. We evaluate that the proposed strategy with Tandem Feature and WaveNet vocoder and show that it provides performance improvement consistently over the traditional sparse representations framework in objective and subjective evaluations.

## WaveNet Vocoder with Limited Training Data for Voice Conversion

*Li-Juan Liu[1], Zhen-Hua Ling[2], Yuan Jiang[1], Ming Zhou[1] and Li-Rong Dai[2]*

[1]iFLYTEK Research, iFLYTEK Co., Ltd.
[2]National Engineering Laboratory of Speech and Language Information Processing, University of Science and Technology of China, Hefei, P.R.China

**Wed-P-1-1-4, Time: 10:00-12:00**
This paper investigates the approaches of building WaveNet vocoders with limited training data for voice conversion (VC). Current VC systems using statistical acoustic models always suffer from the quality degradation of converted speech. One of the major causes is the use of hand-crafted vocoders for waveform generation. Recently, with the emergence of WaveNet for waveform modeling, speaker-dependent WaveNet vocoders have been proposed and they can reconstruct speech with better quality than conventional vocoders, such as STRAIGHT. Because training a WaveNet vocoder in the speaker-dependent way requires a relatively large training dataset, it remains a challenge to build a high-quality WaveNet vocoder for VC tasks when the training data of target speakers is limited. In this paper, we propose to build WaveNet vocoders by combining the initialization using a multi-speaker corpus and the adaptation using a small amount of target data and evaluate this proposed method on the Voice Conversion Challenge (VCC) 2018 dataset which contains approximately 5 minute recordings for each target speaker. Experimental results show that the WaveNet vocoders built using our proposed method outperform conventional STRAIGHT vocoder. Furthermore, our system achieves an average naturalness MOS of 4.13 in VCC 2018, which is the highest among all submitted systems.

## Collapsed Speech Segment Detection and Suppression for WaveNet Vocoder

*Yi-Chiao Wu[1], Kazuhiro Kobayashi[2], Tomoki Hayashi[3], Patrick Lumban Tobing[1] and Tomoki Toda[2]*

[1]Graduate School of Informatics, Nagoya University, Japan

**Notes**

[2]Information Technology Center, Nagoya University, Japan
[3]Graduate School of Information Science, Nagoya University, Japan

**Wed-P-1-1-5, Time: 10:00-12:00**

In this paper, we propose a technique to alleviate the quality degradation caused by collapsed speech segments sometimes generated by the WaveNet vocoder. The effectiveness of the WaveNet vocoder for generating natural speech from acoustic features has been proved in recent works. However, it sometimes generates very noisy speech with collapsed speech segments when only a limited amount of training data is available or significant acoustic mismatches exist between the training and testing data. Such a limitation on the corpus and limited ability of the model can easily occur in some speech generation applications, such as voice conversion and speech enhancement. To address this problem, we propose a technique to automatically detect collapsed speech segments. Moreover, to refine the detected segments, we also propose a waveform generation technique for WaveNet using a linear predictive coding constraint. Verification and subjective tests are conducted to investigate the effectiveness of the proposed techniques. The verification results indicate that the detection technique can detect most collapsed segments. The subjective evaluations of voice conversion demonstrate that the generation technique significantly improves the speech quality while maintaining the same speaker similarity.

## High-quality Voice Conversion Using Spectrogram-Based WaveNet Vocoder

*Kuan Chen, Bo Chen, Jiahao Lai and Kai Yu*
Key Lab. of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering,
SpeechLab, Department of Computer Science and Engineering,
Brain Science and Technology Research Center
Shanghai Jiao Tong University, Shanghai, China

**Wed-P-1-1-6, Time: 10:00-12:00**

Waveform generator is a key component in voice conversion. Recently, WaveNet waveform generator conditioned on the Mel-cepstrum (Mcep) has shown better quality over standard vocoder. In this paper, an enhanced WaveNet model based on spectrogram is proposed to further improve voice conversion performance. Here, Mel-frequency spectrogram is converted from source speaker to target speaker using an LSTM-RNN based frame-to-frame feature mapping. To evaluate the performance, the proposed approach is compared to an Mcep based LSTM-RNN voice conversion system. Both STRAIGHT vocoder and Mcep-based WaveNet vocoder are elected to produce the converted speech for Mcep conversion system. The fundamental frequency (F0) of the converted speech in different systems is analyzed. The naturalness, similarity and intelligibility are evaluated in subjective measures. Results show that the spectrogram based WaveNet waveform generator can achieve better voice conversion quality compared to traditional WaveNet approaches. The Mel-spectrogram based voice conversion can achieve significant improvement in speaker similarity and inherent F0 conversion.

## Spanish Statistical Parametric Speech Synthesis Using a Neural Vocoder

*Antonio Bonafonte, Santiago Pascual and Georgina Dorca*
Universitat Politècnica de Catalunya, Barcelona, Spain

**Wed-P-1-1-7, Time: 10:00-12:00**

During the 2000s decade, unit-selection based text-to-speech was the dominant commercial technology. Meanwhile, the TTS research community has made a big effort to push statistical-parametric speech synthesis to get similar quality and more flexibility on the generated voice. During last years, deep learning advances applied to speech synthesis have filled the gap, specially when neural vocoders substitute traditional signal-processing based vocoders.

In this paper we substitute the waveform generation vocoder of MUSA, our Spanish TTS, with SampleRNN, a neural vocoder which was recently proposed as a deep autoregressive raw waveform generation model. MUSA uses recurrent neural networks to predict vocoder parameters (MFCC and logF0) from linguistic features. Then, the Ahocoder vocoder is used to recover the speech waveform out of the predicted parameters. In the first system SampleRNN is extended to generate speech conditioned on the Ahocoder generated parameters, where two configurations have been considered to train the system. First, the parameters derived from the signal using Ahocoder are used. Secondly, the system is trained with the parameters predicted by MUSA, where SampleRNN and MUSA are jointly optimized. The subjective evaluation shows that the second system outperforms both the original Ahocoder and SampleRNN as an independent neural vocoder.

## Experiments with Training Corpora for Statistical Text-to-speech Systems

*Monika Podsiadło[1] and Victor Ungureanu[2]*
[1]Google Inc.
[2]Google Switzerland

**Wed-P-1-1-8, Time: 10:00-12:00**

Common text-to-speech (TTS) systems rely on training data for modelling human speech. The quality of this data can range from professional voice actors recording hand-curated sentences in high-quality studio conditions, to found voice data representing arbitrary domains. For years, the unit selection technology dominant in the field required many hours of data that was expensive and time-consuming to collect. With the advancement of statistical methods of waveform generation, there have been experiments with more noisy and often much larger datasets, testing the inherent flexibility of such systems. In this paper we examine the relationship between training data and speech synthesis quality. We then hypothesise that statistical text-to-speech benefits from high acoustic quality corpora with high level of prosodic variation, but that beyond the first few hours of training data we do not observe quality gains. We then describe how we engineered a training dataset containing optimized distribution of features and how these features were defined. Lastly, we present results from a series of evaluation tests. These confirm our hypothesis and show how a carefully engineered training corpus of a smaller size yields the same speech quality as much larger datasets, particularly for voices that use WaveNet.

## Multi-task WaveNet: A Multi-task Generative Model for Statistical Parametric Speech Synthesis without Fundamental Frequency Conditions

*Yu Gu and Yongguo Kang*
Baidu Speech Department,
Baidu Technology Park, Beijing, 100193, China

**Wed-P-1-1-9, Time: 10:00-12:00**

This paper introduces an improved generative model for statistical parametric speech synthesis (SPSS) based on WaveNet under a multi-task learning framework. Different from the original WaveNet model, the proposed Multi-task WaveNet employs the frame-level acoustic feature prediction as the secondary task and the external fundamental frequency prediction model for the original WaveNet can be removed. Therefore the improved WaveNet can generate high-quality speech waveforms only conditioned on linguistic features. Multi-task WaveNet can produce more natural and expressive speech by addressing the pitch prediction error accumulation issue and possesses more succinct inference procedures than the original WaveNet. Experimental results prove that the SPSS method proposed in this paper can achieve better performance than the state-of-the-art approach utilizing the original WaveNet in both objective and subjective preference tests.

## Speaker-independent Raw Waveform Model for Glottal Excitation

*Lauri Juvela[1], Vassilis Tsiaras[2], Bajibabu Bollepalli[1], Manu Airaksinen[1], Junichi Yamagishi[3] and Paavo Alku[1]*
[1]Aalto University, Finland
[2]University of Crete, Greece
[3]National Institute of Informatics, Japan

**Wed-P-1-1-10, Time: 10:00-12:00**

Recent speech technology research has seen a growing interest in using

**Notes**

WaveNets as statistical vocoders, i.e., generating speech waveforms from acoustic features. These models have been shown to improve the generated speech quality over classical vocoders in many tasks, such as text-to-speech synthesis and voice conversion. Furthermore, conditioning WaveNets with acoustic features allows sharing the waveform generator model across multiple speakers without additional speaker codes. However, multi-speaker WaveNet models require large amounts of training data and computation to cover the entire acoustic space. This paper proposes leveraging the source-filter model of speech production to more effectively train a speaker-independent waveform generator with limited resources. We present a multi-speaker 'GlotNet' vocoder, which utilizes a WaveNet to generate glottal excitation waveforms, which are then used to excite the corresponding vocal tract filter to produce speech. Listening tests show that the proposed model performs favourably to a direct WaveNet vocoder trained with the same model architecture and data.

### A New Glottal Neural Vocoder for Speech Synthesis

*Yang Cui, Xi Wang, Lei He and Frank K. Soong*

Microsoft AI & Research, Beijing, China

**Wed-P-1-1-11, Time: 10:00-12:00**

Direct modeling of waveform generation for speech synthesis, e.g. WaveNet, has made significant progress on improving the naturalness and clarity of TTS. Such deep neural network-based models can generate highly realistic speech but at high computational and memory costs. We propose here a novel neural glottal vocoder which tends to bridge the gap between the traditional parametric vocoder and end-to-end speech sample generation. In the analysis, speech signals are decomposed into corresponding glottal source signals and vocal tract filters by the glottal inverse filtering. Glottal pulses are parameterized into energy, DCT coefficients (shape) and phase. The phase trajectory of successive glottal pulses is rendered with a trainable weighting matrix to keep a smooth pitch synchronous phase trajectory. We design a hybrid, i.e., both feed-forward and recurrent, neural network to reconstruct the glottal waveform including the optimized weighting matrix. Speech is then synthesized by filtering the generated glottal waveform with the vocal tract filter. The new neural glottal vocoder can generate high-quality speech with efficient computations. Subjective tests show that it gets an MOS score of 4.12 and 75% preference over the conventional glottal vocoder with a perceived quality comparable to WaveNet and natural recording in analysis-by-synthesis.

### Exemplar-based Speech Waveform Generation

*Oliver Watts, Cassia Valentini-Botinhao, Felipe Espic and Simon King*

The Centre for Speech Technology Research, Edinburgh University, UK

**Wed-P-1-1-12, Time: 10:00-12:00**

This paper presents a simple but effective method for generating speech waveforms by selecting small units of stored speech to match a low-dimensional target representation. The method is designed as a drop-in replacement for the vocoder in a deep neural network-based text-to-speech system. Most previous work on hybrid unit selection waveform generation relies on phonetic annotation for determining unit boundaries, or for specifying target cost, or for candidate preselection. In contrast, our waveform generator requires no phonetic information, annotation, or alignment. Unit boundaries are determined by epochs and spectral analysis provides representations which are compared directly with target features at runtime. As in unit selection, we minimise a combination of target cost and join cost, but find that greedy left-to-right nearest-neighbour search gives similar results to dynamic programming. The method is fast and can generate the waveform incrementally. We use publicly available data and provide a permissively-licensed open source toolkit for reproducing our results.

### Frequency Domain Variants of Velvet Noise and Their Application to Speech Processing and Synthesis

*Hideki Kawahara[1], Ken-Ichi Sakakibara[2], Masanori Morise[3], Hideki Banno[4], Tomoki Toda[5] and Toshio Irino[1]*

[1]Wakayama University, Japan

[2]Health Science University of Hokkaido, Japan

[3]University of Yamanashi, Japan

[4]Meijo University, Japan

[5]Nagoya University, Japan

**Wed-P-1-1-13, Time: 10:00-12:00**

We propose a new excitation source signal for VOCODERs and an all-pass impulse response for post-processing of synthetic sounds and pre-processing of natural sounds for data-augmentation. The proposed signals are variants of velvet noise, which is a sparse discrete signal consisting of a few non-zero (1 or -1) elements and sounds smoother than Gaussian white noise. One of the proposed variants, FVN (Frequency domain Velvet Noise) applies the procedure to generate a velvet noise on the cyclic frequency domain of DFT (Discrete Fourier Transform). Then, by smoothing the generated signal to design the phase of an all-pass filter followed by inverse Fourier transform yields the proposed FVN. Temporally variable frequency weighted mixing of FVN generated by frozen and shuffled random number provides a unified excitation signal which can span from random noise to a repetitive pulse train. The other variant, which is an all-pass impulse response, significantly reduces "buzzy" impression of VOCODER output by filtering. Finally, we will discuss applications of the proposed signal for watermarking and psychoacoustic research.

## Wed-P-1-2: Extracting Information from Audio

Hall 4-6: Poster-2; 10:00-12:00; Wednesday, 5 September, 2018

Chair: Thiago Fraga da Silva

### Joint Learning of Interactive Spoken Content Retrieval and Trainable User Simulator

*Pei-Hung Chung[1], Kuan Tung[2], Ching-Lun Tai[2] and Hung-yi Lee[1]*

[1]Graduate Institute of Communication Engineering, National Taiwan University

[2]Department of Electrical Engineering, National Taiwan University

**Wed-P-1-2-1, Time: 10:00-12:00**

User-machine interaction is crucial for information retrieval, especially for spoken content retrieval, because spoken content is difficult to browse and speech recognition has a high degree of uncertainty. In interactive retrieval, the machine takes different actions to interact with the user to obtain better retrieval results; here it is critical to select the most efficient action. In previous work, deep Q-learning techniques were proposed to train an interactive retrieval system but rely on a hand-crafted user simulator; building a reliable user simulator is difficult. In this paper, we further improve the interactive spoken content retrieval framework by proposing a learnable user simulator which is jointly trained with interactive retrieval system, making the hand-crafted user simulator unnecessary. The experimental results show that the learned simulated users not only achieve larger rewards than the hand-crafted ones but act more like real users.

### Attention-based End-to-End Models for Small-Footprint Keyword Spotting

*Changhao Shan[1,2], Junbo Zhang[2], Yujun Wang[2] and Lei Xie[1]*

[1]Shaanxi Provincial Key Laboratory of Speech and Image Information Processing, School of Computer Science, Northwestern Polytechnical University, Xi'an, China

[2]Xiaomi Inc., Beijing, China

**Wed-P-1-2-2, Time: 10:00-12:00**

In this paper, we propose an attention-based end-to-end neural approach for small-footprint keyword spotting (KWS), which aims to simplify the pipelines of building a production-quality KWS system. Our model consists of an encoder and an attention mechanism. The encoder transforms the input signal into a high level representation using RNNs. Then the attention

**Notes**

mechanism weights the encoder features and generates a fixed-length vector. Finally, by linear transformation and softmax function, the vector becomes a score used for keyword detection. We also evaluate the performance of different encoder architectures, including LSTM, GRU and CRNN. Experiments on real-world wake-up data show that our approach outperforms the recent Deep KWS approach by a large margin and the best performance is achieved by CRNN. To be more specific, with ~84K parameters, our attention-based model achieves 1.02% false rejection rate (FRR) at 1.0 false alarm (FA) per hour.

## Prediction of Aesthetic Elements in Karnatic Music: A Machine Learning Approach

*Ragesh Rajan M[1], Ashwin Vijayakumar[2] and Deepu Vijayasenan[1]*

[1]National Institute of Technology Karnataka, India

[2]Georgia Institute of Technology, Atlanta, USA

**Wed-P-1-2-3, Time: 10:00-12:00**

Gamakas, the embellishments and ornamentations used to enhance musical experience, are defining features of Karnatic Music (KM). The appropriateness of using gamaka is determined by aesthetics and is often developed by musicians with experience. Therefore, understanding and modeling gamaka is a significant bottleneck in applications like music synthesis, automatic accompaniment, etc. in the context of KM. To this end, we propose to learn both the presence and the type of gamaka in a data-driven manner using annotated symbolic music. In particular, we explore the efficacy of three classes of features – note-based, phonetic and structural – and train a Random Forest

Classifier to predict the existence and the type of gamaka. The observed accuracy is ~70% for gamaka detection and ~60% for gamaka classification. Finally, we present an analysis of the features and find that frequency and duration of the neighbouring notes prove to be the most important features.

## Topic and Keyword Identification for Low-resourced Speech Using Cross-Language Transfer Learning

*Wenda Chen[1,2], Mark Hasegawa-Johnson[1] and Nancy F. Chen[2]*

[1]Beckman Institute, University of Illinois at Urbana-Champaign, USA

[2]Institute for Infocomm Research, A*STAR, Singapore

**Wed-P-1-2-4, Time: 10:00-12:00**

This paper studies topic and keyword identification for languages in which we have no transcribed speech data. We adopt a transfer learning framework to transfer what is learned from rich-resourced languages (RRL) to low-resourced languages (LRL). Specifically, we propose that a convolutional neural network (CNN) trained as a topic classifier in an RRL learns features (hidden layer activations) that can be used for the same purpose in an LRL. The CNN observes acoustic features, RRL phones, or segment clusters generated by an unsupervised phone clustering system; its hidden layers are retained and its output layer re-trained from scratch on the LRL. Our results are compared with the state-of-the-art topic classification methods on cross-language ASR transcripts. We also discuss the successful detection of topic dependent keywords and the use of unsupervised learning based clusters in our approach for low-resourced language topic detection.

## Automatic Speech Recognition and Topic Identification from Speech for Almost-Zero-Resource Languages

*Matthew Wiesner[1,2], Chunxi Liu[1,2], Lucas Ondel[3], Craig Harman[1], Vimal Manohar[1,2], Jan Trmal[1,2], Zhongqiang Huang[4], Najim Dehak[1] and Sanjeev Khudanpur[1,2]*

[1]Center for Language and Speech Processing, The Johns Hopkins University, USA

[2]Human Language Technology Center of Excellence, The Johns Hopkins University, USA

[3]Brno University of Technology, Czech Republic

[4]Raytheon BBN Technologies, USA

**Wed-P-1-2-5, Time: 10:00-12:00**

Automatic speech recognition (ASR) systems often need to be developed for extremely low-resource languages to serve end-uses such as audio content categorization and search. While universal phone recognition is natural to consider when no transcribed speech is available to train an ASR system in a language, adapting universal phone models using very small amounts (minutes rather than hours) of transcribed speech also needs to be studied, particularly with state-of-the-art DNN-based acoustic models. The DARPA LORELEI program provides a framework for such very-low-resource ASR studies and provides an extrinsic metric for evaluating ASR performance in a humanitarian assistance, disaster relief setting. This paper presents our Kaldi-based systems for the program, which employ a universal phone modeling approach to ASR and describes recipes for very rapid adaptation of this universal ASR system. The results we obtain significantly outperform results obtained by many competing approaches on the NIST LoReHLT 2017 Evaluation datasets.

## Play Duration Based User-Entity Affinity Modeling in Spoken Dialog System

*Bo Xiao[1], Nicholas Monath[2], Shankar Ananthakrishnan[1] and Abishek Ravi[1]*

[1]Amazon.com, Cambridge, MA, U.S.A.

[2]CICS, University of Massachusetts Amherst, Amherst, MA, U.S.A

**Wed-P-1-2-6, Time: 10:00-12:00**

Multimedia streaming services over spoken dialog systems have become ubiquitous. User-entity affinity modeling is critical for the system to understand and disambiguate user intents and personalize user experiences. However, fully voice-based interaction demands quantification of novel behavioral cues to determine user affinities. In this work, we propose using play duration cues to learn a matrix factorization based collaborative filtering model. We first binarize play durations to obtain implicit positive and negative affinity labels. The Bayesian Personalized Ranking objective and learning algorithm are employed in our low-rank matrix factorization approach. To cope with uncertainties in the implicit affinity labels, we propose to apply a weighting function that emphasizes the importance of high confidence samples. Based on a large-scale database of Alexa music service records, we evaluate the affinity models by computing Spearman correlation between play durations and predicted affinities. Comparing different data utilizations and weighting functions, we find that employing both positive and negative affinity samples with a convex weighting function yields the best performance. Further analysis demonstrates the model's effectiveness on individual entity level and provides insights on the temporal dynamics of observed affinities.

## Empirical Analysis of Score Fusion Application to Combined Neural Networks for Open Vocabulary Spoken Term Detection

*Shi-wook Lee[1], Kazuyo Tanaka[2] and Yoshiaki Itoh[3]*

[1]National Institute of Advanced Industrial Science and Technology, Japan

[2]University of Tsukuba, Japan

[3]Iwate Prefectural University, Japan

**Wed-P-1-2-7, Time: 10:00-12:00**

System combination, which combines the outputs of multiple systems or internal representations, is a powerful method to improve the performance of machine learning tasks and has been widely adopted in recent knowledge transfer learning. In this study, to describe how to extract effective knowledge from an ensemble of neural networks, we first examine several score fusions from an ensemble of neural networks tasked with open vocabulary spoken term detection, where the class probability of the neural network is utilized as a similarity metric; then, we investigate the trade-off between confusion and dark knowledge. From the experimental evaluation on open vocabulary spoken term detection, we obtain 2.09% absolute gain as compared to the best result from single systems. Furthermore, the performance gains achieved via score fusion of class probabilities exactly match the mathematical inequality for sum and power means

**Notes**

results and that the gain achieved via summation of class probabilities is consistently better than that achieved via score fusion of power means. The experimental analysis confirms that summation, which enhances the discriminative capability of the superior class probability, can implement smoothed probability distribution to yield more effective dark knowledge, while adequately suppressing undesirable effects.

## Phonological Posterior Hashing for Query by Example Spoken Term Detection

*Afsaneh Asaei[1], Dhananjay Ram[1,2] and Hervé Bourlard[1,2]*

[1]Idiap Research Institute, Martigny, Switzerland

[2]École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

**Wed-P-1-2-8, Time: 10:00-12:00**

State of the art query by example spoken term detection (QbE-STD) systems in zero-resource conditions rely on representation of speech in terms of sequences of class-conditional posterior probabilities estimated by deep neural network (DNN). The posteriors are often used for pattern matching or dynamic time warping (DTW). Exploiting posterior probabilities as speech representation propounds diverse advantages in a classification system. One key property of the posterior representations is that they admit a highly effective hashing strategy that enables indexing a large audio archive in divisions for reducing the search complexity. Moreover, posterior indexing leads to a compressed representation and enables pronunciation dewarping and partial detection with no need for DTW. We exploit these characteristics of the posterior space in the context of redundant hash addressing for query-by-example spoken term detection (QbE-STD). We evaluate the QbE-STD system on AMI corpus and demonstrate that tremendous speedup and superior accuracy is achieved compared to the state-of-the-art pattern matching solution based on DTW. The system has the potential to enable massively large scale spoken query detection.

## Term Extraction via Neural Sequence Labeling a Comparative Evaluation of Strategies Using Recurrent Neural Networks

*Maren Kucza, Jan Niehues, Thomas Zenkel, Alex Waibel and Sebastian Stüker*

Karlsruhe Institute of Technology

**Wed-P-1-2-9, Time: 10:00-12:00**

Traditionally systems for term extraction use a two stage approach of first identifying candiate terms and the scoring them in a second process for identifying actual terms. Thus, research in this field has often mainly focused on refining and improving the scoring process of term candidates, which commonly are identified using linguistic and statistical features. Machine learning techniques and especially neural networks are currently only used in the second stage, that is to score candidates and classify them.

In contrast to that we have built a system that identifies terms via directly performing sequence-labeling with a BILOU scheme on word sequences. To do so we have worked with different kinds of recurrent neural networks and word embeddings.

In this paper we describe how one can built a state-of-the-art term extraction systems with this single-stage technique and compare different network types and topologies and also examine the influence of the type of input embedding used for the task. We further investigated which network types and topologies are best suited when applying our term extraction systems to other domains than that of the training data of the networks.

## Semi-supervised Learning for Information Extraction from Dialogue

*Anjuli Kannan, Kai Chen, Diana Jaunzeikare and Alvin Rajkomar*

Google Brain, Google, Inc., USA

**Wed-P-1-2-10, Time: 10:00-12:00**

In this work we present a method for semi-supervised learning from transcripts of dialogue between humans. We consider the scenario in which a large amount of transcripts are available and we would like to extract some semantic information from them; however, only a small number of transcripts have been labeled with this information. We present a method for leveraging the unlabeled data to learn a better model than could be learned from the labeled data alone. First, a recurrent neural network (RNN) encoder-decoder is trained on the task of predicting nearby turns on the full dialogue corpus; next, the RNN encoder is reused as a feature representation for the supervised learning problem.

While previous work has explored the use of pre-training for non-dialogue corpora, our method is specifically geared toward the dialogue use case. We demonstrate an improvement on a clinical documentation task, particularly in the regime of small amounts of labeled data. We compare several types of encoders, both in the context of a classification task and in a human-evaluation of their learned representations. We show that our method significantly improves the classification task in the case where only a small amount of labeled data is available.

## Slot Filling with Delexicalized Sentence Generation

*Youhyun Shin, Kang Min Yoo and Sang-goo Lee*

Department of Computer Science and Engineering, Seoul National University, Seoul, Korea

**Wed-P-1-2-11, Time: 10:00-12:00**

We introduce a novel approach that jointly learns slot filling and delexicalized sentence generation. There have been recent attempts to tackle slot filling as a type of sequence labeling problem, with encoder-decoder attention framework. We further improve the framework by training the model to generate delexicalized sentences, in which words according to slot values are replaced with slot labels. Slot filling with delexicalization shows better results compared to models having a single learning objective of filling slots. The proposed method achieves state-of-the-art slot filling performance on ATIS dataset. We experiment different variants of our model and find that delexicalization encourages generalization by sharing weights among the words with same labels and helps the model to further leverage certain linguistic features.

## Music Genre Recognition Using Deep Neural Networks and Transfer Learning

*Deepanway Ghosal and Maheshkumar H. Kolekar*

Indian Institute of Technology Patna, India

**Wed-P-1-2-12, Time: 10:00-12:00**

Music genre recognition is a very interesting area of research in the broad scope of music information retrieval and audio signal processing. In this work we propose a novel approach for music genre recognition using an ensemble of convolutional long short term memory based neural networks (CNN LSTM) and a transfer learning model. The neural network models are trained on a diverse set of spectral and rhythmic features whereas the transfer learning model was originally trained on the task of music tagging. We compare our system with a number of recently published works and show that our model outperforms them and achieves new state of the art results.

## Efficient Voice Trigger Detection for Low Resource Hardware

*Siddharth Sigtia, Rob Haynes, Hywel Richards, Erik Marchi and John Bridle*

Siri Speech, Apple

**Wed-P-1-2-13, Time: 10:00-12:00**

We describe the architecture of an always-on keyword spotting (KWS) system for battery-powered mobile devices used to initiate an interaction with the device. An always-available voice assistant needs a carefully designed voice keyword detector to satisfy the power and computational constraints of battery powered devices. We employ a multi-stage system that uses a low-power primary stage to decide when to run a more accurate (but more power-hungry) secondary detector. We describe a straightforward primary detector and explore variations that result in very useful reductions in computation (or increased accuracy for the same computation). By reducing the set of target labels from three to one per phone and reducing the rate at which the acoustic model is operated, the compute rate can be reduced by a factor of six while maintaining the same accuracy.

**Notes**

# Wed-P-1-3: Signal Analysis for the Natural, Biological and Social Sciences

Hall 4-6: Poster-3; 10:00-12:00; Wednesday, 5 September, 2018

Chair: V Ramasubramanian

## A Novel Normalization Method for Autocorrelation Function for Pitch Detection and for Speech Activity Detection

*Qiguang Lin[1,2] and Yiwen Shao[1,3]*

[1]Baihu Technology, Co., Ltd, China

[2]ZIBOT, Inc., Santa Clara, California, USA

[3]Dept. Computer Science, Johns Hopkins University, USA

**Wed-P-1-3-1, Time: 10:00-12:00**

Autocorrelation functions (ACF) have been used in various pitch detection algorithms (PDA) and voicing-feature based speech activity detection (SAD) techniques. Speech is assumed to be stationary over a short-term window and a Hanning window is typically applied in the calculation of ACF. As a result of windowing, the ACF tapers as the autocorrelation lags increase. Boersma demonstrated that the tapering effect could be compensated for by dividing the ACF of the windowed signal by the autocorrelation of the windowing function itself, referred to as wACF hereafter. We recently found that wACF could cause overcompensation and therefore, result in errors in pitch detection. In this paper, a novel normalization method, eACF, is proposed that can both mitigate the tapering effect and minimize the overcompensation. The new method is evaluated on synthetic speech and on the TIMIT database with various types of additive noise at different signal-to-noise (SNR) ratios. The results show that the new method leads to better performance both in terms of pitch detection and speech activity detection. In this paper, we also investigate the scenarios where applying the wACF method is advantageous and where it is not.

## Estimation of the Vocal Tract Length of Vowel Sounds Based on the Frequency of the Significant Spectral Valley

*TV Ananthapadmanabha[1] and Ramakrishnan A G[2]*

*[1]Voice and Speech Systems, Malleswaram, Bangalore 560003, India*

*[2]Department of Electrical Engineering, Indian Institute of Science, Bangalore 560012, India*

**Wed-P-1-3-2, Time: 10:00-12:00**

Estimating the vocal tract length (VTL), given the acoustic signal of a vowel sound, is an important problem, which is useful in speaker normalization for vowel recognition, in the inversion problem and in acoustic-phonetic studies. The common approach of using the formant data to estimate VTL works for a neutral vowel approximating a uniform tube. However, for natural vowels, formant data shift considerably away from the resonant frequencies of a uniform tube. The proposed method is motivated from these observations: (a) the frequency of a spectral valley, F_v, depends inversely on VTL; (b) there is much smaller shift in F_v, across vowels, from the corresponding valley frequency of a uniform tube; (c) F_v can be estimated from the spectral envelope itself. VTL has been estimated for the Peterson and Barney (33 male and 28 female speakers) and the TIMIT (326 male and 136 female speakers) databases. When the estimated F_v is used for normalization, the spread in the formant data due to gender differences is considerably reduced. The normalization procedure is vowel and speaker intrinsic. Additionally, we report applications such as Front/Back classification, gender recognition and phonetic feature mapping.

## Deep Learning Techniques for Koala Activity Detection

*Ivan Himawan[1], Michael Towsey[1], Bradley Law[2] and Paul Roe[1]*

[1]Queensland University of Technology, Brisbane, Australia

[2]Forest Science Unit, NSW Department of Industry-Lands, Australia

**Wed-P-1-3-3, Time: 10:00-12:00**

Automatically detecting koalas in the real-life environment from audio recordings will immensely help ecologists, conservation groups and government departments interested in their preservation and the protection of their habitat. Inspired by the success of deep learning approaches in various audio classification tasks, in this paper, the feasibility of recognizing koalas' calls using a convolutional recurrent neural network architecture (CNN+RNN) is studied. The benefit of this architecture is twofold: firstly, convolutional layers learn local time-frequency patterns from the audio spectrogram and secondly, recurrent layers model longer temporal dependencies of the extracted features. In our datasets, the performance of CNN+RNN is evaluated and compared with standard convolutional neural networks (CNNs). The experimental results show that hybrid CNN+RNN architecture is beneficial for learning long-term patterns in spectrogram exhibited by koalas' calls in unseen conditions. The proposed method is also applicable for detecting other animal calls such as bird sound where it achieves 87.46% area under curve score on the bird audio detection challenge evaluation data.

## Glottal Closure Instant Detection from Speech Signal Using Voting Classifier and Recursive Feature Elimination

*Jindřich Matoušek[1,2] and Daniel Tihelka[2]*

[1]Department of Cybernetics, Faculty of Applied Sciences, University of West Bohemia, Pilsen, Czech Rep.

[2]New Technology for the Information Society (NTIS), Faculty of Applied Sciences, University of West Bohemia, Pilsen, Czech Rep.

**Wed-P-1-3-4, Time: 10:00-12:00**

In our previous work, we introduced a classification-based method for the automatic detection of glottal closure instants (GCIs) from the speech signal and we showed it was able to perform very well on several test datasets. In this paper, we investigate whether adding more features (voiced/unvoiced, harmonic/noise, spectral etc.) and/or using an ensemble of classifiers such as a voting classifier can further improve GCI detection performance. We show that using additional features leads to a better detection accuracy; best results were obtained when recursive feature elimination was applied on the whole feature set. In addition, a voting classifier is shown to outperform other classifiers and other existing GCI detection algorithms on publicly available databases.

## Assessing Speaker Engagement in 2-Person Debates: Overlap Detection in United States Presidential Debates

*Midia Yousefi[1], Navid Shokouhi[2] and John H.L. Hansen[1]*

[1]Center for Robust Speech Systems(CRSS), The University of Texas at Dallas, Richardson, Texas, USA

[2]University of Melbourne

**Wed-P-1-3-5, Time: 10:00-12:00**

Co-channel speech contain significant amounts of overlap in which the intelligibility and quality of the desired speech can be degraded. Convolutive Non-negative Matrix Factorization (CNMF) has been shown to be a successful approach in detecting overlap by extracting specific acoustic basis dimensions for each speaker from an audio stream. While the results of CNMF have been successful, it requires isolated single speech recordings for each speaker to derive their corresponding bases functions/dimensions. In our previous work, Teager-Kaiser Energy Operator (TEO)-based Pyknogram has been introduced. In this study, Pyknogram and CNMF based solutions for overlap detection within audio streams have been examined using the GRID dataset. TEO-based Pyknogram is shown to achieve a relative 8-10% lower Equal Error Rate (EER) compared to CNMF features. In addition, a secondary evaluation was also performed based on naturalistic audio streams with overlap. Specifically, we collected a real-world audio database of US Presidential debates stemming from the last 12 years that are very challenging due to various forms of overlaps, changing Signal to Interference Ratio (SIR) and, environmental noise among

**Notes**

others. Our experiments indicate that TEO-based Pyknogram is well suited for detecting overlap in challenging real world scenarios such as the US presidential debates.

## All-Conv Net for Bird Activity Detection: Significance of Learned Pooling

*Arjun Pankajakshan, Anshul Thakur, Daksh Thapar, Padmanabhan Rajan and Aditya Nigam*

School of Computing and Electrical Engineering, Indian Institute of Technology, Mandi

**Wed-P-1-3-6, Time: 10:00-12:00**

Bird activity detection (BAD) deals with the task of predicting the presence or absence of bird vocalizations in a given audio recording. In this paper, we propose an all-convolutional neural network (all-conv net) for bird activity detection. All the layers of this network including pooling and dense layers are implemented using convolution operations. The pooling operation implemented by convolution is termed as learned pooling. This learned pooling takes into account the inter feature-map correlations which are ignored in traditional max-pooling. This helps in learning a pooling function which aggregates the complementary information in various feature maps, leading to better bird activity detection. Experimental observations confirm this hypothesis. The performance of the proposed all-conv net is evaluated on the BAD Challenge 2017 dataset. The proposed all-conv net achieves state-of-art performance with a simple architecture and does not employ any data pre-processing or data augmentation techniques.

## Deep Convex Representations: Feature Representations for Bioacoustics Classification

*Anshul Thakur[1], Vinayak Abrol[2], Pulkit Sharma[1] and Padmanabhan Rajan[1]*

[1]IIT Mandi, India

[2]Idiap Research Institute, Martigny, Switzerland

**Wed-P-1-3-7, Time: 10:00-12:00**

In this paper, a deep convex matrix factorization framework is proposed for bioacoustics classification. Archetypal analysis, a form of convex non-negative matrix factorization, is used for acoustic modelling at each level of this framework. At first level, the input feature matrix is factorized into an archetypal dictionary and corresponding convex representations. The representation matrix obtained at the first level is further factorized into a dictionary and convex representations at second level. This hierarchical factorization continues until a desired depth is achieved. We observe that the dictionaries at different levels model complimentary information present in the data. The atoms of the dictionary learned at the first layer lie on convex hull of the data, thus try to model the extremal behaviour. On the contrary, atoms of the deeper dictionaries lie on the convex hull as well as inside the convex hull. Hence, these dictionaries can simultaneously model the extremal and average behavior of the data. The convex representations obtained from these deeper dictionaries are referred as deep convex representations. Due to inherent sparsity, they result in efficient classification performance. Through experiments on two available bioacoustics datasets, we show that the proposed approach yield comparable or better results than state-of-art approaches.

## Detection of Glottal Excitation Epochs in Speech Signal Using Hilbert Envelope

*Hirak Dasgupta, Prem C. Pandey and K S Nataraj*

Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, India

**Wed-P-1-3-8, Time: 10:00-12:00**

A technique, suitable for real-time processing, is presented for detection of glottal excitation epochs in voiced speech. It uses Hilbert envelope to enhance saliency of the glottal excitation epochs and to reduce the ripples due to the vocal tract filter. The processing comprises the steps of dynamic range compression, calculation of the Hilbert envelope and epoch marking. The first step reduces amplitude variation by applying A-law on the signal envelope. The second step calculates the Hilbert envelope using the output of an FIR filter-based Hilbert transformer and the delay-compensated signal. The third step uses a dynamic peak detector with fast rise and slow fall and nonlinear smoothing using a two-step median-mean filter to further enhance the saliency of the epochs, followed by a differentiator to mark them. The technique is tested using the CMU-ARCTIC database with simultaneously recorded speech and EGG signals. The results showed a good match in the performance of the proposed technique with those of the state-of-the-art techniques and its robustness against highpass filtering. It may be useful for diagnosis of voice disorders and high-quality voice conversion.

## Analyzing Thai Tone Distribution through Functional Data Analysis

*Hong Zhang*

University of Pennsylvania

**Wed-P-1-3-9, Time: 10:00-12:00**

This paper reports an analysis of tonal properties of Thai using a method based on functional data analysis on a large collection of TIMIT-like corpus. Both density estimation pooled across syllable-wise F0 contours and Functional Principle Component Analysis (FPCA) were applied. The results suggest that the simple two dimensional representation of tones: pitch target height and contour slope, is not able to capture context dependent variations of tonal contour within and across tone categories. In addition, the shape and timing of pitch target are also crucial both in differentiating tonal categories and explaining variations associated with syllable structure. The third and fourth dimension of the functional basis have been shown to be able to represent these higher-order properties. Thus FPCA can provide a compact yet interpretable low dimension representation for the tonal property of Thai. These findings are also helpful for understanding tone distribution properties and coarticulation.

## Articulatory Feature Classification Using Convolutional Neural Networks

*Danny Merkx and Odette Scharenborg*

Centre for Language Studies, Radboud University, Nijmegen, The Netherlands

**Wed-P-1-3-10, Time: 10:00-12:00**

The ultimate goal of our research is to improve an existing speech-based computational model of human speech recognition on the task of simulating the role of fine-grained phonetic information in human speech processing. As part of this work we are investigating articulatory feature classifiers that are able to create reliable and accurate transcriptions of the articulatory behaviour encoded in the acoustic speech signal. Articulatory feature (AF) modelling of speech has received a considerable amount of attention in automatic speech recognition research. Different approaches have been used to build AF classifiers, most notably multi-layer perceptrons. Recently, deep neural networks have been applied to the task of AF classification. This paper aims to improve AF classification by investigating two different approaches: 1) investigating the usefulness of a deep Convolutional neural network (CNN) for AF classification; 2) integrating the Mel filtering operation into the CNN architecture. The results showed a remarkable improvement in classification accuracy of the CNNs over state-of-the-art AF classification results for Dutch, most notably in the minority classes. Integrating the Mel filtering operation into the CNN architecture did not further improve classification performance.

## A New Frequency Coverage Metric and a New Subband Encoding Model, with an Application in Pitch Estimation

*Shoufeng Lin*

School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, Kent Street, Bentley, Western Australia, 6102

**Wed-P-1-3-11, Time: 10:00-12:00**

The auditory filterbank has been a well-accepted and important tool for speech feature extraction. It decomposes the speech signal into subbands usually on an equivalent rectangular bandwidth frequency scale before further subband analysis and processing, such as auto-correlation and cross-correlation. However, the choice of the number of subbands and

**Notes**

subband center frequencies for a given frequency range has been essentially empirical in the literature. Moreover, correlation of subband signals may not produce distinct peaks for feature extraction. This paper proposes a novel frequency coverage metric to calculate the required number of subbands. It also presents a new subband encoding model for correlation processing, inspired by psychoacoustic studies and statistical analysis. The proposed frequency coverage metric and the subband encoding model are applied to a pitch estimation method as an example of their possible implementations in the speech feature extraction. Compared with state-of-the-art methods, evaluation results demonstrate the benefits of the proposed methods.

### Improved Epoch Extraction from Telephonic Speech Using Chebfun and Zero Frequency Filtering

*Ganga Gowri B, Soman K.P and Govind D*

Center for Computational Engineering and Networking (CEN), Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, India

**Wed-P-1-3-12, Time: 10:00-12:00**

Epoch in speech, represent the instant where maximum excitation at the vocal tract is obtained. Existing epoch extraction algorithms are capable of accurately extracting epoch information from clean speech signals. However, epoch extraction of band limited signals such as telephonic speech is challenging due to the attenuation of the fundamental frequency components. The present work is focused on improving the performance of epoch extraction from telephonic speech signals by exploiting the properties of Chebyshev polynomial interpolation and by reinforcing the frequency components around the fundamental frequency through the Hilbert envelope (HE). The proposed algorithm brings a refinement of the existing Zero Frequency Filtering (ZFF) method by incorporating Chebyshev interpolation. The proposed refinements to the ZFF algorithm confirmed to provide improved epoch identification rate, identification accuracy, reduced miss rate and false alarm rate. The epoch identification rate of the proposed method is observed to be better than existing methods like Dynamic Programming Phase Slope Algorithm (DYPSA), Speech Event Detection using the Residual Excitation And a Mean-based Signal (SEDREAMS), Dynamic Plosion Index (DPI) and Single Pole Filtering (SPF) methods for telephonic speech quality.

## Wed-P-1-4: Speech Prosody

Hall 4-6: Poster-4; 10:00-12:00; Wednesday, 5 September, 2018

Chair: Volker Dellwo

### An Empirical Analysis of the Correlation of Syntax and Prosody

*Arne Köhn[1], Timo Baumann[2] and Oskar Dörfler[1]*

[1]Natural Language Systems Group, Department of Informatics, Universität Hamburg, Germany

[2]Language Technology Institute, Carnegie Mellon University, Pittsburgh, USA

**Wed-P-1-4-1, Time: 10:00-12:00**

The relation of syntax and prosody (the syntax-prosody interface) has been an active area of research, mostly in linguistics and typically studied under controlled conditions. More recently, prosody has also been successfully used in the data-based training of syntax parsers. However, there is a gap between the controlled and detailed study of the individual effects between syntax and prosody and the large-scale application of prosody in syntactic parsing with only a shallow analysis of the respective influences. In this paper, we close the gap by investigating the significance of correlations of prosodic realization with specific syntactic functions using linear mixed effects models in a very large corpus of read-out German encyclopedic texts. Using this corpus, we are able to analyze prosodic structuring performed by a diverse set of speakers while they try to optimize factual content delivery. After normalization by speaker, we obtain significant effects, e.g. confirming that the subject function, as compared to the object function, has a positive effect on pitch and duration of a word, but a negative effect on loudness.

### Analysing the Focus of a Hierarchical Attention Network: the Importance of Enjambments When Classifying Post-modern Poetry

*Timo Baumann[1], Hussein Hussein[2] and Burkhard Meyer-Sickendiek[2]*

[1]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA

[2]Department of Literary Studies, Free University of Berlin, Germany

**Wed-P-1-4-2, Time: 10:00-12:00**

After overcoming the traditional metrics, modern and postmodern poetry developed a large variety of 'free verse prosodies' that falls along a spectrum from a more fluent to a more disfluent and choppy style. We present a method, grounded in philological analysis and theories on cognitive (dis) fluency, to analyze this 'free verse spectrum' into six classes of poetic styles as well as to differentiate three types of poems with enjambments. We use a model for automatic prosodic analysis of spoken free verse poetry which uses deep hierarchical attention networks to integrate the source text and audio and predict the assigned class. We then analyze and fine-tune the model with a particular focus on enjambments and in two ways: we drill down on classification performance by analyzing whether the model focuses on similar traits of poems as humans would, specifically, whether it internally builds a notion of enjambment. We find that our model is similarly good as humans in finding enjambments; however, when we employ the model for classifying enjambment-dominated poem types, it does not pay particular attention to those lines. Adding enjambment labels to the training only marginally improves performance, indicating that all other lines are similarly informative for the model.

### Language-Dependent Melody Embeddings

*Daniil Kocharov and Alla Menshikova*

Saint Petersburg State University, Russia

**Wed-P-1-4-3, Time: 10:00-12:00**

The paper explores the perspectives of applying the distributional approach to prosodic typology of languages. The method discussed here is an adaptation of the distributional semantics approach, as suggested by Mikolov, to melodic features of speech. The paper contains a detailed description of the new method, as well as a comparison of five European languages (English, Czech, German, Russian and Finnish) in terms of melody embeddings. The total amount of speech data was over 500 hours.

The experimental results show that melody embeddings are language dependent. The proposed melody embedding model has shown reasonable results in language comparison.

### Stress Distribution of Given Information in Chinese Reading Texts

*Yuan Jia[1] and Xiaoxiao Ma[1,2]*

[1] Institute of Linguistic, Chinese Academy of Social Science, Beijing, China

[2] Nankai University, Tianjin, China

**Wed-P-1-4-4, Time: 10:00-12:00**

Using an information structure annotation System, namely RefLex Scheme, the present study annotates the information structure of Chinese reading discourse and explores the relationship between information status and stress distribution. Our analysis results show that given information could bear stresses as well as new information. Specifically, the stress distribution of given information is significantly affected by the sub-category of information status on the referential level, i.e., r-given, while r-given-generic and r-given-displaced show different stress distributions. However, the sub-category of information status on the lexical level

**Notes**

exhibits no such effect. Besides, the given information of proper nouns and personal pronouns on the lexical level can attract stresses. The reason is that a proper noun often serves as the topic of a sentence and a personal pronoun usually processes a center shift. Furthermore, the inconsistency of information status on both referential and lexical levels causes the stress on the given information unit.

## Acoustic-prosodic Entrainment in Structural Metadata Events

*Vera Cabarrão[1,2], Fernando Batista[1,3], Helena Moniz[1,2], Isabel Trancoso[1,4] and Ana Isabel Mata[2]*

[1]INESC-ID Lisboa, Laboratório de Sistemas de Língua Falada, Portugal

[2]FLUL - Faculdade de Letras da Universidade de Lisboa; CLUL - Centro de Linguística da, Universidade de Lisboa, Portugal

[3]Instituto Universitário de Lisboa (ISCTE-IUL), Lisboa, Portugal

[4]Instituto Superior Técnico, Universidade de Lisboa, Portugal

**Wed-P-1-4-5, Time: 10:00-12:00**

This paper presents an acoustic-prosodic analysis of entrainment in a Portuguese map-task corpus. Our aim is to analyze how turn-by-turn entrainment varies with distinct structural metadata events: types of sentence-like units (SU) in consecutive turns (e.g. interrogatives followed by declaratives, or both declaratives) and with the presence of discourse markers, affirmative cue words and disfluencies in the beginning of turns. Entrainment at turn-exchanges may be observed in terms of pitch, energy, duration and voice quality. Regarding SU types, question-answer turns are the ones with stronger similarity and declarative-interrogative pairs are the ones where less entrainment occurs, as expected. Moreover, in question-answer pairs, there is also stronger evidence of entrainment with Yes/No and Tag questions than with Wh-questions. In fact, these subtypes are coded in distinctive prosodic ways (moreover, the first subtype has no associated lexical-syntactic cues in Portuguese, only prosodic). As for turn-initial structures, entrainment is stronger when the second turn begins with an affirmative cue word; less strong with ambiguous structures (such as 'OK'), emphatic affirmative answers and negative answers; and scarce with disfluencies and discourse markers. The different degrees of local entrainment may be related with the informative structure of distinct structural metadata events.

## Formant Measures of Vowels Adjacent to Alveolar and Retroflex Consonants in Arrernte: Stressed and Unstressed Position

*Marija Tabain[1], Richard Beare[2,3] and Andrew Butcher[4]*

[1]Department of Languages and Linguistics, La Trobe University, Melbourne, Australia

[2]Department of Medicine, Monash University, Melbourne, Australia

[3]Murdoch Children's Research Institute, Melbourne, Australia

[4]Flinders University, Adelaide, Australia

**Wed-P-1-4-6, Time: 10:00-12:00**

This study presents formant data for six speakers of Arrernte, a language of central Australia. The focus of the study is the (marginal) phonemic contrast between two sets of apical consonants: alveolar and retroflex. The apical contrast is studied for the stop, nasal and lateral manners of articulation: /t ʈ/, /n ɳ/ and /l ɭ/. The apical consonants are examined both in strong prosodic context (preceding a stressed vowel) and in weak prosodic context (preceding an unstressed vowel). Formant data are sampled 10 ms before the onset of the consonant and 10 ms after the offset of the consonant. Results show no differences in F2 or F4 in the various conditions studied and results for F1 show differences between obstruents and sonorants. F3 is lower at consonant onset than consonant offset for retroflex stops in the weak prosodic context and to a lesser extent for retroflex stops in the strong prosodic context; it is also lower for laterals in the weak prosodic context. Other effects on F3 suggest that the apical contrast is most clearly realized for the stop manner of articulation.

## Automatic Assessment of L2 English Word Prosody Using Weighted Distances of F0 and Intensity Contours

*Quy-Thao Truong, Tsuneo Kato and Seiichi Yamamoto*

Graduate School of Science and Engineering, Doshisha University, Kyoto, Japan

**Wed-P-1-4-7, Time: 10:00-12:00**

In the current paper, an automatic prosody assessment method for learners of English using a weighted comparison of fundamental frequency (F0) and intensity contours is proposed. Patterns of F0 and intensity of learners are compared to that of native using a proposed metric - a weighted distance - in which the error around the high values of prosodic features have more weight in the computation of the final distance. Gold-standard native references are built using the k-means clustering algorithm. Therefore, we also propose a data-driven criterion called weighted variance based on the weighted similarity within the whole set of native utterances to determine the optimal number of clusters k. In comparison with baseline contour comparison metrics which resulted in a subjective-objective score correlation of 0.278, our method combining the proposed metric and criterion led to a final subjective-objective score correlation of 0.304. In comparison, subjective scores correlated at 0.480.

## Homogeneity vs Heterogeneity in Indian English: Investigating Influences of L1 on f0 Range

*Olga Maxwell[1], Elinor Payne[2] and Rosey Billington[1,3]*

[1]University of Melbourne, Australia

[2]University of Oxford, UK

[3]ARC Centre of Excellence for the Dynamics of Language, Australia

**Wed-P-1-4-8, Time: 10:00-12:00**

We present an exploratory analysis of several long-term distributional measures of f0 range in the speech of university-educated speakers of Indian English from four L1 backgrounds (Telugu, Tamil, Hindi and Bengali). The aim of this study is to investigate the degree of homogeneity in Indian English prosody and any similarities between the speakers' productions in English and their L1. Following recent studies, we examine three aspects of f0 range: pitch level (relative height of habitual f0), pitch span and pitch dynamism. Overall, across varieties, pitch level measures reveal individual speaker differences and only weak L1 effects on max f0 and median f0. Some speakers show higher f0 in their L1 productions compared to their English productions. More robust patterns were found for pitch span and dynamism: for all measures (maximum-minimum f0, pitch dynamism quotient and standard deviation), significant differences were found between L1 and English ($p<0.001$) for Bengali and Telugu L1 speakers. The relative weakness of L1 effects would suggest a degree of homogeneity in Indian English, at least for the prosodic parameters investigated. Evidence of a shift in pitch span when talking in English, regardless of L1, further suggests a convergent speech variety.

## Emotional Prosody Perception in Mandarin-speaking Congenital Amusics

*Yixin Zhang, Tianzhu Geng and Jinsong Zhang*

Beijing Advanced Innovation Center for Language Resources, Beijing Language and Culture University, Beijing 100083, China

**Wed-P-1-4-9, Time: 10:00-12:00**

Congenital amusia, which is a neurogenetic disorder affecting musical pitch processing, was found recently to affect not only human speech perception, but also emotional perception. Since previous studies only examined participants with non-tonal languages, they cannot easily generalize the finding to people with tonal language background, due to the fact that those people utilize pitch cues much more heavily in daily communication compared with others. To make clear the doubt, this paper investigates emotional prosody perception of Mandarin speakers with congenital amusia. We tried to recruit 19 amusics and matched control group of similar number of normal speakers and carried out emotional perception experiments in which speech and non-speech stimuli with six kinds of emotions were used, including happy, sad, fear, angry, surprise and neutral. Results showed that the amusics performed significantly worse than matched controls. This indicated that tone-language expertise cannot compensate for pitch deficits in amusia

**Notes**

for emotional perception. Further analyses demonstrated that there was a positive correlation between emotion prosody performance and pitch perceptional ability. These findings further support previous hypothesis that music and language share cognitive and neural resources and provide a new perspective on the proposition of the relation between music and language.

### Cultural Differences in Pattern Matching: Multisensory Recognition of Socio-affective Prosody

*Takaaki Shochi[1,2], Jean-Luc Rouas[1], Marine Guerry[2] and Donna Erickson[3]*

[1]Univ. Bordeaux, CNRS, LaBRI, UMR 5800, 33405 Talence, France

[2]CLLE-ERRSaB CNRS UMR 5263, Bordeaux, France

[3]Kanazawa Medical University, Japan

**Wed-P-1-4-10, Time: 10:00-12:00**

This study focuses on the cross-cultural differences in perception of audio visual prosodic recordings of Japanese social affects. The study compares cultural differences of perceptual patterns of 21 Japanese subjects with 20 French subjects who have no knowledge of Japanese language or Japanese social affects. The test material is a semantically affectively neutral utterance expressed in 9 various social affects by 2 Japanese speakers (one male, one female) who were chosen as best performers in our previous recognition experiment. The task was to create a specific audio-visual affect by choosing one video stimuli among 9 choices and one audio stimuli, again among 9 choices. The participants could preview each audio and video stimuli individually and also the combination of chosen stimuli. The results reveal that native subjects can correctly combine auditory and visually expressed social affects, showing some confusion inside semantic categories. Different matching patterns are observed for non-native subjects especially for a type of cultural-specific politeness.

## Perspective Talk-3

Hall 3; 12:00-12:30; Wednesday, 5 September, 2018

Chair: Chandra Sekhar Seelamantula

### Speech Processing in the Human Brain Meets Deep Learning

*Nima Mesgarani*

Electrical Engineering Department, Columbia University in the City of New York

**Wed-Perspective-3, Time: 12:00-12:30**

Speech processing technologies have seen tremendous progress since the advent of deep learning, where the most challenging problems no longer seem out of reach. In parallel, deep learning has advanced the state-of-the-art in processing the neural signals to speech in the human brain. This talk reports progress in three important areas of research: I) Decoding (reconstructing) speech from the human auditory cortex to establish a direct interface with the brain. Such an interface not only can restore communication for paralyzed patients, but also has the potential to transform human-computer interaction technologies, II) Auditory Attention Decoding, which aims to create a mind-controlled hearing aid that can track the brain-waves of a listener to identify and amplify the voice of the attended speaker in a crowd. Such a device could help hearing-impaired listeners communicate more effortlessly with others in noisy environments, and III) More accurate models of the transformations that the brain applies to speech at different stages of the human auditory pathway. This is achieved by training deep neural networks to learn the mapping from sound to the neural responses. Using a novel method to study the exact function learned by these neural networks has led to new insights on how the human brain processes speech. On the other hand, these new insights motivate distinct computational properties that can be incorporated into the neural network models to better capture the properties of speech processing in the human auditory cortex.

**Notes**

## Industry Presentation-4

Hall 3; 12:30-13:00; Wednesday, 5 September, 2018

Chairs: S R Mahadeva Prasanna and K T Deepak

**Presenter:** Yifan Gong

**Company:** Microsoft

## Industry Presentation-5

Hall 1; 12:30-13:00; Wednesday, 5 September, 2018

Chairs: C. Krishna Mohan and Venkata Subramanian Viraraghavan

**Presenter:** Wang Yujun

**Company:** Xiaomi

## Industry Presentation-6

Hall 2; 12:30-13:00; Wednesday, 5 September, 2018

Chairs: Anil Vuppala and Pratibha Moogi

**Presenter:** Joint Secretary, MeitY

**Company:** Ministry of Electronics and Information Technology (MeitY), Government of India

## Wed-O-2-1: Recurrent Neural Models for ASR

Hall 3; 14:30-16:30; Wednesday, 5 September, 2018

Chairs: Michael Riley and Patrick Nguyen

## ESPnet: End-to-End Speech Processing Toolkit

*Shinji Watanabe[1], Takaaki Hori[2], Shigeki Karita[3], Tomoki Hayashi[4], Jiro Nishitoba[5], Yuya Unno[6], Nelson Enrique Yalta Soplin[7], Jahn Heymann[8], Matthew Wiesner[1], Nanxin Chen[1], Adithya Renduchintala[1] and Tsubasa Ochiai[9]*

[1]Johns Hopkins University

[2]Mitsubishi Electric Research Laboratories

[3]NTT Communication Science Laboratories

[4]Nagoya University

[5]Retrieva, Inc.

[6]Preferred Networks, Inc.

[7]Waseda University

[8]Paderborn University

[9]Doshisha University

**Wed-O-2-1-1, Time: 14:30-14:50**

This paper introduces a new open source platform for end-to-end speech processing named ESPnet. ESPnet mainly focuses on end-to-end automatic speech recognition (ASR) and adopts widely-used dynamic neural network toolkits, Chainer and PyTorch, as a main deep learning engine. ESPnet also follows the Kaldi ASR toolkit style for data processing, feature extraction/format and recipes to provide a complete setup for speech recognition and other speech processing experiments. This paper explains a major architecture of this software platform, several important functionalities, which differentiate ESPnet from other open source ASR toolkits and experimental results with major ASR benchmarks.

## A GPU-based WFST Decoder with Exact Lattice Generation

*Zhehuai Chen[1,3], Justin Luitjens[2], Hainan Xu[3], Yiming Wang[3], Daniel Povey[3] and Sanjeev Khudanpur[3]*

[1]SpeechLab, Department of Computer Science and Engineering, Shanghai Jiao Tong University

[2]NVIDIA, USA

[3]Center for Language and Speech Processing, Johns Hopkins University

**Wed-O-2-1-2, Time: 14:50-15:10**

We describe initial work on an extension of the Kaldi toolkit that supports weighted finite-state transducer (WFST) decoding on Graphics Processing Units (GPUs). We implement token recombination as an atomic GPU operation in order to fully parallelize the Viterbi beam search and propose a dynamic load balancing strategy for more efficient token passing scheduling among GPU threads. We also redesign the exact lattice generation and lattice pruning algorithms for better utilization of the GPUs. Experiments on the Switchboard corpus show that the proposed method achieves identical 1-best results and lattice quality in recognition and confidence measure tasks, while running 3 to 15 times faster than the single process Kaldi decoder. The above results are reported on different GPU architectures. Additionally we obtain a 46-fold speedup with sequence parallelism and multi-process service (MPS) in GPU.

## Automatic Speech Recognition System Development in the "Wild"

*Anton Ragni and Mark Gales*

Department of Engineering, University of Cambridge

Trumpington Street, Cambridge CB2 1PZ, UK

**Wed-O-2-1-3, Time: 15:10-15:30**

The standard framework for developing an automatic speech recognition (ASR) system is to generate training and development
data for building the system and evaluation data for the final performance analysis. All the data is assumed to come from the domain of interest. Though this framework is matched to some tasks, it is more challenging for systems that are required to operate over broad domains, or where the ability to collect the required data is limited. This paper discusses ASR work performed under the IARPA MATERIAL program, which is aimed at cross-language information retrieval and examines this challenging scenario. In terms of available data, only limited narrow-band conversational telephone speech data was provided.

However, the system is required to operate over a range of domains, including broadcast data. As no data is available for the broadcast domain, this paper proposes an approach for system development based on scraping "related" data from the web and using ASR system confidence scores as the primary metric for developing the acoustic and language model components. As an initial evaluation of the approach, the Swahili development language is used, with the final system performance assessed on the IARPA MATERIAL Analysis Pack 1 data.

## Semantic Lattice Processing in Contextual Automatic Speech Recognition for Google Assistant

*Leonid Velikovich, Ian Williams, Justin Scheiner, Petar Aleksic, Pedro Moreno and Michael Riley*

Google, Inc

**Wed-O-2-1-4, Time: 15:30-15:50**

Recent interest in intelligent assistants has increased demand for Automatic Speech Recognition (ASR) systems that can utilize contextual information to adapt to the user's preferences or the current device state. For example, a user might be more likely to refer to their favorite songs when giving a "music playing" command or request to watch a movie starring a particular favorite actor when giving a "movie playing" command. Similarly, when a device is in a "music playing" state, a user is more likely to give volume control commands.

In this paper, we explore using semantic information inside the ASR word lattice by employing Named Entity Recognition (NER) to identify and boost contextually relevant paths in order to improve speech recognition accuracy. We use broad semantic classes comprising millions of entities, such as songs and musical artists, to tag relevant semantic entities in the lattice. We show that our method reduces Word Error Rate (WER) by 12.0% relative on a Google Assistant "media playing" commands test set, while not affecting WER on a test set containing commands unrelated to media.

## Contextual Speech Recognition in End-to-End Neural Network Systems Using Beam Search

*Ian Williams, Anjuli Kannan, Petar Aleksic, David Rybach and Tara Sainath*

Google, Inc.

**Wed-O-2-1-5, Time: 15:50-16:10**

Recent work has shown that end-to-end (E2E) speech recognition architectures such as Listen Attend and Spell (LAS) can achieve state-of-the-art quality results in LVCSR tasks. One benefit of this architecture is that it does not require a separately trained pronunciation model, language model and acoustic model. However, this property also introduces a drawback: it is not possible to adjust language model contributions separately from the system as a whole. As a result, inclusion of dynamic, contextual information (such as nearby restaurants or upcoming events) into recognition requires a different approach from what has been applied in conventional systems.

We introduce a technique to adapt the inference process to take advantage of contextual signals by adjusting the output likelihoods of the neural network at each step in the beam search. We apply the proposed method to a LAS E2E model and show its effectiveness in experiments on a voice search task with both artificial and real contextual information. Given optimal context, our system reduces WER from 9.2% to 3.8%. The results show that this technique is effective at incorporating context into the prediction of an E2E system.

## Forward-Backward Attention Decoder

*Masato Mimura, Shinsuke Sakai and Tatsuya Kawahara*

Kyoto University, School of Informatics, Sakyo-ku, Kyoto 606-8501, Japan

**Notes**

**Wed-O-2-1-6, Time: 16:10-16:30**
This paper investigates how forward and backward attentions can be integrated to improve the performance of attention-based sequence-to-sequence (seq2seq) speech recognition systems. In the proposed approach, speech is decoded from left to right as well as from right to left utilizing forward and backward attention vectors and the best sentence hypothesis is searched for according to combined probabilities provided by the decoders of two directions.

Our method takes advantage of two distinct and complementary ways of extracting information from the asymmetric time structure of speech. It also mitigates a drawback of attention-based models that they tend to output less reliable labels due to error accumulation when the utterance becomes longer.

We also show the effectiveness of a multitask learning in which the forward decoder is jointly trained with backward decoding sharing a single encoder. The proposed forward-backward decoding improved word error rates (WERs) of word-level attention models by up to 12.7% relative in speech recognition experiments using large-scale spontaneous speech corpora. They achieve much higher performances than a state-of-the-art hybrid DNN-HMM system while retaining the advantage of very low latency.

# Wed-O-2-2: Speaker Verification Using Neural Network Methods I

Hall 1; 14:30-16:30; Wednesday, 5 September, 2018
Chairs: Héctor Delgado and Jan Černocký

## Learning Discriminative Features for Speaker Identification and Verification

*Sarthak Yadav and Atul Rai*
Staqu Technologies, India

**Wed-O-2-2-1, Time: 14:30-14:50**
The success of any Text Independent Speaker Identification and/or Verification system relies upon the system's capability to learn discriminative features.

In this paper we propose a Convolutional Neural Network (CNN) Architecture based on the popular Very Deep VGG [1] CNNs, with key modifications to accommodate variable length spectrogram inputs, reduce the model disk space requirements and reduce the number of parameters, resulting in significant reduction in training times. We also propose a unified deep learning system for both Text-Independent Speaker Recognition and Speaker Verification, by training the proposed network architecture under the joint supervision of Softmax loss and Center loss [2] to obtain highly discriminative deep features that are suited for both Speaker Identification and Verification Tasks.

We use the recently released VoxCeleb dataset [3], which contains hundreds of thousands of real world utterances of over 1200 celebrities belonging to various ethnicities, for benchmarking our approach. Our best CNN model achieved a Top-1 accuracy of 84.6%, a 4% absolute improvement over VoxCeleb's approach, whereas training in conjunction with Center Loss improved the Top-1 accuracy to 89.5%, a 9% absolute improvement over Voxceleb's approach.

## Triplet Loss Based Cosine Similarity Metric Learning for Text-independent Speaker Recognition

*Sergey Novoselov[1,2], Vadim Shchemelinin[1,2], Andrey Shulipa[2], Alexandr Kozlov[1] and Ivan Kremnev[1]*
[1]STC Ltd., St. Petersburg, Russia
[2]ITMO University, St. Petersburg, Russia

**Wed-O-2-2-2, Time: 14:50-15:10**
Deep neural network based speaker embeddings become increasingly popular in the text-independent speaker recognition task. In contrast to a generatively trained i-vector extractor, a DNN speaker embedding extractor is usually trained discriminatively in the closed set classification scenario using softmax. The problem we addressed in the paper is choosing a dnn based speaker embedding backend solution for speaker verification scoring. There are several options to perform speaker verification in the dnn embedding space. One of them is using a simple heuristic speaker similarity metric for the scoring (e.g. cosine metric). Similarly in the i-vector based systems, the standard Linear Discriminant Analisys (LDA) followed by the Probabilistic Linear Discriminant Analisys (PLDA) can be used for segregating speaker information. As an alternative, the discriminative metric learning approach can be considered. This work demonstrates that performance of deep speaker embeddings based systems can be improved by using Cosine Similarity Metric Learning (CSML) with the triplet loss training scheme. Results obtained on Speakers in the Wild and NIST SRE 2016 evaluation sets demonstrate superiority and robustness of CSML based systems.

## Speaker Embedding Extraction with Phonetic Information

*Yi Liu[1], Liang He[1], Jia Liu[1] and Michael T. Johnson[2]*
[1]Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[2]Department of Electrical and Computer Engineering, University of Kentucky, 453 F. Paul Anderson Tower, Lexington, KY 40506-0046, USA

**Wed-O-2-2-3, Time: 15:10-15:30**
Speaker embeddings achieve promising results on many speaker verification tasks. Phonetic information, as an important component of speech, is rarely considered in the extraction of speaker embeddings. In this paper, we introduce phonetic information to the speaker embedding extraction based on the x-vector architecture. Two methods using phonetic vectors and multi-task learning are proposed. On the Fisher dataset, our best system outperforms the original x-vector approach by 20% in EER and by 15%, 15% in minDCF08 and minDCF10, respectively. Experiments conducted on NIST SRE10 further demonstrate the effectiveness of the proposed methods.

## Attentive Statistics Pooling for Deep Speaker Embedding

*Koji Okabe[1], Takafumi Koshinaka[1] and Koichi Shinoda[2]*
[1]Data Science Research Laboratories, NEC Corporation, Japan
[2]Department of Computer Science, Tokyo Institute of Technology, Japan

**Wed-O-2-2-4, Time: 15:30-15:50**
This paper proposes attentive statistics pooling for deep speaker embedding in text-independent speaker verification. In conventional speaker embedding, frame-level features are averaged over all the frames of a single utterance to form an utterance-level feature. Our method utilizes an attention mechanism to give different weights to different frames and generates not only weighted means but also weighted standard deviations. In this way, it can capture long-term variations in speaker characteristics more effectively. An evaluation on the NIST SRE 2012 and the VoxCeleb data sets shows that it reduces equal error rates (EERs) from the conventional method by 7.5% and 8.1%, respectively.

## Robust and Discriminative Speaker Embedding via Intra-Class Distance Variance Regularization

*Nam Le[1,2] and Jean-Marc Odobez[1,2]*
[1]Idiap Research Institute, Martigny, Switzerland
[2]École Polytechnique Fédérale de Lausanne, Switzerland

**Wed-O-2-2-5, Time: 15:50-16:10**
Learning a good speaker embedding is critical for many speech processing tasks, including recognition, verification and diarization. To this end, we propose a complementary optimizing goal called intra-class loss to improve deep speaker embeddings learned with triplet loss. This loss

**Notes**

function is formulated as a soft constraint on the averaged pair-wise distance between samples from the same class. Its goal is to prevent the scattering of these samples within the embedding space to increase the intra-class compactness.When intra-class loss is jointly optimized with triplet loss, we can observe 2 major improvements: the deep embedding network can achieve a more robust and discriminative representation and the training process is more stable with a faster convergence rate. We conduct experiments on 2 large public benchmarking datasets for speaker verification, VoxCeleb and VoxForge. The results show that intra-class loss helps accelerating the convergence of deep network training and significantly improves the overall performance of the resulted embeddings.

### Deep Discriminative Embeddings for Duration Robust Speaker Verification

*Na Li, Deyi Tuo, Dan Su, Zhifeng Li and Dong Yu*
Tencent AI lab

**Wed-O-2-2-6, Time: 16:10-16:30**

The embedding-based deep convolution neural networks (CNNs) have demonstrated effective for text-independent speaker verification systems with short utterances. However, the duration robustness of the existing deep CNNs based algorithms has not been investigated when dealing with utterances of arbitrary duration. To improve robustness of embedding-based deep CNNs for longer duration utterances, we propose a novel algorithm to learn more discriminative utterance-level embeddings based on the Inception-ResNet speaker classifier. Specifically, the discriminability of embeddings is enhanced by reducing intra-speaker variation with center loss and simultaneously increasing inter-speaker discrepancy with softmax loss. To further improve system performance when long utterances are available, at test stage long utterances are segmented into shorter ones, where utterance-level speaker embeddings are extracted by an average pooling layer. Experimental results show that when cosine distance is employed as the measure of similarity for a trial, the proposed method outperforms ivector/PLDA framework for short utterances and is effective for long utterances.

## Wed-O-2-3: Speech Perception in Adverse Conditions

Hall 2; 14:30-16:30; Wednesday, 5 September, 2018
Chairs: Jeesun Kim and Samudravijaya K

### Impact of Different Speech Types on Listening Effort

*Olympia Simantiraki[1], Martin Cooke[1] and Simon King[2]*
[1]Language and Speech Laboratory, Universidad del País Vasco, Vitoria, Spain
[2]Center for Speech Technology Research, University of Edinburgh, Edinburgh, UK

**Wed-O-2-3-1, Time: 14:30-14:50**

Listeners are exposed to different types of speech in everyday life, from natural speech to speech that has undergone modifications or has been generated synthetically. While many studies have focused on measuring the intelligibility of these distinct speech types, their impact on listening effort is not known. The current study combined an objective measure of intelligibility, a physiological measure of listening effort (pupil size) and listeners' subjective judgements, to examine the impact of four speech types: plain (natural) speech, speech produced in noise (Lombard speech), speech enhanced to promote intelligibility and synthetic speech. For each speech type, listeners responded to sentences presented in one of three levels of speech-shaped noise. Subjective effort ratings and intelligibility scores showed an inverse ranking across speech types, with synthetic speech being the most demanding and enhanced speech the least. Pupil size measures indicated an increase in listening effort with decreasing signal-to-noise ratio for all speech types apart from synthetic speech, which required significantly more effort at the most favourable noise level. Naturally and artificially modified speech were less effortful than plain speech at the more adverse noise levels. These outcomes indicate a clear impact of speech type on the cognitive demands required for comprehension.

### Who Are You Listening to? Towards a Dynamic Measure of Auditory Attention to Speech-on-speech

*Moïra-Phoebé Huet [1,2], Christophe Micheyl [1,3], Etienne Gaudrain [1,4] and Etienne Parizet [2]*
[1]Université de Lyon, CNRS UMR5292, Inserm U1028, Lyon Neuroscience Research Center, Lyon, France
[2]Université de Lyon, Institut National des Sciences Appliquées de Lyon, Laboratoire Vibrations Acoustique, F-69621 Villeurbanne, France
[3]Starkey, Créteil, France
[4]University of Groningen, University Medical Center Groningen, Department of Otorhinolaryngology, Groningen, Netherlands

**Wed-O-2-3-2, Time: 14:50-15:10**

When studying speech-on-speech perception, even when participants are explicitly instructed to focus selectively on a single voice, they can spuriously find themselves listening to the wrong voice. These paradigms generally do not allow to infer, retrospectively, which of the speakers was listened to, at different times during presentation. The present study sought to develop a psychophysical test paradigm and a set of speech stimuli to that purpose. In this paradigm, after listening to two simultaneous stories, the participant had to identify, among a set of words, those that were present in the target story. Target and masker stories were presented dichotically or diotically. F0 and vocal-tract length were manipulated in order to parametrically vary the distance between the target and masker voices. Consistent with the hypothesis that correct-identification performance for target words depends on selective attention, performance decreases with the distance between the target and masker voices. These results indicate that the paradigm and stimuli described here can be used to infer which voice a participant is listening to in concurrent-speech listening experiments.

### Investigating the Role of Familiar Face and Voice Cues in Speech Processing in Noise

*Jeesun Kim[1], Sonya Karisma[1], Vincent Aubanel[1,2] and Chris Davis[1]*
[1]The MARCS Institute, Western Sydney University, Australia
[2]Gipsa-Lab, Grenoble-Alpes University, France

**Wed-O-2-3-3, Time: 15:10-15:30**

The speech of a familiar talker is better recognized in noise than an unfamiliar one, suggesting that listeners access talker-specific models to assist with degraded input. This study investigated whether a talker model could be accessed by presenting the face of a talker. In the experiment, participants were trained in recognizing three talkers' faces and voices to ceiling-level. Participants were then given a speech in noise recognition task consisting of four talker conditions: familiar face then familiar voice; unfamiliar face then familiar voice, familiar face then unfamiliar voice; and unfamiliar face then unfamiliar voice. A talker familiarity effect was found, i.e., speech perception was more accurate in the familiar face and familiar voice condition than all other ones. A familiar voice did not produce a talker familiarity effect when paired with an unfamiliar face. The familiar face and unfamiliar voice condition had the poorest performance, indicating that pairing a familiar face and unfamiliar voice had a disruptive effect. The results suggest that listeners develop a talker model that includes details of both the voice and the face; and that accessing this model can in some circumstances be wholly determined by face cues.

### The Conversation Continues: the Effect of Lyrics and Music Complexity of Background Music on Spoken-Word Recognition

*Odette Scharenborg[1,2,3] and Martha Larson[1,3,4]*
[1]Centre for Language Studies, Radboud University Nijmegen, Netherlands
[2]Donders Institute for Brain, Cognition, & Behavior, Radboud University Nijmegen, Netherlands

**Notes**

[3]Intelligent Systems Department, Delft University of Technology, Netherlands
[4]Institute for Computing and Information Sciences, Radboud University Nijmegen, Netherlands

**Wed-O-2-3-4, Time: 15:30-15:50**

Background music in social interaction settings can hinder conversation. Yet, little is known of how specific properties of music impact speech processing. This paper addresses this knowledge gap by investigating the effect of the 1) complexity of the background music and 2) the presence versus absence of sung lyrics on spoken-word recognition in background music. To answer these questions, a word identification experiment was run in which Dutch participants listened to Dutch CVC words embedded in stretches of background music in four conditions: low/high complexity and with lyrics/music-only and at three SNRs. Music stretches with and without lyrics were sampled from the same song in order to control for factors beyond the complexity of the music and the presence of lyrics. The results showed a clear negative impact of more complex music and the presence of lyrics in background music on spoken-word recognition. The results open a path for future work and suggest that social spaces (e.g., restaurants, cafés and bars) should make careful choices of music to promote conversation.

### Loud and Shouted Speech Perception at Variable Distances in a Forest

*Julien Meyer[1], Fanny Meunier[2], Laure Dentel[3], Noelia Do Carmo Blanco[2] and Frédéric Sèbe[4]*
[1]Univ. Grenoble Alpes, CNRS, GIPSA-lab, Grenoble 38000, France
[2]Université Côte d'Azur, CNRS, BCL, France
[3]The World Whistles Research Association, Paris, France
[4]Equipe de Neuro-Ethologie Sensorielle, Neuro-PSI, CNRS UMR 9197, Univ. Lyon/Saint Etienne

**Wed-O-2-3-5, Time: 15:50-16:10**

To increase the range of modal speech in natural ambient noise, individuals increase their vocal effort and may pass into the 'shouted speech' register. To date, most studies concerning the influence of distance on spoken communication in outdoor natural environments have focused on the 'productive side' of the human ability to tacitly adjust vocal output to compensate for acoustic losses due to sound propagation. Our study takes a slightly different path as it is based on an adaptive speech production/perception experiment. The setting was an outdoor natural soundscape (a plane forest in altitude). The stimuli were produced live during the interaction: each speaker adapted speech to transmit French disyllabic words in isolation to an interlocutor/listener who was situated at variable distances in the course of the experiment (30m, 60m, 90m). Speech recognition was explored by evaluating the ability of 16 normal-hearing French listeners to recognize these words and their constituent vowels and consonants. Results showed that in such conditions, speech adaptation was rather efficient as word recognition remained around 95% at 30m, 85% at 60m and 75% at 90m. We also observed striking differences in patterns of answers along several lines: different distances, speech registers, vowels and consonants.

### Phoneme Resistance and Phoneme Confusion in Noise: Impact of Dyslexia

*Noelia Do Carmo Blanco[1], Julien Meyer[2], Michel Hoen[3] and Fanny Meunier[1]*
[1]Université Côte d'Azur, CNRS, BCL, France
[2]Université Grenoble Alpes, CNRS, GIPSA-Lab, Grenoble, France
[3]Oticon Medical, Vallauris, France

**Wed-O-2-3-6, Time: 16:10-16:30**

Understanding speech in noisy environments is a challenge for almost everyone and particularly so for people with dyslexia. To better understand the phonological processing deficit, which has been posited as a core trait of dyslexia, we wanted to further characterize the impact of noise on speech perception. In this paper we investigated phoneme resistance to noise for dyslexic and control adults and explored the pattern of errors produced by noise interference. Our aim was to examine differences between phoneme confusion matrices of the two populations.

Disyllabic nouns were embedded in noise and participants had to perform an auditory word identification task. Error rates, phoneme resistance and phoneme confusions were compared between a dyslexic and a group of matched controls.

Error rate was higher in the dyslexic group. However, no qualitative differences in the profile of errors were found. The coronals /ʃ and s/ were the most resistant phoneme in both groups while the labials /f, m and v/ were the most vulnerable. Although dyslexics showed a more scattered pattern of confusions, the matrices were correlated. Our results confirm a phonological deficit in dyslexia whereas they do not support the hypothesis of qualitative differences in phonological representation between the two groups.

## Wed-O-2-4: Measuring Pitch and Articulation

MR G.01-G.02; 14:30-16:30; Wednesday, 5 September, 2018
Chairs: Balamurali B T and Qiguang Lin

### Conditional End-to-End Audio Transforms

*Albert Haque[1], Michelle Guo[1] and Prateek Verma[2]*
[1]Department of Computer Science, Stanford University, USA
[2]Center for Computer Research in Music and Acoustics, Stanford University, USA

**Wed-O-2-4-1, Time: 14:30-14:50**

We present an end-to-end method for transforming audio from one style to another. For the case of speech, by conditioning on speaker identities, we can train a single model to transform words spoken by multiple people into multiple target voices. For the case of music, we can specify musical instruments and achieve the same result. Architecturally, our method is a fully-differentiable sequence-to-sequence model based on convolutional and hierarchical recurrent neural networks. It is designed to capture long-term acoustic dependencies, requires minimal post-processing and produces realistic audio transforms. Ablation studies confirm that our model can separate acoustic properties from musical and language content at different receptive fields. Empirically, our method achieves competitive performance on community-standard datasets.

### Detection of Glottal Closure Instants in Degraded Speech Using Single Frequency Filtering Analysis

*Gunnam Aneeja, Sudarsana Reddy Kadiri and Bayya Yegnanarayana*
Speech Processing Laboratory,
International Institute of Information Technology, Hyderabad, India

**Wed-O-2-4-2, Time: 14:50-15:10**

Impulse-like characteristics of excitation occur at the glottal closure instant (GCI) due to sharp closure of the vibrating vocal folds in each glottal cycle. The GCIs are detected from the excitation component of the speech signal and the excitation component is derived using inverse filtering or its variants. In this paper we propose a method for GCI detection based on single frequency filtering (SFF) of the speech signal. The SFF output has high signal-to-noise ratio (SNR) property in speech regions. The variance (across frequency) contour computed from the SFF output show rapid changes around the GCIs and these rapid changes can be observed even when the speech signal is degraded. Thus the GCI locations can be extracted even from degraded speech using the SFF analysis. The robustness of the method is demonstrated for several cases of degradation of speech signal.

### Tone Recognition Using Lifters and CTC

*Loren Lugosch and Vikrant Singh Tomar*
Fluent.ai Inc., Montréal, Québec, Canada

**Wed-O-2-4-3, Time: 15:10-15:30**

**Notes**

In this paper, we present a new method for recognizing tones in continuous speech for tonal languages. The method works by converting the speech signal to a cepstrogram, extracting a sequence of cepstral features using a convolutional neural network and predicting the underlying sequence of tones using a connectionist temporal classification (CTC) network. The performance of the proposed method is evaluated on a freely available Mandarin Chinese speech corpus, AISHELL-1 and is shown to outperform the existing techniques in the literature in terms of tone error rate (TER).

### Epoch Extraction from Pathological Children Speech Using Single Pole Filtering Approach

*Vikram C M[1] and S R Mahadeva Prasanna[1,2]*

[1]Indian Institute of Technology Guwahati, Guwahati, India

[2]Indian Institute of Technology Dharwad, Dharwad, India

**Wed-O-2-4-4, Time: 15:30-15:50**

The instant of significant excitation of the vocal tract system is referred to the epoch of the speech signal. The presence of high pitch and aperiodicity are the major challenges for the epoch extraction from the speech of pathological children. In this work, impulse-like characteristics of epochs derived from single pole filter based time-frequency representation are exploited to propose an epoch extraction algorithm for the pathological children speech. The sharp transitions present in the single pole filtered envelope at the epochs are enhanced using multi-scale product computation. Further, the combined evidence derived from the multi-scale product of the filtered envelopes at different frequencies is used to locate the epochs. The proposed algorithm is evaluated over the Saarbruecken Voice Database containing pathological children speech and simultaneously recorded electroglottographic signals. The proposed method showed better identification accuracy for pathological children speech when compared to state-of-the-art techniques.

### Automated Classification of Vowel-Gesture Parameters Using External Broadband Excitation

*Balamurali B T and Jer-Ming Chen*

Audio Research Group, Singapore University of Technology & Design, Singapore

**Wed-O-2-4-5, Time: 15:50-16:10**

External broadband signal excitation applied at the speaker (or singer)'s mouth has previously been successfully used to estimate acoustic resonances of the vocal tract during speaking and singing. In this study, we used a modified, low cost, light-weight, pocket-sized and simplified version of this measurement technique, with reduced sampling time and improved low frequency detection, so that such vocal tract measurements may be easily deployed 'in the field' and facilitate a more 'ecological/natural' tracking of phonatory gestures. This system was investigated with 6 volunteer speakers phonating 17 English vowels and the relative impedance spectrum $\gamma$ ('gamma') was measured. Although the $\gamma(f)$ signal measured here for each phonatory gesture is somewhat noisier than the original technique, it is still believed to carry some important cues associated with vocal tract configuration that produce these vowels. Features were identified both in the amplitude and phase of $\gamma(f)$ and three ensemble classifiers namely random forest, gradient boosting and adaboost were trained using them. The prediction output from these classifiers were combined using soft voting to predict a class label (front-central-back; open-close). This yielded an accuracy exceeding 80% in classifying the six nominal regions of the vowel plane.

### Estimation of Fundamental Frequency from Singing Voice Using Harmonics of Impulse-like Excitation Source

*Sudarsana Reddy Kadiri and Bayya Yegnanarayana*

Speech Processing Laboratory, International Institute of Information Technology, Hyderabad, India

**Wed-O-2-4-6, Time: 16:10-16:30**

This paper focuses on the problem of estimating fundamental frequency from singing voice. Estimation of fundamental frequency is a well studied topic in the speech research community. From the recent studies on fundamental frequency estimation from singing voice with state-of-art methods proposed for speech, there exists a significant gap in accuracy for singing voice. This is mainly because of the wider and rapid variations in pitch in singing voice compared to that in speech. To overcome this, in this paper we propose a method to derive the fundamental frequency from singing voice by exploiting the harmonics of impulse-like excitation in sequence of glottal cycles. The proposed method is compared with the eight state-of-art methods such as YIN, SWIPE, YAAPT, RAPT, SRH, SFF_CEP, PEFAC and SHRP on the LYRICS singing database. From the experimental results, it is observed that the accuracy of fundamental frequency by the proposed method is better than many state-of-art methods in various singing categories and laryngeal mechanisms.

## Wed-O-2-5: Speech and Language Analytics for Mental Health

MR G.03-G.04; 14:30-16:30; Wednesday, 5 September, 2018

Chairs: Emily Mower Provost and Tanja Schultz

### Investigating the Effect of Audio Duration on Dementia Detection Using Acoustic Features

*Jochen Weiner[1], Miguel Angrick[1], Srinivasan Umesh[2] and Tanja Schultz[1]*

[1]Cognitive Systems Lab, University of Bremen, Germany

[2]Department of Electrical Engineering, Indian Institute of Technology (IIT) Madras, India

**Wed-O-2-5-1, Time: 14:30-14:50**

This paper presents recent progress toward our goal to enable area-wide pre-screening methods for the early detection of dementia based on automatically processing conversational speech of a representative group of more than 200 subjects. We focus on conversational speech since it is the natural form of communication that can be recorded unobtrusively, without adding stress to subjects and without the need of controlled clinical settings. We describe our unsupervised process chain consisting of voice activity detection and speaker diarization followed by extraction of features and detection of early signs of dementia. The unsupervised system achieves up to 0.645 unweighted average recall (UAR) and compares favorably to a system that was carefully designed on manually annotated data. To further lower the burden for subjects, we investigate UAR over speech duration and find that about 12 minutes of interview are sufficient to achieve the best UAR.

### An Interlocutor-Modulated Attentional LSTM for Differentiating between Subgroups of Autism Spectrum Disorder

*Yun-Shao Lin[1,3], Susan Shur-Fen Gau[2] and Chi-Chun Lee[1,3]*

[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan

[2]Department of Psychiatry, National Taiwan University Hospital and College of Medicine, Taiwan

[3]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan

**Wed-O-2-5-2, Time: 14:50-15:10**

Recalling and discussing personal emotional experiences is one of the key procedures in assessing complex affect processing of individuals with Autism Spectrum Disorder (ASD). This procedure is a standard subpart of a diagnostic interview to assess ASD - the Autism Diagnostic Observation Schedule (ADOS). Previous work has demonstrated that the behavior features computed from this procedure in ADOS possess discriminative information between the three distinct ASD subgroups: Autistic Disorder (AD), High Functioning Autism (HFA) and Asperger Syndrome (AS). In this work, we propose an interlocutor-modulated attentional long short term memory network (IM-aLSTM) that models the ASD individual's acoustic features with a novel interlocutor-modulated attention mechanism. Our IM-

**Notes**

aLSTM achieves ASD subgroup categorization accuracy of 66.5%, which is a 14% absolute improvement over baseline method on the same database. Our analyses further indicate that the attention weights are concentrated more on interaction segments where the ASD individual is being asked to recall and discuss his/her own negative emotional experiences.

### Recognition of Echolalic Autistic Child Vocalisations Utilising Convolutional Recurrent Neural Networks

*Shahin Amiriparian[1], Alice Baird[1], Sahib Julka[1], Alyssa Alcorn[2], Sandra Ottl[1], Sunčica Petrović[3], Eloise Ainger[2], Nicholas Cummins[1] and Björn Schuller[1,4]*

[1]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany

[2]Centre for Research in Autism and Education, UCL Institute of Education, U. K.

[3]Serbian Society of Autism, Belgrade, Serbia

[4]GLAM – Group on Language, Audio & Music, Imperial College London, U. K.

**Wed-O-2-5-3, Time: 15:10-15:30**

Autism spectrum conditions (ASC) are a set of neuro-developmental conditions partly characterised by difficulties with communication. Individuals with ASC can show a variety of atypical speech behaviours, including echolalia or the `echoing' of another's speech. We herein introduce a new dataset of 15 Serbian ASC children in a human-robot interaction scenario, annotated for the presence of echolalia amongst other ASC vocal behaviours. From this, we propose a four-class classification problem and investigate the suitability of applying a 2D convolutional neural network augmented with a recurrent neural network with bidirectional long short-term memory cells to solve the proposed task of echolalia recognition. In this approach, log Mel-spectrograms are first generated from the audio recordings and then fed as input into the convolutional layers to extract high-level spectral features. The subsequent recurrent layers are applied to learn the long-term temporal context from the obtained features. Finally, we use a feed forward neural network with softmax activation to classify the dataset. To evaluate the performance of our deep learning approach, we use leave-one-subject-out cross-validation. Key results presented indicate the suitability of our approach by achieving a classification accuracy of 83.5% unweighted average recall.

### Modeling Interpersonal Influence of Verbal Behavior in Couples Therapy Dyadic Interactions

*Sandeep Nallan Chakravarthula[1], Brian Baucom[2] and Panayiotis Georgiou[1]*

[1]Dept. of Electrical Engineering, University of Southern California

[2]Dept. of Psychology, University of Utah

**Wed-O-2-5-4, Time: 15:30-15:50**

Dyadic interactions among humans are marked by speakers continuously influencing and reacting to each other in terms of responses and behaviors, among others. Understanding how interpersonal dynamics affect behavior is important for successful treatment in psychotherapy domains. Traditional schemes that automatically identify behavior for this purpose have often looked at only the target speaker. In this work, we propose a text-based Markov model of how a target speaker's behavior is influenced by their own past behavior as well as their perception of their partner's behavior, based on lexical features. Apart from incorporating additional potentially useful information, our model can also control the degree to which the partner affects the target speaker. We evaluate our proposed model on the task of classifying Negative behavior in Couples Therapy and show that it is more accurate than the single-speaker model. Furthermore, we investigate the degree to which the optimal influence relates to how well a couple does on the long-term, via relating to relationship outcomes.

### Computational Modeling of Conversational Humor in Psychotherapy

*Anil Ramakrishna[1], Timothy Greer[1], David Atkins[2] and Shrikanth Narayanan[1]*

[1]Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, USA

[2]Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, USA

**Wed-O-2-5-5, Time: 15:50-16:10**

Humor is an important social construct that serves several roles in human communication. Though subjective, it is culturally ubiquitous and is often used to diffuse tension, specially in intense conversations such as those in psychotherapy sessions. Automatic recognition of humor has been of considerable interest in the natural language processing community thanks to its relevance in conversational agents. In this work, we present a model for humor recognition in Motivational Interviewing based psychotherapy sessions. We use a Long Short Term Memory (LSTM) based recurrent neural network sequence model trained on dyadic conversations from psychotherapy sessions and our model outperforms a standard baseline with linguistic humor features.

### Multimodal I-vectors to Detect and Evaluate Parkinson's Disease

*Nicanor Garcia[1], Juan Camilo Vásquez Correa[1,2], Juan Rafael Orozco-Arroyave[1,2] and Elmar Nöth[2]*

[1]Faculty of Engineering, Universidad de Antioquia UdeA, Medellín, Colombia

[2]Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

**Wed-O-2-5-6, Time: 16:10-16:30**

Parkinson's Disease (PD) is a neurodegenerative disorder characterized by a variety of motor symptoms. PD patients show several motor deficits, including speech deficits, impaired handwriting and gait disturbances. In this work we propose a methodology to fuse i-vectors extracted from three different bio-signals: speech, handwriting and gait. These i-vectors are used to classify Parkinson's Disease patients and healthy controls and to evaluate the neurological state of the patients. Speech i-vectors are extracted from MFCCs, handwriting i-vectors are extracted from kinematic features and gait i-vectors are extracted from modified MFCCs computed from inertial sensor signals. Two fusion strategies are tested: concatenating the i-vectors of a subject to form a super-i-vector with information from the three bio-signals and score pooling. The proposed fusion methods leads to better classification results respect to the separate analysis with each bio-signal, reaching an accuracy of up to 85%.

## Wed-SS-2-1: Special Session: Spoken CALL Shared Task, Second Edition

MR 1.01-1.02; 14:30-16:30; Wednesday, 5 September, 2018

Chairs: Johanna Gerlach, Manny Rayner, Martin Russell and Helmer Strik

### Overview of the 2018 Spoken CALL Shared Task

*Claudia Baur[1], Andrew Caines[2], Cathy Chua[3], Johanna Gerlach[1], Mengjie Qian[4], Manny Rayner[1], Martin Russell[4], Helmer Strik[5] and Xizi Wei[4]*

[1]FTI/TIM, University of Geneva, Switzerland

[2]Automated Language Teaching & Assessment Institute, University of Cambridge

[3]Independent researcher

[4]Department of Electronic, Electrical and Systems Engineering, University of Birmingham

[5]Centre for Language Studies (CLS), Radboud University Nijmegen

**Wed-SS-2-1-1, Time: 14:30-14:50**

**Notes**

We present an overview of the second edition of the Spoken CALL Shared Task. Groups competed on a prompt-response task using English-language data collected, through an online CALL game, from Swiss German teens in their second and third years of learning English. Each item consists of a written German prompt and an audio file containing a spoken response. The task is to accept linguistically correct responses and reject linguistically incorrect ones, with "linguistically correct" defined by a gold standard derived from human annotations. Scoring was performed using a metric defined as the ratio of the relative rejection rates on incorrect and correct responses. The second edition received eighteen entries and showed very substantial improvement on the first edition; all entries were better than the best entry from the first edition and the best score was about four times higher. We present the task, the resources, the results, a discussion of the metrics used and an analysis of what makes items challenging. In particular, we present quantitative evidence suggesting that incorrect responses are much more difficult to process than correct responses and that the most significant factor in making a response challenging is its distance from the closest training example.

## The CSU-K Rule-Based System for the 2nd Edition Spoken CALL Shared Task

*Dominik Jülg[1], Mario Kunstek[1], Cem Philipp Freimoser[1], Kay Berkling[1] and Mengjie Qian[2]*

[1]Cooperative State University, Karlsruhe (CSU-K), Germany

[2]Department of Electronic, Electrical & Systems Engineering, The University of Birmingham, UK

**Wed-SS-2-1-2, Time: 14:50-15:08**

This paper presents the set-up and results of the rule-based Cooperative State University Karlsruhe (CSU-K) system for the 2nd edition of the shared spoken CALL ESL task. The data was collected from Swiss teenage students using a speech-enabled online tool for English conversation practice. The tool should eventually be able to judge student input with respect to syntactic and semantic correctness. The tasks consisted of training data of a German text prompt with the associated audio file containing an English language response by the students.

In the second edition of the task, 6.698 utterances were provided in addition to the 2017 task. The contribution of this paper is a further look at how rule-based systems can be employed for these sorts of tasks. Meaning and grammar are treated separately in order to classify the language as correct. A number of experts were constructed to deal separately with different POS such as nouns, adjectives, verb usage and pronouns or determiners. Distance measurements derived from Doc2Vec where then employed between utterance and prompt responses. A D-value of 10.08 is reported on the final 2nd Edition evaluation test files.

## Liulishuo's System for the Spoken CALL Shared Task 2018

*Huy Nguyen, Lei Chen, Ramon Prieto, Chuan Wang and Yang Liu*

Liulishuo www.liulishuo.com/en

**Wed-SS-2-1-3, Time: 15:08-15:26**

The Spoken CALL (Computer-Assisted Language Learning) 2018 shared task requires systems to automatically accept or reject each single-sentence spoken response depending on whether the response is correct given a prompt. Spoken responses are first recognized into texts and then classified as 'accept' or 'reject' based on their language and meaning. This paper describes our system for the shared task. We focused on improving speech recognition performance, developing a rich set of features to capture the linguistic and semantic meaning of the responses and optimizing classification results for various factors (training set, n-best hypotheses of speech recognition, decision threshold, model ensemble). Our system achieves the best performance among the participating teams.

## An Optimization Based Approach for Solving Spoken CALL Shared Task

*Mohammad Ateeq, Abualsoud Hanani and Aziz Qaroush*

Birzeit University, Palestine

**Wed-SS-2-1-4, Time: 15:26-15:44**

In this paper, we are describing our developed systems for the 2018 SLaTE CALL Shared Task on grammatical and linguistic assessment of English spoken by German-speaking Swiss teenagers. The English spoken response is converted to text using baseline English DNN-HMM ASR trained on the shared task training data and another two commercial ASRs (Google and Microsoft Bing). The produced transcription is assessed in terms of language and meaning errors. In this work, we focused on the text-processing component. Grammatical errors are detected using English grammar checker, part of speech analysis and extracting incorrect bi-grams from grammatically incorrect responses. Errors related to the meaning are detected using novel approaches which measure the similarity between the given response and stored set of reference responses.

The outputs of several systems have been fused together into one overall system, where the fusion weights and parameters are tuned using genetic algorithm. The best result on the 2018 shared task test dataset is D-score of 14.41, which was achieved by the fused system and the optimized set of incorrect bi-grams.

## The University of Birmingham 2018 Spoken CALL Shared Task Systems

*Mengjie Qian, Xizi Wei, Peter Jančovič and Martin Russell*

Department of Electronic, Electrical & Systems Engineering, The University of Birmingham, UK

**Wed-SS-2-1-5, Time: 15:44-16:02**

This paper describes the systems developed by the University of Birmingham for the 2018 CALL Shared Task (ST) challenge. The task is to perform automatic assessment of grammatical and linguistic aspects of English spoken by German-speaking Swiss teenagers. Our developed systems consist of two components, automatic speech recognition (ASR) and text processing (TP). We explore several ways of building a DNN-HMM ASR system using out-of-domain AMI speech corpus plus a limited amount of ST data. In development experiments on the initial ST data, our final ASR system achieved the word-error-rate (WER) of 12.00%, compared to 14.89% for the official ST baseline DNN-HMM system. The WER of 9.28% was achieved on the test set data. For TP component, we first post-process the ASR output to deal with hesitations and then pass this to a template-based grammar, which we expanded from the provided baseline. We also developed a TP system based on machine learning methods, which enables to better accommodate variability of spoken language. We also fused outputs from several systems using a linear logistic regression. Our best system submitted to the challenge achieved F-measure of 0.914, D of 10.764 and D_{full} score of 5.691 on the final test set.

## Improvements to an Automated Content Scoring System for Spoken CALL Responses: the ETS Submission to the Second Spoken CALL Shared Task

*Keelan Evanini[1], Matthew Mulholland[1], Rutuja Ubale[2], Yao Qian[2], Robert Pugh[2], Vikram Ramanarayanan[2] and Aoife Cahill[1]*

[1]Educational Testing Service R&D, 660 Rosedale Rd., Princeton, NJ, USA

[2]Educational Testing Service R&D, 90 New Montgomery Street, Suite 1500, San Francisco, CA, USA

**Wed-SS-2-1-6, Time: 16:02-16:20**

This paper describes the details of the ETS submission to the 2018 Spoken CALL Shared Task. We employed a system using word and character n-gram features in a random forest machine learning framework based on the system that achieved the second-highest score in the text processing track of the 2017 Spoken CALL Shared Task. This system was augmented with additional features based on comparing the learner's responses to language models trained on text written by both native English speakers and L1-German English learners. In addition, we developed a set of sequence-to-label models using bidirectional LSTM-RNNs with an attention layer. The RNN model predictions were combined with the other feature sets using feature-level and score-level fusion approaches resulting in a best-performing system that achieved a D score of 7.397 on the test set (ranking 5th out of 12 submissions to the text processing track of the Shared Task). Subsequent experiments resulted in higher D scores when the model parameters were optimized for D score instead of F-score and the paper presents an error

**Notes**

analysis of these models in an attempt to determine which metric is more appropriate for evaluating spoken CALL systems.

## Discussion

**Wed-SS-2-1-7, Time: 16:20-16:30**

## Wed-S&T-2-1: Show and Tell 6

MR G.05-G.06; 14:30-16:30; Wednesday, 5 September, 2018
Chairs: Noelia do Carmo Blanco and Chitralekha Bhat

### Extracting Speaker's Gender, Accent, Age and Emotional State from Speech

*Nagendra Goel, Mousmita Sarma, Tejendra Kushwah, Dharmesh Agarwal, Zikra Iqbal and Surbhi Chauhan*
Go-Vivace Inc., McLean, VA, USA

**Wed-S&T-2-1-1, Time: 14:30-16:30**

We demonstrate a speaker characteristics assessment solution to extract speaker's information like gender, age, emotion, language and accent from telephone quality speech. The solution has been designed using machine learning algorithms ranging from Gaussian mixture models to deep neural networks and utilize websocket technology for real-time bidirectional interface to provide live updates in a scalable manner. The service is utilized on our demonstration web-page where user can upload or record audio file and obtain the speaker's characteristics. Such speaker characteristics information can be used as metadata in many real life applications designed for an emotionally sensitive human to machine interaction and human to human interaction.

### Determining Speaker Location from Speech in a Practical Environment

*BHVS Narayanamurthy[1], JV Satyanarayana[1] and Bayya Yegnanarayana[2]*
[1]Research Center Imarat, Hyderabad
[2]INSA Senior Scientist, IIIT-Hyderabad

**Wed-S&T-2-1-2, Time: 14:30-16:30**

The objective of the study is to show that a speaker's location in a practical environment can be obtained from the time delays
of the speech received at spatially distributed microphones. The time delay at a pair of microphones is estimated reliably
using a recently proposed single frequency filtering (SFF) analysis of speech even when the speech collected in a live room is
degraded due to echoes, reverberation and audio signals from other sources. The reliability is due to evidence of time delay from multiple frequency components obtained in the SFF analysis. The effectiveness of the proposed method for determining the speaker location can be demonstrated using a pair of microphones for picking up the speech signals and then processing the signals using SFF analysis.

### An Automatic Speech Transcription System for Manipuri Language

*Tanvina Patel, Krishna D N, Noor Fathima, Nisar Shah, Mahima C, Deepak Kumar and Anuroop Iyengar*
Cogknit Semantics, Bangalore, India

**Wed-S&T-2-1-3, Time: 14:30-16:30**

Development of speech technologies in Indian languages has witnessed a steep improvement recently. In this work, we present our efforts in building various speech technology applications for Manipuri language. For the language at hand, we initially perform Language identification (LID) task. This is followed by speech-to-text (STT) and Keyword Search (KWS). In addition, we build a Speaker Diarization (SD) framework as well. The speech modules are integrated together to extract information from the speech signal. Currently, the platform is build for Manipuri and English language and can be extended to other languages as well. A visual User Interface (UI) is available for demonstration purpose where given a set of speech files the services from all the mentioned speech modules can be used.

### SPIRE-SST: An Automatic Web-based Self-learning Tool for Syllable Stress Tutoring (SST) to the Second Language Learners

*Chiranjeevi Yarra[1], Anand P A[2], Kausthubha N K[1] and Prasanta Kumar Ghosh[1]*
[1]Electrical Engineering, Indian Institute of Science (IISc), Bangalore-560012, India
[2]National Institute of Technology Karnataka (NITK), Surathkal-575025, India

**Wed-S&T-2-1-4, Time: 14:30-16:30**

Correct stress placement on the syllables in a word or word groups is important in the spoken communication. Thus, incorrect syllable stress, typically made by second language (L2) learners, could result in miscommunication. In this demo, we present SPIRE-SST tool that tutors to learn correct stress patterns in a self-learning manner. Thus, the proposed tool could also benefit the learners without any access to the effective training methods. For this, we design a front-end containing self-explanatory instructions that can be easily followed by the user. Using the front-end, learners can submit their audio to the back-end and can view the corresponding feedback from the back-end. In the back-end, we divide the entire audio from the learner into syllable segments and detect each syllable as stressed or unstressed. Using these stress markings, we compute a score representing the stress quality in comparison with the ground-truth stress markings and send it to the front-end as a feedback. We also send a set of three features by comparing the audio from the expert and learner as the feedback, which we assume to be useful for correcting the pronunciation errors.

### Glotto Vibrato Graph: A Device and Method for Recording, Analysis and Visualization of Glottal Activity

*Kishalay Chakraborty[1], Senjam Shantirani Devi[1], Sanjeevan Devnath[1], S R Mahadeva Prasanna[1,2] and Priyankoo Sarmah[1,3]*
[1]DFM InfoAnalytics Private Limited, India
[2]Indian Institute of Technology Dharwad, India
[3]Indian Institute of Technology Guwahati, India

**Wed-S&T-2-1-5, Time: 14:30-16:30**

Opening and closing of the glottis is the primary and the most crucial step in converting airstream into speech sounds. The opening and closing process results in glottal vibration, filtered through the oral and nasal cavities to produce a variety of speech sounds. In this work we demonstrate the working of Glotto Vibrato Graph (GVG), that records, analyzes and visualizes the glottal activity with the help of a single low-cost hardwaresoftware package. The measurements are carried out by using piezo electric sensors to detect glottal vibration. A modular graphical computer software based on custom algorithm is used for the analysis and visualization. Results from the low cost GVG are comparable to results obtained from proprietary electroglottograph (EGG) devices.

**Notes**

## Wed-P-2-1: Adjusting to Speaker, Accent and Domain

Hall 4-6: Poster-1; 14:30-16:30; Wednesday, 5 September, 2018
Chair: Mikko Kurimo

### Multi-Modal Data Augmentation for End-to-End ASR

*Adithya Renduchintala, Shuoyang Ding, Matthew Wiesner and Shinji Watanabe*

Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD 21218, USA

**Wed-P-2-1-1, Time: 14:30-16:30**

We present a new end-to-end architecture for automatic speech recognition (ASR) that can be trained using symbolic input in addition to the traditional acoustic input. This architecture utilizes two separate encoders: one for acoustic input and another for symbolic input, both sharing the attention and decoder parameters. We call this architecture a multi-modal data augmentation network (MMDA), as it can support multi-modal (acoustic and symbolic) input and enables seamless mixing of large text datasets with significantly smaller transcribed speech corpora during training. We study different ways of transforming large text corpora into a symbolic form suitable for training our MMDA network. Our best MMDA setup obtains small improvements on character error rate (CER) and as much as 7-10% relative word error rate (WER) improvement over a baseline both with and without an external language model.

### Multi-task Learning with Augmentation Strategy for Acoustic-to-word Attention-based Encoder-decoder Speech Recognition

*Takafumi Moriya[1], Sei Ueno[1,2], Yusuke Shinohara[1], Marc Delcroix[3], Yoshikazu Yamaguchi[1] and Yushi Aono[1]*

[1]NTT Media Intelligence Laboratories, NTT Corporation, Japan

[2]Graduate School of Informatics, Kyoto University, Japan

[3]NTT Communication Science Laboratories, NTT Corporation, Japan

**Wed-P-2-1-2, Time: 14:30-16:30**

In this paper, we propose a novel training strategy for attention-based encoder-decoder acoustic-to-word end-to-end systems. Accuracy of end-to-end systems has greatly improved thanks to careful tuning of model structure and the introduction of novel training strategies to stabilize training. For example, multi-task learning using a shared-encoder is often used to escape from bad local optima. However, multi-task learning usually relies on a linear interpolation of the losses for each sub-task and consequently, the shared-encoder is not optimized for each task. To solve the above problem, we propose a multi-task learning with augmentation strategy. We augment the training data by creating multiple copies of the original training data to suit different output targets associated with each sub-task. We use each target loss sequentially to update the parameters of the shared-encoder so as to enhance the versatility of capturing acoustic features. This strategy enables better learning of the shared-encoder as each task is trained with a dedicated loss. The parameters of the word-decoder are jointly updated via the shared-encoder when optimizing the word prediction task loss. We evaluate our proposal on various speech data sets and show that our models achieve lower word error rates than both single-task and conventional multi-task approaches.

### Training Augmentation with Adversarial Examples for Robust Speech Recognition

*Sining Sun[1], Ching-Feng Yeh[2], Mari Ostendorf[3], Mei-Yuh Hwang2 and Lei Xie[1]*

[1]School of Computer Science, Northwestern Polytechnical University, Xi'an, China

[2]Mobvoi AI Lab, Seattle, USA

[3]Department of Electrical Engineering, University of Washington, Seattle, USA

**Wed-P-2-1-3, Time: 14:30-16:30**

This paper explores the use of adversarial examples in training speech recognition systems to increase robustness of deep neural network acoustic models. During training, the fast gradient sign method is used to generate adversarial examples augmenting the original training data. Different from conventional data augmentation based on data transformations, the examples are dynamically generated based on current acoustic model parameters. We assess the impact of adversarial data augmentation in experiments on the Aurora-4 and CHiME-4 single-channel tasks, showing improved robustness against noise and channel variation. Further improvement is obtained when combining adversarial examples with teacher/student training, leading to a 23% relative word error rate reduction on Aurora-4.

### Data Augmentation Improves Recognition of Foreign Accented Speech

*Takashi Fukuda[1], Raul Fernandez[1], Andrew Rosenberg[1], Samuel Thomas[1], Bhuvana Ramabhadran[2], Alexander Sorin[1] and Gakuto Kurata[1]*

[1]IBM Research AI

[2]Google

**Wed-P-2-1-4, Time: 14:30-16:30**

Speech recognition of foreign accented (non-native or L2) speech remains a challenge to the state-of-the-art. The most common approach to address this scenario involves the collection and transcription of accented speech and incorporating this into the training data. However, the amount of accented data is dwarfed by the amount of material from native (L1) speakers, limiting the impact of the additional material. In this work, we address this problem via data augmentation. We create modified copies of two accents, Latin American and Asian accented English speech with voice transformation (modifying glottal source and vocal tract parameters), noise addition and speed modification. We investigate both supervised (where transcription of the accented data is available) and unsupervised approaches to using the accented data and associated augmentations. We find that all augmentations provide improvements, with the largest gains coming from speed modification, then voice transformation and noise addition providing the least improvement. The improvements from training accent specific models with the augmented data are substantial. Improvements from supervised and unsupervised adaptation (or training with soft labels) with the augmented data are relatively minor. Overall, we find speed modification to be a remarkably reliable data augmentation technique for improving recognition of foreign accented speech. Our strategies with associated augmentations provide Word Error Rate (WER) reductions of up to 30% relative over a baseline trained with only the accented data.

### Speaker Adaptive Training and Mixup Regularization for Neural Network Acoustic Models in Automatic Speech Recognition

*Natalia Tomashenko[1,2], Yuri Khokhlov[3] and Yannick Estève[1]*

[1]LIUM, University of Le Mans, France

[2]ITMO University, Saint-Petersburg, Russia

[3]STC-Innovations Ltd, Saint-Petersburg, Russia

**Wed-P-2-1-5, Time: 14:30-16:30**

This work investigates speaker adaptation and regularization techniques for deep neural network acoustic models (AMs) in automatic speech recognition (ASR) systems. In previous works, GMM-derived (GMMD) features have been shown to be an efficient technique for neural network AM adaptation. In this paper, we propose and investigate a novel way to improve speaker adaptive training (SAT) for neural network AMs using GMMD features. The idea is based on using inaccurate transcriptions from ASR for adaptation during neural network training, while keeping the exact transcriptions for targets of neural networks. In addition, we apply a mixup technique, recently proposed for classification tasks, to acoustic models for ASR and investigate the impact of this technique on speaker adapted acoustic models. Experimental results on the TED-LIUM corpus show that the proposed approaches provide an additional gain in speech recognition performance in comparison with the speaker adapted AMs.

**Notes**

## Neural Language Codes for Multilingual Acoustic Models

*Markus Müller[1], Sebastian Stüker[1] and Alex Waibel[1,2]*

[1]Karlsruhe Institute of Technology, Karlsruhe, Germany

[2]Carnegie Mellon University, Pittsburgh PA, USA

**Wed-P-2-1-6, Time: 14:30-16:30**

Multilingual Speech Recognition is one of the most costly AI problems, because each language (7,000+) and even different accents require their own acoustic models to obtain best recognition performance. Even though they all use the same phoneme symbols, each language and accent imposes its own coloring or "twang". Many adaptive approaches have been proposed, but they require further training, additional data and generally are inferior to monolingually trained models. In this paper, we propose a different approach that uses a large multilingual model that is modulated by the codes generated by an ancillary network that learns to code useful differences between the "twangs" or human language.

We use Meta-Pi networks to have one network (the language code net) gate the activity of neurons in another (the acoustic model nets). Our results show that during recognition multilingual Meta-Pi networks quickly adapt to the proper language coloring without retraining or new data and perform better than monolingually trained networks. The model was evaluated by training acoustic modeling nets and modulating language code nets jointly and optimize them for best recognition performance.

## Encoder Transfer for Attention-based Acoustic-to-word Speech Recognition

*Sei Ueno[1,2], Takafumi Moriya[1], Masato Mimura[2], Shinsuke Sakai[2], Yusuke Shinohara[1], Yoshikazu Yamaguchi[1], Yushi Aono[1] and Tatsuya Kawahara[2]*

[1]NTT Media Intelligence Laboratories, NTT Corporation, Japan

[2]Graduate School of Informatics, Kyoto University, Sakyo-ku, Kyoto, Japan

**Wed-P-2-1-7, Time: 14:30-16:30**

Acoustic-to-word speech recognition based on attention-based encoder-decoder models achieves better accuracies with much lower latency than the conventional speech recognition systems. However, acoustic-to-word models require a very large amount of training data and it is difficult to prepare one for a new domain such as elderly speech. To address the problem, we propose domain adaptation based on transfer learning with layer freezing. Layer freezing first pre-trains a network with the source domain data and then a part of parameters is re-trained for the target domain while the rest is fixed. In the attention-based acoustic-to-word model, the encoder part is frozen to maintain the generality and only the decoder part is re-trained to adapt to the target domain. This substantially allows for adaptation of the latent linguistic capability of the decoder to the target domain. Using a large-scale Japanese spontaneous speech corpus as source, the proposed method is applied to three target domains: a call center task and two voice search tasks by adults and by elderly. The models trained with the proposed method achieved better accuracy than the baseline models, which are trained from scratch or entirely re-trained with the target domain.

## Empirical Evaluation of Speaker Adaptation on DNN Based Acoustic Model

*Ke Wang[1,2], Junbo Zhang[2], Yujun Wang[2] and Lei Xie[1]*

[1]Shaanxi Provincial Key Laboratory of Speech and Image Information Processing, School of Computer Science, Northwestern Polytechnical University, Xi'an, China

[2]Xiaomi, Beijing, China

**Wed-P-2-1-8, Time: 14:30-16:30**

Speaker adaptation aims to estimate a speaker specific acoustic model from a speaker independent one to minimize the mismatch between the training and testing conditions arisen from speaker variabilities. A variety of neural network adaptation methods have been proposed since deep learning models have become the main stream. But there still lacks an experimental comparison between different methods, especially when DNN-based acoustic models have been advanced greatly. In this paper, we aim to close this gap by providing an empirical evaluation of three typical speaker adaptation methods: LIN, LHUC and KLD. Adaptation experiments, with different size of adaptation data, are conducted on a strong TDNN-LSTM acoustic model. More challengingly, here, the source and target we are concerned with are standard Mandarin speaker model and accented Mandarin speaker model. We compare the performances of different methods and their combinations. Speaker adaptation performance is also examined by speaker's accent degree.

## Improving DNNs Trained with Non-Native Transcriptions Using Knowledge Distillation and Target Interpolation

*Amit Das and Mark Hasegawa-Johnson*

University of Illinois, USA

**Wed-P-2-1-9, Time: 14:30-16:30**

Often, it is quite hard to find native transcribers in under-resourced languages. However, Turkers (crowd workers) available in online marketplaces can serve as valuable alternative resources by providing transcriptions in the target language. Since the Turkers may neither speak nor have any familiarity with the target language, their transcriptions are non-native by nature and are usually filled with incorrect labels. After some post-processing, these transcriptions can be converted to Probabilistic Transcriptions (PT). Conventional Deep Neural Networks (DNN) trained using PTs do not necessarily improve error rates over Gaussian Mixture Models (GMMs) due to the presence of label noise. Previously reported results have demonstrated some success by adopting Multi-Task Learning (MTL) training for PTs. In this study, we report further improvements using Knowledge Distillation (KD) and Target Interpolation (TI) to alleviate transcription errors in PTs. In the KD method, knowledge is transfered from a well-trained multilingual DNN to the target language DNN trained using PTs. In the TI method, the confidences of the labels provided by PTs are modified using the confidences of the target language DNN. Results show an average absolute improvement in phone error rates (PER) by about 1.9% across Swahili, Amharic, Dinka and Mandarin using each proposed method.

## Improving Cross-Lingual Knowledge Transferability Using Multilingual TDNN-BLSTM with Language-Dependent Pre-Final Layer

*Siyuan Feng and Tan Lee*

Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong

**Wed-P-2-1-10, Time: 14:30-16:30**

Multilingual acoustic modeling for improved automatic speech recognition (ASR) has been extensively researched. It's widely acknowledged that the shared-hidden-layer multilingual deep neural network (SHL-MDNN) acoustic model (AM) could outperform the conventional monolingual AM, due to its effectiveness in cross-lingual knowledge transfer. In this work, two research aspects are investigated, with the goal of improving multilingual acoustic modeling. Firstly, in the SHL-MDNN architecture, the shared hidden layer configuration is replaced by a combined TDNN-BLSTM structure. Secondly, the improvement of cross-lingual knowledge transferability is achieved through adding the proposed language-dependent pre-final layer under each network output. The pre-final layer, rarely adopted in past works, is expected to increase nonlinear modeling capability between universal transformed features generated by shared hidden layers and language-specific outputs. Experiments are carried out with CUSENT, WSJ and RASC-863 corpora, covering Cantonese, English and Mandarin. A Cantonese ASR task is chosen for evaluation. Experimental results show that SHL-MTDNN-BLSTM achieves the best performance. The proposed additional language-dependent pre-final layer brings moderate while consistent performance gains in various multilingual training corpora settings, thus demonstrates its effectiveness in improving cross-lingual knowledge transferability.

**Notes**

### Auxiliary Feature Based Adaptation of End-to-End ASR Systems

*Marc Delcroix[1], Shinji Watanabe[2], Atsunori Ogawa[1], Shigeki Karita[1] and Tomohiro Nakatani[1]*

[1]NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan

[2]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA

**Wed-P-2-1-11, Time: 14:30-16:30**

Acoustic model adaptation has been widely used to adapt models to speakers or environments. For example, appending auxiliary features representing speakers such as i-vectors to the input of a deep neural network (DNN) is an effective way to realize unsupervised adaptation of DNN-hybrid automatic speech recognition (ASR) systems. Recently, end-to-end (E2E) models have been proposed as an alternative to conventional DNN-hybrid ASR systems. E2E models map a speech signal to a sequence of characters or words using a single neural network, which greatly simplifies the ASR pipeline. However, adaptation of E2E models has received little attention yet. In this paper, we investigate auxiliary feature based adaptation for encoder-decoder E2E models. We employ a recently proposed sequence summary network to compute auxiliary features instead of i-vectors, as it can be easily integrated into E2E models and keep the ASR pipeline simple. Indeed, the sequence summary network allows the auxiliary feature extraction module to be a part of the computational graph of the E2E model. We demonstrate that the proposed adaptation scheme consistently improves recognition performance of three publicly available recognition tasks.

### Leveraging Native Language Information for Improved Accented Speech Recognition

*Shahram Ghorbani and John H.L. Hansen*

Center for Robust Speech Systems (CRSS), University of Texas at Dallas, Richardson, TX 75080

**Wed-P-2-1-12, Time: 14:30-16:30**

Recognition of accented speech is a long-standing challenge for ASR systems, given the increasing worldwide population of bi-lingual speakers with English as their second language. If we consider foreign-accented speech as an interpolation of the native language(L1) and English(L2), using a model that can simultaneously recognize both languages would perform better at the acoustic level for accented speech. In this study, we explore how an end-to-end recurrent neural network (RNN) trained system with English and native languages (Spanish and Indian languages) could leverage the data of native languages to perform better for accented English speech. To this end, we examine using pre-training with native languages, as well as multitask learning in which the main task is trained with native English data and the secondary task is trained with Spanish or Indian Languages. We show that the multitask setting performs better than the former approach. We suggest a new setting for multitask learning in which the secondary task is trained with both English and the native language, using the same output set. This proposed scenario yields better performance than the first setting which provides +11.95% and +17.55% character error rate (CER) gain over the baseline, for Hispanic and Indian accents, respectively.

### Improved Accented Speech Recognition Using Accent Embeddings and Multi-task Learning

*Abhinav Jain, Minali Upreti and Preethi Jyothi*

Department of Computer Science and Engineering, Indian Institute of Technology Bombay, India

**Wed-P-2-1-13, Time: 14:30-16:30**

One of the major remaining challenges in modern automatic speech recognition (ASR) systems for English is to be able to handle speech from users with a diverse set of accents. ASR systems that are trained on speech from multiple English accents still underperform when confronted with a new speech accent. In this work, we explore how to use accent embeddings and multi-task learning to improve speech recognition for accented speech. We propose a multi-task architecture that jointly learns an accent classifier and a multi-accent acoustic model. We also consider augmenting the speech input with accent information in the form of embeddings extracted by a separate network. These techniques together give significant relative performance improvements of 15% and 10% over a multi-accent baseline system on test sets containing seen and unseen accents, respectively.

### Fast Language Adaptation Using Phonological Information

*Sibo Tong[1,2], Philip N. Garner[1] and Hervé Bourlard [1,2]*

[1]Idiap Research Institute, Martigny, Switzerland

[2]Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

**Wed-P-2-1-14, Time: 14:30-16:30**

Phoneme-based multilingual connectionist temporal classification (CTC) model is easily extensible to a new language by concatenating parameters of the new phonemes to the output layer. In the present paper, we improve cross-lingual adaptation in the context of phoneme-based CTC models by using phonological information. A universal (IPA) phoneme classifier is first trained on phonological features generated from a phonological attribute detector. When adapting the multilingual CTC to a new, never seen, language, phonological attributes of the unseen phonemes are derived based on phonology and fed into the phoneme classifier. Posteriors given by the classifier are used to initialize the parameters of the unseen phonemes when extending the multilingual CTC output layer to the target language. Adaptation experiments show that the proposed initialization approaches further improve the cross-lingual adaptation on CTC models and yield significant improvements over Deep Neural Network / Hidden Markov Model (DNN/HMM)-based adaptation using limited data.

## Wed-P-2-2: Speech Synthesis Paradigms and Methods

Hall 4-6: Poster-2; 14:30-16:30; Wednesday, 5 September, 2018

Chair: Masami Akamine

### Naturalness Improvement Algorithm for Reconstructed Glossectomy Patient's Speech Using Spectral Differential Modification in Voice Conversion

*Hiroki Murakami[1], Sunao Hara[1], Masanobu Abe[1], Masaaki Sato[2] and Shogo Minagi[2]*

[1]Graduate School of Natural Science and Technology, Okayama University, Japan

[2]Graduate School of Medicine Dentistry and Pharmaceutical Sciences, Okayama University, Japan

**Wed-P-2-2-1, Time: 14:30-16:30**

In this paper, we propose an algorithm to improve the naturalness of the reconstructed glossectomy patient's speech that is generated by voice conversion to enhance the intelligibility of speech uttered by patients with a wide glossectomy. While existing VC algorithms make it possible to improve intelligibility and naturalness, the result is still not satisfying. To solve the continuing problems, we propose to directly modify the speech waveforms using a spectrum differential. The motivation is that glossectomy patients mainly have problems in their vocal tract, not in their vocal cords. The proposed algorithm requires no source parameter extractions for speech synthesis, so there are no errors in source parameter extractions and we are able to make the best use of the original source characteristics. In terms of spectrum conversion, we evaluate with both GMM and DNN. Subjective evaluations show that our algorithm can synthesize more natural speech than the vocoder-based method. Judging from observations of the spectrogram, power in high-frequency bands of fricatives and stops is reconstructed to be similar to that of natural speech.

**Notes**

## Audio-visual Voice Conversion Using Deep Canonical Correlation Analysis for Deep Bottleneck Features

*Satoshi Tamura[1], Kento Horio[1], Hajime Endo[1], Satoru Hayamizu[1] and Tomoki Toda[2]*

[1]Gifu University, Japan

[2]Nagoya University, Japan

**Wed-P-2-2-2, Time: 14:30-16:30**

This paper proposes Audio-Visual Voice Conversion (AVVC) methods using Deep BottleNeck Features (DBNF) and Deep Canonical Correlation Analysis (DCCA). DBNF has been adopted in several speech applications to obtain better feature representations. DCCA can generate much correlated features in two views and enhance features in one modality based on another view. In addition, DCCA can make projections from different views ideally to the same vector space. Firstly, in this work, we enhance our conventional AVVC scheme by employing the DBNF technique in the visual modality. Secondly, we apply the DCCA technology to DBNFs for new effective visual features. Thirdly, we build a cross-modal voice conversion model available for both audio and visual DCCA features. In order to clarify effectiveness of these frameworks, we carried out subjective and objective evaluations and compared them with conventional methods. Experimental results show that our DBNF- and DCCA-based AVVC can successfully improve the quality of converted speech waveforms.

## An Investigation of Convolution Attention Based Models for Multilingual Speech Synthesis of Indian Languages

*Pallavi Baljekar, SaiKrishna Rallabandi and Alan W Black*

Carnegie Mellon University

**Wed-P-2-2-3, Time: 14:30-16:30**

In this paper we investigate multi-speaker, multi-lingual speech synthesis for 4 Indic languages (Hindi, Marathi, Gujarathi, Bengali) as well as English in a fully convolutional attention based model. We show how factored embeddings can allow cross lingual transfer and investigate methods to adapt the model in a low resource scenario for the case of Marathi and Gujarati. We also show results on how effectively the model scales to a new language and how much data is required to train the system on a new language.

## The Effect of Real-Time Constraints on Automatic Speech Animation

*Danny Websdale, Sarah Taylor and Ben Milner*

University of East Anglia

**Wed-P-2-2-4, Time: 14:30-16:30**

Machine learning has previously been applied successfully to speech-driven facial animation. To account for carry-over and anticipatory coarticulation a common approach is to predict the facial pose using a symmetric window of acoustic speech that includes both past and future context. Using future context limits this approach for animating the faces of characters in real-time and networked applications, such as online gaming. An acceptable latency for conversational speech is 200ms and typically network transmission times will consume a significant part of this. Consequently, we consider asymmetric windows by investigating the extent to which decreasing the future context effects the quality of predicted animation using both deep neural networks (DNNs) and bi-directional LSTM recurrent neural networks (BiLSTMs). Specifically we investigate future contexts from 170ms (fully-symmetric) to 0ms (fully-asymmetric). We find that a BiLSTM trained using 70ms of future context is able to predict facial motion of equivalent quality as a DNN trained with 170ms, while introducing increased processing time of only 5ms. Subjective tests using the BiLSTM show that reducing the future context from 170ms to 50ms does not significantly decrease perceived realism. Below 50ms, the perceived realism begins to deteriorate, generating a trade-off between realism and latency.

## Joint Learning of Facial Expression and Head Pose from Speech

*David Greenwood, Iain Matthews and Stephen Laycock*

School of Computing Science, University of East Anglia, United Kingdom

**Wed-P-2-2-5, Time: 14:30-16:30**

Natural movement plays a significant role in realistic speech animation and numerous studies have demonstrated the contribution visual cues make to the degree human observers find an animation acceptable. Natural, expressive, emotive and prosodic speech exhibits motion patterns that are difficult to predict with considerable variation in visual modalities. Recently, there have been some impressive demonstrations of face animation derived in some way from the speech signal. Each of these methods have taken unique approaches, but none have included rigid head pose in their predicted output.

We observe a high degree of correspondence with facial activity and rigid head pose during speech and exploit this observation to jointly learn full face animation and head pose rotation and translation combined. From our own corpus, we train Deep Bi-Directional LSTMs (BLSTM) capable of learning long-term structure in language to model the relationship that speech has with the complex activity of the face. We define a model architecture to encourage learning of rigid head motion via the latent space of the speaker's facial activity. The result is a model that can predict lip sync and other facial motion along with rigid head motion directly from audible speech.

## Acoustic-dependent Phonemic Transcription for Text-to-speech Synthesis

*Kévin Vythelingum[1,2], Yannick Estève[2] and Olivier Rosec[1]*

[1]Voxygen, Pleumeur-Bodou, France

[2]LIUM, Le Mans University, France

**Wed-P-2-2-6, Time: 14:30-16:30**

Text-to-speech synthesis (TTS) purpose is to produce a speech signal from an input text. This implies the annotation of speech recordings with word and phonemic transcriptions. The overall quality of TTS highly depends on the accuracy of phonemic transcriptions. However, they are generally automatically produced by grapheme-to-phoneme conversion systems, which don't deal with speaker variability. In this work, we explore ways to obtain signal-dependent phonemic transcriptions. We investigate forced-alignment with enriched pronunciation lexicon and multimodal phonemic transcription. We then apply our results on error detection of grapheme-to-phoneme conversion hypotheses in order to find where the phonemic transcriptions may be erroneous. On a French TTS dataset, we show that we can detect up to 90.5% of errors of a state-of-the-art grapheme-to-phoneme conversion system by annotating less than 15.8% of phonemes as erroneous. This can help a human annotator to correct most of grapheme-to-phoneme conversion errors without checking a lot of data. In other words, our method can significantly reduce the cost of high quality TTS data creation.

## Multimodal Speech Synthesis Architecture for Unsupervised Speaker Adaptation

*Hieu-Thi Luong[1] and Junichi Yamagishi[1,2]*

[1]National Institute of Informatics, Tokyo, Japan

[2]University of Edinburgh, Edinburgh, UK

**Wed-P-2-2-7, Time: 14:30-16:30**

This paper proposes a new architecture for speaker adaptation of multi-speaker neural-network speech synthesis systems in which an unseen speaker's voice can be synthesized using a relatively small amount of speech data without transcriptions for adaptation. This is sometimes called "unsupervised speaker adaptation". More specifically, we concatenate the layers to the audio inputs when performing unsupervised speaker adaptation while we concatenate them to the text inputs when synthesizing speech from a text. Two new training schemes for this new architecture are also proposed in this paper. These training schemes are not limited to speech synthesis; other applications are suggested. Experimental results show that the proposed model not only enables adaptation to unseen speakers using untranscribed speech but it also improves the performance of multi-speaker modeling and speaker adaptation using transcribed audio files.

**Notes**

## Articulatory-to-speech Conversion Using Bi-directional Long Short-term Memory

*Fumiaki Taguchi1 and Tokihiko Kaburagi2*

[1]Graduate School of Design,

[2]Faculty of Design, Kyushu University Shiobaru 4–9–1, Minami-ku, Fukuoka-shi, Fukuoka,815–8540 Japan

**Wed-P-2-2-8, Time: 14:30-16:30**

Methods for synthesizing speech sounds from the motion of articulatory organs can be used to produce substitute speech for people who have undergone laryngectomy. To achieve this goal, feature parameters representing the spectral envelope of speech, directly related to the acoustic characteristics of the vocal tract, has been estimated from articulatory movements. Within this framework, speech can be synthesized by driving the filter obtained from a spectral envelope with noise signals. In the current study, we examined an alternative method that generates speech sounds directly from the motion pattern of articulatory organs based on the implicit relationships between articulatory movements and the source signal of speech. These implicit relationships were estimated by considering that articulatory movements are involved in phonological representations of speech that are also related to sound source information such as the temporal pattern of pitch and voiced/unvoiced flag. We developed a method for simultaneously estimating the spectral envelope and sound source parameters from articulatory data obtained with an electromagnetic articulography (EMA) sensor. Furthermore, objective evaluation of estimated speech parameters and subjective evaluation of the word error rate were performed to examine the effectiveness of our method.

## Implementation of Respiration in Articulatory Synthesis Using a Pressure-Volume Lung Model

*Keisuke Tanihara, Shogo Yonekura and Yasuo Kuniyoshi*

Graduate School of Information Science and Technology, The University of Tokyo, Japan

**Wed-P-2-2-9, Time: 14:30-16:30**

In previous studies of the 1D vocal tract model of articulatory synthesis, subglottal pressure is typically regarded as constant, ignoring its dynamics. However, human vocalization is initially generated by glottal airflow via subglottal pressure change. This change is caused by the expansion and contraction of the lungs.

In the current study, we propose a new pressure-volume model that relates pressure changes to volume changes of the human lung. Using this model, the behavior of the human lung can be integrated with articulatory synthesis. This model produces positive and negative subglottal pressure corresponding to expiration and inspiration respectively. In addition, breathing could be implemented in the proposed model. This implementation would expand the possibilities for articulatory synthesis.

## Learning and Modeling Unit Embeddings for Improving HMM-based Unit Selection Speech Synthesis

*Xiao Zhou, Zhen-Hua Ling, Zhi-Ping Zhou and Li-Rong Dai*

National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei, P.R. China

**Wed-P-2-2-10, Time: 14:30-16:30**

This paper presents a method of learning and modeling unit embeddings using deep neutral networks (DNNs) to improve the performance of HMM-based unit selection speech synthesis. First, a DNN with an embedding layer is built to learn a fixed-length embedding vector for each phone-sized candidate unit in the corpus from scratch. Then, another two DNNs are constructed to map linguistic features toward the extracted unit vector of each phone. One of them employs the unit vectors of preceding phones as model input. At synthesis time, the L2 distances between the unit vectors predicted by these two DNNs and the ones derived from candidate units are integrated into the target cost and the concatenation cost of HMM-based unit selection speech synthesis respectively. Experimental results demonstrate that the unit vectors estimated using only acoustic features display phone-dependent clustering properties. Furthermore, integrating unit vector distances into cost functions, especially the concatenation cost, improves the naturalness of HMM-based unit selection speech synthesis in our experiments.

## Deep Metric Learning for the Target Cost in Unit-Selection Speech Synthesizer

*Ruibo Fu[1,2], Jianhua Tao[1,2,3], Yibin Zheng[1,2] and Zhengqi Wen[1]*

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China

[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

[3]CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China

**Wed-P-2-2-11, Time: 14:30-16:30**

This paper describes a unified Deep Metric Learning (DML) framework to predict the target cost directly by supervised learning method. The conventional methods to calculate the target cost include two separate steps: feature extraction and standard distance measurement. The proposed DML framework aims to measure the similarity between the candidate units and the target units more reasonably and directly. Firstly, the symmetrical DML framework is pre-trained to learn the metric between pairs of candidate units and target units. The relabeling procedure is added to correct the initial designed labels of the target cost. Secondly, the acoustic features of the target units are removed, which fits the runtime of the unit-selection synthesizer. The asymmetrical DML is fine-tuned to learn the metric between candidate units and target units. Compared with the conventional methods, the proposed unified DML framework can avoid the accumulation of errors in separate steps and improve the accuracy in labeling and predicting the target cost. The evaluation results demonstrate that the naturalness of synthetic speech has been improved by adopting DML framework to predict target cost.

## DNN-based Speech Synthesis for Small Data Sets Considering Bidirectional Speech-Text Conversion

*Kentaro Sone and Toru Nakashika*

Graduate School of Informatics and Engineering, The University of Electro-Communications, Japan

**Wed-P-2-2-12, Time: 14:30-16:30**

In statistical parametric speech synthesis, approaches based on deep neural networks (DNNs) have improved qualities of the synthesized speech. General DNN-based approaches require a large amount of training data to synthesize natural speech. However, it is not practical to record speech for many hours from a single speaker. To address this problem, this paper presents a novel pre-training method of DNN-based speech synthesis systems for small data sets. In this method, a Gaussian-Categorical deep relational model (GCDRM), which represents a joint probability of two visible variables, is utilized to describe the joint distribution of acoustic features and linguistic features. During the maximum-likelihood-based training, the model attempts to obtain parameters of a deep architecture considering the bidirectional conversion between 1) generated acoustic features given linguistic features and 2) re-generated linguistic features given acoustic features generated from itself. Owing to considering whether the generated acoustic features are recognizable, our method can obtain reasonable parameters from small data sets. Experimental results show that pre-trained DNN-based systems using our proposed method outperformed randomly-initialized DNN-based systems. This method also outperformed DNN-based systems in a speaker-dependent speech recognition task.

## A Weighted Superposition of Functional Contours Model for Modelling Contextual Prominence of Elementary Prosodic Contours

*Branislav Gerazov[1,2], Gérard Bailly[2] and Yi Xu[3]*

[1]FEEIT, University of Ss. Cyril and Methodius in Skopje, Macedonia

[2]Univ. Grenoble-Alpes, CNRS, Grenoble-INP, GIPSA-lab, 38000

**Notes**

Grenoble, France
[3]Division of Psychology and Language Sciences, University College London, UK

**Wed-P-2-2-13, Time: 14:30-16:30**

The way speech prosody encodes linguistic, paralinguistic and non-linguistic information via multiparametric representations of the speech signals is still an open issue. The Superposition of Functional Contours (SFC) model proposes to decompose prosody into elementary multiparametric functional contours through the iterative training of neural network contour generators using analysis-by-synthesis. Each generator is responsible for computing multiparametric contours that encode one given linguistic, paralinguistic and non-linguistic information on a variable scope of rhythmic units. The contributions of all generators' outputs are then overlapped and added to produce the prosody of the utterance. We propose an extension of the contour generators that allows them to model the prominence of the elementary contours based on contextual information. WSFC jointly learns the patterns of the elementary multiparametric functional contours and their weights dependent on the contours' contexts. The experimental results show that the proposed weighted SFC (WSFC) model can successfully capture contour prominence and thus improve SFC modelling performance. The WSFC is also shown to be effective at modelling the impact of attitudes on the prominence of functional contours cuing syntactic relations in French and that of emphasis on the prominence of tone contours in Chinese.

### LSTBM: A Novel Sequence Representation of Speech Spectra Using Restricted Boltzmann Machine with Long Short-Term Memory

*Toru Nakashika*
The University of Electro-Communications

**Wed-P-2-2-14, Time: 14:30-16:30**

In this paper, we propose a novel probabilistic model, namely long short-term Boltzmann memory (LSTBM), to represent sequential data like speech spectra. The LSTBM is an extension of a restricted Boltzmann machine (RBM) that has generative long short-term memory (LSTM) units. The original RBM automatically learns relationships between visible and hidden units and is widely used as a feature extractor, a generator, a classifier, a pre-training method of deep neural networks, etc. However, the RBM is not sufficient to represent sequential data because it assumes that each frame from sequential data is completely independent of the others. Unlike conventional RBMs, the LSTBM has connections over time via LSTM units and represents time dependencies in sequential data. Our speech coding experiments demonstrated that the proposed LSTBM outperformed the other conventional methods: an RBM and a temporal RBM.

## Wed-P-2-3: Second Language Acquisition and Code-switching

Hall 4-6: Poster-3; 14:30-16:30; Wednesday, 5 September, 2018
Chair: Alejandrina Cristia

### Should Code-switching Models Be Asymmetric?

*Barbara E. Bullock[1], Gualberto Guzmán[1], Jacqueline Serigos[2] and Almeida Jacqueline Toribio[1]*
[1]University of Texas at Austin
[2]George Mason University

**Wed-P-2-3-1, Time: 14:30-16:30**

Since the work of Joshi [1], most models of code-switching (C-S) have assumed asymmetry of the participating languages. While there exist patterns of language mixing in which a dominant or matrix language (ML) may not be discernible, these more complex signatures are rarely modeled [2, 3]. We use a series of metrics to characterize the switching in corpora as asymmetrical (insertional C-S) or symmetrical (alternational C-S). We test the efficacy of a linguistic model that assumes no ML in predicting the syntax of C-S in three Spanish–English corpora that vary according to whether the ML is Spanish, English or indeterminate. Our results show that the same constraints on the grammatical junctures and on the directionality of switching hold irrespective of the symmetry of the data. The length of the alternating language spans varies according to POS with noun phrases comprising the shortest spans. This suggests that insertional C-S may be subsumed under alternational C-S, as spontaneous borrowing. These results invite researchers to reconsider the linguistic theories they adopt and to expand the typology of training data used in creating language models and processing tools for C-S.

### Cross-language Perception of Mandarin Lexical Tones by Mongolian-speaking Bilinguals in the Inner Mongolia Autonomous Region, China

*Kimiko Tsukada[1,2] and Yu Rong[3]*
[1]Macquarie University, Australia
[2]University of Oregon, USA
[3]Inner Mongolia University, China

**Wed-P-2-3-2, Time: 14:30-16:30**

Mandarin is a representative tonal language with four contrastive tone categories (Tone 1 (T1): high level (ā), Tone 2 (T2): high rising (á), Tone 3 (T3): dipping (ǎ), Tone 4 (T4): high falling (à)). Learning Mandarin tones is known to be difficult for speakers from diverse linguistic backgrounds. The purpose of this research was to examine how native Mongolian-speaking bilinguals perceive Mandarin lexical tones. The 24 (17 females, 7 males) participants studied Mandarin for 15 years on average in the Inner Mongolia Autonomous Region, China. A discrimination experiment was conducted to assess Mongolian bilinguals' perception of six tone pairs (T1-T2, T1-T3, T1-T4, T2-T3, T2-T4, T3-T4). The Mongolian group was less accurate than the control group of ten native Mandarin listeners for all six pairs and the between-group difference was particularly large for T2-T3. However, large individual variation was observed and some Mongolian bilinguals perceived Mandarin tones as accurately as native Mandarin listeners, suggesting that native-like tone perception is attainable in subsequently acquired languages.

### Automatically Measuring L2 Speech Fluency without the Need of ASR: A Proof-of-concept Study with Japanese Learners of French

*Lionel Fontan[1], Maxime Le Coz[1] and Sylvain Detey[2]*
[1]Archean LABS, Archean Technologies, Montauban, France
[2]Waseda University – SILS, Tokyo, Japan

**Wed-P-2-3-3, Time: 14:30-16:30**

This research work investigates the possibility of using automatic acoustic measures to assess speech fluency in the context of second language (L2) acquisition.

To this end, three experts rated speech recordings of Japanese learners of French who were instructed to read aloud a 21-sentence-long text. A Forward-Backward Divergence Segmentation (FBDS) algorithm was used to segment speech recordings (sentences) into acoustically homogeneous units at a subphonemic scale. The FBDS processing results were used — along with more classic measures such as raw percentage of speech and length/standard deviation of silent pauses — to estimate speech rate and regularity of speech rate, while a formant tracking algorithm was used to estimate speech fluidity (i.e., quality of coarticulation). A step-by-step multiple linear regression was finally computed to predict the experts' mean fluency ratings.

Results show that FBDS-derived measures, raw percentage of speech and standard deviation of the first formant curve derivative can be combined together to calculate accurate estimates of speakers' fluency scores ($R = .92$; $P < .001$). As only low-level signal features were used in the study, the method could also be relevant for the assessment of speakers of other target languages, as well as for the assessment of disordered speech.

**Notes**

## Analysis of L2 Learners' Progress of Distinguishing Mandarin Tone 2 and Tone 3

*Yue Sun[1], Win Thuzar Kyaw[1], Jinsong Zhang[2] and Yoshinori Sagisaka[1]*

[1]Graduate School of Fundamental Science and Engineering, Waseda University, Japan

[2]College of Information Science, Beijing Language and Culture University, China

**Wed-P-2-3-4, Time: 14:30-16:30**

Many studies have shown the effectiveness of perceptual training to improve L2 learners' ability to distinguish Mandarin tones. In this paper, we quantified learners perceptual characteristics on discriminating the most difficult tone pair, Tone 2 and 3 in Mandarin before and after training. L2 learners' categorical perception is measured by fitting a sigmoid curve to the identification responses with average F0 height be the acoustic dimension. The boundary location of the two tones in L2 learners' perception space is significantly improved to a higher F0 height after training. Regression analysis indicated that $\delta$F0 and $\delta$t of the initial falling of the concave F0 shape are the key acoustic features for native speakers in discrimination. L2 learners rely on not only the initial fall but also the $\delta$F0 of the final rise to discriminate the tones. A detailed analysis using cognitive measurements reports an increasing attention on the initial fall of the F0 contour for L2 learners after perceptual training. These results confirmed that directing the attention to key acoustic features is essential for L2 learners to improve their categorical perception of novel speech contrasts.

## Unsupervised Discovery of Non-native Phonetic Patterns in L2 English Speech for Mispronunciation Detection and Diagnosis

*Xu Li[1], Shaoguang Mao[2], Xixin Wu[1], Kun Li[3], Xunying Liu[1] and Helen Meng[1,2]*

[1]Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong

[2]Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University

[3]SpeechX Limited, Shenzhen

**Wed-P-2-3-5, Time: 14:30-16:30**

Second language (L2) speech is often annotated with the native phoneme categories. However, we often observe that an L2 speech segment generally deviates from a canonical phoneme and sometimes it is very difficult for linguists to annotate with any canonical phoneme label. We refer to these segments as non-native phonetic patterns. Existing approaches to mispronunciation detection and diagnosis (MDD) focus mainly on canonical mispronunciations, i.e. one canonical phoneme is substituted for another, aside from those deleted or inserted. To better represent L2 speech, this work explores non-native phonetic patterns (NN-PPs) of each native phoneme by an unsupervised approach. We apply an optimized k-means algorithm to cluster state-based phonemic posterior-grams, which are generated with a deep neural network. Then, to discover the NN-PPs related to each native phoneme, we perform forced alignment to divide L2 speech into segments grouped by native phonemes. We use the cluster sequences within segments derived from clustering results to represent different phonetic patterns of each native phoneme. Finally, we apply Cluster Sequence Analysis to discover each phoneme's potential NN-PPs. We verified experimentally that NN-PPs can extend the native phoneme categories to better describe L2 speech, which can enrich the existing approaches to MDD for better performance.

## Wuxi Speakers' Production and Perception of Coda Nasals in Mandarin

*Lei Wang, Jie Cui and Ying Chen*

School of Foreign Studies, Nanjing University of Science and Technology, China

**Wed-P-2-3-6, Time: 14:30-16:30**

Wuxi natives speak a dialect of Wu, which has only one coda nasal /n/ but allows allophones depending on the pre-nasal vowel (Qian, 1992), whereas in Mandarin, there are two coda nasals—alveolar /n/ and velar /ŋ/. Two perception experiments were conducted to investigate Wuxi speakers' perception and production of coda nasals in their second language (L2) Mandarin. First, two groups of Wuxi native speakers, age around 20 and 50, produced monosyllabic words with nasal coda in Mandarin and their production was used as the stimuli for native Mandarin speakers to identify. Second, the same Wuxi speakers participated in an identification task to judge the place of articulation of the nasal coda in monosyllabic words in standard Mandarin. The results of the first experiment indicate that young Wuxi speakers' Mandarin production was identified with higher accuracy by native Mandarin speakers than older Wuxi speakers', suggesting the young speakers produced more nativelike Mandarin than the older speakers. The results of the second experiment reveal that young Wuxi speakers identified coda nasals in Mandarin more accurately than older Wuxi speakers did, suggesting Wuxi speakers' production of Mandarin coda nasals is associated with their perception.

## The Diphthongs of Formal Nigerian English: A Preliminary Acoustic Analysis

*Natalia Dyrenko[1] and Robert Fuchs[2]*

[1]Englisches Seminar, Westfälische Wilhelms-Universität Münster, Germany

[2]Institut für Anglistik und Amerikanistik, Universität Hamburg, Germany

**Wed-P-2-3-7, Time: 14:30-16:30**

Postcolonial varieties of English, used in countries such as Nigeria, India and Singapore, are subject to both local ("endonormative") and external ("exonormative") forces, the latter often in the form of British/American English. This gives rise to a stylistic continuum, where informal speech is more endonormatively oriented than formal/educated speech, which, nevertheless, is clearly distinguishable from British/American English. The formal end of the continuum is often regarded as the incipient local standard. Nigerian English is the most widely spoken African variety of English, but empirical/quantitative descriptions are rare. In this pilot study, we present an acoustic analysis of eight phonological diphthongs produced in formal contexts by nine educated speakers of NigE with L1 Yoruba and drawn from the ICE Nigeria corpus.

Results show that the NigE speakers produced more monophthongal realisations of English phonological diphthongs than speakers of British English do, as measured by trajectory length in F1-F2 space. Phonetically, most of these vowels can be considered monophthongs.

The results can be explained through two factors at work during the foundation phase of NigE: (1) historical L1 influence and (2) the native English input present in the country, which involved more monophthongal realisations of some phonological diphthongs than in present-day BrE.

## Characterizing Rhythm Differences between Strong and Weak Accented L2 Speech

*Chris Davis and Jeesun Kim*

The MARCS Institute, Western Sydney University, Australia

**Wed-P-2-3-8, Time: 14:30-16:30**

This study examined the rhythmic characteristics of accented L2 speech by using two relatively novel measures of prosodic rhythm: The S-AMPH measure, an index of the degree of synchrony between the stress and syllable amplitude modulation rates; and the Allan Factor measure, that determines the nested clustering of temporal events (in this case peaks in the amplitude envelope) over different timescales. An extreme-group design was used to select strong versus weak foreign accent recordings from a group of Korean and French L2 English talkers saying the same 69-word English passage. For the Korean talkers, both the S-AMPH and the Allan Factor measures differed as a function of the strength of foreign accent. This was not the case for the French talkers, where neither measure differed as a function of the strength of the foreign accent. The difference in outcome between the Korean and French talkers suggests that the measures may not be indexing a general property of L2 accent (e.g., production fluency)

**Notes**

but rather that they may be picking up a property specific to the strongly accented Korean talkers. We consider several options.

## Analysis of Phone Errors Attributable to Phonological Effects Associated With Language Acquisition Through Bottleneck Feature Visualisations

*Eva Fringi and Martin Russell*

Department of Electronic Electrical and Systems Engineering, University of Birmingham B15 2TT, UK

**Wed-P-2-3-9, Time: 14:30-16:30**

Previous work aimed to investigate the extent to which errors attributable to phonological effects associated with language acquisition (PEALA) contribute to the output of children's ASR. Opposite to what was intuitively expected, the proportion of errors predictable from PEALA was positively correlated with recognition accuracy, therefore increased across ages. In order to interpret this finding, the present paper employs a DNN-HMM automatic speech recognition system, built on the CSLU children's speech corpus, to produce bottleneck feature (BNF) visualisations of phones and examine how these relate with respect to PEALA. The focus is drawn particularly on ASR errors caused by phone confusions, which are compared against phone substitution pairs indicated by PEALA. The ASR results confirm the previously observed interaction between errors predictable from PEALA and rising accuracy, but also suggest that these errors only account for a small percentage of the total phone substitution error. The BNF visualisations for the most part outline the age progression smoothly and demonstrate clear clusters of neighbouring phones consistently. The distance between PEALA related phones can be partitioned in four sets; two that increase with age (at a higher or lower rate), one that roughly remains constant and one that decreases with age.

## Category Similarity in Multilingual Pronunciation Training

*Jacques Koreman*

Norwegian University of Science and Technology, Trondheim, Norway

**Wed-P-2-3-10, Time: 14:30-16:30**

Learners with different native languages (L1) meet different challenges when they learn a foreign language (L2). The Speech Learning Model and the Perceptual Assimilation Model PAM-L2 have led to important insights about these challenges. Among other things, they have shown that the learnability of L2 sounds depends on their similarity to sounds in the L1: L2 sounds are more likely to lead to the formation of new phonetic categories if they differ strongly from L1 categories than if they are similar. The similarity of sounds is hard to quantify objectively, especially if the aim is to do this for many L1-L2 pairs. This limits the models' practical applicability.

The multilingual pronunciation training platform CALST offers exercises for all new L2 sounds. Two implementations of category (dis)similarity are proposed to identify new sounds, one at the level of functional similarity maintaining all L2 phonemic contrasts, the other based on a more fine-grained, multilingual similarity measure, where L2 sounds are considered new if they can contrast phonemically with the most similar L1 sound in any one language. This level of granularity reflects phonetically salient differences between sounds which, when perceived and produced adequately, suffice for high intelligibility and comprehensibility in L2.

## Talker Diarization in the Wild: the Case of Child-centered Daylong Audio-recordings

*Alejandrina Cristia[1], Shobhana Ganesh[2], Marisa Casillas[3] and Sriram Ganapathy[2]*

[1]LSCP, Département d'études cognitives, ENS, EHESS, CNRS, PSL University, Paris, 75005, France.

[2]Learning and Extraction of Acoustic Patterns (LEAP) lab, Electrical Engineering, Indian Institute of Science, Bangalore, 560012, India.

[3]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

**Wed-P-2-3-11, Time: 14:30-16:30**

Speaker diarization (answering "who spoke when") is a widely researched subject within speech technology. Numerous experiments have been run on datasets built from broadcast news, meeting data and call centers - the task sometimes appears close to being solved. Much less work has begun to tackle the hardest diarization task of all: spontaneous conversations in real-world settings. Such diarization would be particularly useful for studies of language acquisition, where researchers investigate the speech children produce and hear in their daily lives. In this paper, we study audio gathered with a recorder worn by small children as they went about their normal days. As a result, each child was exposed to different acoustic environments with a multitude of background noises and a varying number of adults and peers. The inconsistency of speech and noise within and across samples poses a challenging task for speaker diarization systems, which we tackled via retraining and data augmentation techniques. We further studied sources of structured variation across raw audio files, including the impact of speaker type distribution, proportion of speech from children and child age on diarization performance. We discuss the extent to which these findings might generalize to other samples of speech in the wild.

## Automated Classification of Children's Linguistic versus Non-Linguistic Vocalisations

*Zixing Zhang[1], Alejandrina Cristia[2], Anne Warlaumont[3] and Björn Schuller[1,4]*

[1]GLAM – Group on Language, Audio & Music, Imperial College London, UK

[2]LSCP, Département d'études cognitives, ENS, EHESS, CNRS, PSL University, France

[3]Department of Communication, University of California, Los Angeles, USA

[4]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany

**Wed-P-2-3-12, Time: 14:30-16:30**

A key outstanding task for speech technology involves dealing with non-standard speakers, notably young children. Distinguishing children's linguistic from non-linguistic vocalisations is crucial for a number of applied and fundamental research goals and yet there are few systems available for such a classification. This paper investigates two large-scale frame-level acoustic feature sets (eGeMAPS and ComParE16) followed by a dynamic model (GRU-RNN) and two kinds of derived static feature sets on the segment level (functional-based and Bag of Audio Words) combined with a static model (SVM) and automatically learnt representations directly from original raw voice signals by using an end-to-end system, which are compared against a simple phonetically-inspired baseline. These are applied to a large database of children's vocalisations (total N = 6,298) drawn from daylong recordings gathered in Namibia, Bolivia and Vanuatu. All of the systems outperform the baseline, with the highest performance in the test set for GRU-RNN using ComParE16 features. We identify promising paths of further research, including the application of a finer-grained classification of children's vocalisations onto these data and the exploration of other feature systems.

## Pitch Characteristics of L2 English Speech by Chinese Speakers: A Large-scale Study

*Jiahong Yuan, Qiusi Dong, Fei Wu, Huan Luan, Xiaofei Yang, Hui Lin and Yang Liu*

Liulishuo Inc.

**Wed-P-2-3-13, Time: 14:30-16:30**

AI-powered English learning apps are used by hundreds of millions of people across the globe on a daily basis. This presents a great opportunity for the study of L2 speech. On one hand, the amount of data accessible for research is very large and rapidly growing; on the other hand, new theories and understanding of L2 speech can be continually tested and revised through real-life and real-time applications.

This paper presents a study of pitch characteristics of L2 English speech using a large-scale dataset from a language learning app. Our dataset

**Notes**

contains 180,000 spoken utterances which amount to 240 hours of speech. The results show that compared to L1, L2 English has narrower pitch range and slower rate of pitch change, but more small "ripples" on the pitch contour. The percentage of F0 rise time is higher in L2 and the maximum F0 in an utterance is realized later (with respect to the onset of the word on which the maximum F0 resides). These results suggest that the influence of L1 on L2 prosody is more complex than previously demonstrated and they shed light on L2 prosody assessment and learning.

# Wed-P-2-4: Topics in Speech Recognition

Hall 4-6: Poster-4; 14:30-16:30; Wednesday, 5 September, 2018
Chair: Dirk Van Compernolle

## Dual Language Models for Code Switched Speech Recognition

*Saurabh Garg, Tanmay Parekh and Preethi Jyothi*
Department of Computer Science and Engineering, Indian Institute of Technology Bombay

**Wed-P-2-4-1, Time: 14:30-16:30**

In this work, we present a simple and elegant approach to language modeling for bilingual code-switched text. Since code-switching is a blend of two or more different languages, a standard bilingual language model can be improved upon by using structures of the monolingual language models. We propose a novel technique called dual language models, which involves building two complementary monolingual language models and combining them using a probabilistic model for switching between the two. We evaluate the efficacy of our approach using a conversational Mandarin-English speech corpus. We prove the robustness of our model by showing significant improvements in perplexity measures over the standard bilingual language model without the use of any external information. Similar consistent improvements are also reflected in automatic speech recognition error rates.

## Multilingual Neural Network Acoustic Modelling for ASR of Under-Resourced English-isiZulu Code-Switched Speech

*Astik Biswas[1], Febe de Wet[1], Ewald van der Westhuizen[1], Emre Yılmaz[2,3] and Thomas Niesler[1]*
[1]Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa
[2]CLS/CLST, Radboud University, Nijmegen, Netherlands
[3]Dept. of Electrical and Computer Engineering, National University of Singapore, Singapore

**Wed-P-2-4-2, Time: 14:30-16:30**

Although isiZulu speakers code-switch with English as a matter of course, extremely little appropriate data is available for acoustic modelling. Recently, a small five-language corpus of code-switched South African soap opera speech was compiled. We used this corpus to evaluate the application of multilingual neural network acoustic modelling to English-isiZulu code-switched speech recognition. Our aim was to determine whether English-isiZulu speech recognition accuracy can be improved by incorporating three other language pairs in the corpus: English-isiXhosa, English-Setswana and English-Sesotho. Since isiXhosa, like isiZulu, belongs to the Nguni language family, while Setswana and Sesotho belong to the more distant Sotho family, we could also investigate the merits of additional data from within and across language groups. Our experiments using both fully connected DNN and TDNN-LSTM architectures show that English-isiZulu speech recognition accuracy as well as language identification after code-switching is improved more by the incorporation of English-isiXhosa data than by the incorporation of the other language pairs. However additional data from the more distant language group remained beneficial and the best overall performance was always achieved with a multilingual neural network trained on all four language pairs.

## Fast ASR-free and Almost Zero-resource Keyword Spotting Using DTW and CNNs for Humanitarian Monitoring

*Raghav Menon[1], Herman Kamper[1], John Quinn[2] and Thomas Niesler[1]*
[1]Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa
[2]UN Global Pulse, Kampala, Uganda

**Wed-P-2-4-3, Time: 14:30-16:30**

We use dynamic time warping (DTW) as supervision for training a convolutional neural network (CNN) based keyword spotting system using a small set of spoken isolated keywords. The aim is to allow rapid deployment of a keyword spotting system in a new language to support urgent United Nations (UN) relief programmes in parts of Africa where languages are extremely under-resourced and the development of annotated speech resources is infeasible.

First, we use 1920 recorded keywords (40 keyword types, 34 minutes of speech) as exemplars in a DTW-based template matching system and apply it to untranscribed broadcast speech.

Then, we use the resulting DTW scores as targets to train a CNN on the same unlabelled speech. In this way we use just 34 minutes of labelled speech, but leverage a large amount of unlabelled data for training. While the resulting CNN keyword spotter cannot match the performance of the DTW-based system, it substantially outperforms a CNN classifier trained only on the keywords, improving the area under the ROC curve from 0.54 to 0.64. Because our CNN system is several orders of magnitude faster at runtime than the DTW system, it represents the most viable keyword spotter on this extremely limited dataset.

## Text-Dependent Speech Enhancement for Small-Footprint Robust Keyword Detection

*Meng Yu[1], Xuan Ji[1], Yi Gao[1], Lianwu Chen[1], Jie Chen[1], Jimeng Zheng[1], Dan Su[1] and Dong Yu[2]*
[1]Tencent AI Lab, Shenzhen, China
[2]Tencent AI Lab, Bellevue, WA, USA

**Wed-P-2-4-4, Time: 14:30-16:30**

Keyword detection (KWD), also known as keyword spotting, is in great demand in small devices in the era of Internet of Things. Albeit recent progresses, the performance of KWD, measured in terms of precision and recall rate, may still degrade significantly when either the non-speech ambient noises or the human voice and speech-like interferences (e.g., TV, background competing talkers) exists. In this paper, we propose a general solution to address all kinds of environmental interferences. A novel text-dependent speech enhancement (TDSE) technique using a recurrent neural network (RNN) with long short-term memory (LSTM) is presented for improving the robustness of the small-footprint KWD task in the presence of environmental noises and interfering talkers. On our large simulated and recorded noisy and far-field evaluation sets, we show that TDSE significantly improves the quality of the target keyword speech and performs particularly well under speech interference conditions. We demonstrate that KWD with TDSE frontend significantly outperforms the baseline KWD system with or without a generic speech enhancement in terms of equal error rate (EER) in the keyword detection evaluation.

## Improved ASR for Under-resourced Languages through Multi-task Learning with Acoustic Landmarks

*Di He[1], Boon Pang Lim[2], Xuesong Yang[3], Mark Hasegawa-Johnson[3] and Deming Chen[1]*
[1]Coordinated Science Lab, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA 61801
[2]Novumind Inc, Santa Clara, USA 95054
[3]Beckman Institute, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA 61801

**Wed-P-2-4-5, Time: 14:30-16:30**

**Notes**

Furui first demonstrated that the identity of both consonant and vowel can be perceived from the C-V transition; later, Stevens proposed that acoustic landmarks are the primary cues for speech perception and that steady-state regions are secondary or supplemental. Acoustic landmarks are perceptually salient, even in a language one doesn't speak and it has been demonstrated that non-speakers of the language can identify features such as the primary articulator of the landmark. These factors suggest a strategy for developing language-independent automatic speech recognition: landmarks can potentially be learned once from a suitably labeled corpus and rapidly applied to many other languages. This paper proposes enhancing the cross-lingual portability of a neural network by using landmarks as the secondary task in multi-task learning (MTL). The network is trained in a well-resourced source language with both phone and landmark labels (English), then adapted to an under-resourced target language with only word labels (Iban). Landmark-tasked MTL reduces source-language phone error rate by 2.9% relative and reduces target-language word error rate by 1.9%-5.9% depending on the amount of target-language training data. These results suggest that landmark-tasked MTL causes the DNN to learn hidden-node features that are useful for cross-lingual adaptation.

## Cross-language Phoneme Mapping for Low-resource Languages: An Exploration of Benefits and Trade-offs

*Nick K Chibuye[1], Todd Rosenstock[2] and Brian DeRenzi[1]*

[1]University of Cape Town

[2]World Agroforestry Centre (ICRAF)

**Wed-P-2-4-6, Time: 14:30-16:30**

Voice-based systems are an essential approach for engaging directly with low-literate and underrepresented populations. Previous work has taken advantage of high-resource speech recognition technology for low-resource language speech recognition through cross-language phoneme mapping. Unfortunately, there is little guidance in how to deploy these systems across a range of languages. We present a systematic exploration of four source languages and five target languages to understand the trade-offs and performance of different source languages and training techniques. We find that one can improve recognition accuracy by selecting a source language that has similar linguistic properties to that of the target language. We also find that the number of alternative pronunciations per word and gender of participants also impact recognition accuracy. Our work will allow other researchers and practitioners to quickly develop high-quality small-vocabulary speech-based applications for under-resourced languages.

## User-centric Evaluation of Automatic Punctuation in ASR Closed Captioning

*Máté Ákos Tündik[1], György Szaszák[1], Gábor Gosztolya[2] and András Beke[3]*

[1]Dept. of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Budapest, Hungary

[2]Research Group on Artificial Intelligence, Hungarian Academy of Sciences, Szeged, Hungary

[3]Spicy Analytics Ltd., Budapest, Hungary

**Wed-P-2-4-7, Time: 14:30-16:30**

Punctuation of ASR-produced transcripts has received increasing attention in the recent years; RNN-based sequence modelling solutions which exploit textual and/or acoustic features show encouraging performance. Switching the focus from the technical side, qualifying and quantifying the benefits of such punctuation from end-user perspective have not been performed yet exhaustively. The ambition of the current paper is to explore to what extent automatic punctuation can improve human readability and understandability.

The paper presents a user-centric evaluation of a real-time closed captioning system enhanced by a lightweight RNN-based punctuation module. Subjective tests involve both normal hearing and deaf or hard-of-hearing (DHH) subjects. Results confirm that automatic punctuation itself significantly increases understandability, even if several other factors interplay in subjective impression. The perceived improvement is even more pronounced in the DHH group. A statistical analysis is carried out to identify objectively measurable factors which are well reflected by subjective scores.

## Punctuation Prediction Model for Conversational Speech

*Piotr Żelasko[1,2], Piotr Szymański[1,3], Jan Mizgajski[1], Adrian Szymczak[1], Yishay Carmiel[1] and Najim Dehak[4]*

[1]Intelligent Wire, USA

[2]Department of Computer Science, Electronics and Telecommunications, AGH University of Science and Technology, al. Mickiewicza 30, Kraków, Poland

[3]Department of Computational Intelligence, Wrocław University of Technology, Wybrzeże Stanisława Wyspiańskiego 27, 50-370 Wrocław, Poland

[4]Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, MD, USA

**Wed-P-2-4-8, Time: 14:30-16:30**

An ASR system usually does not predict any punctuation or capitalization. Lack of punctuation causes problems in result presentation and confuses both the human reader and off-the-shelf natural language processing algorithms. To overcome these limitations, we train two variants of Deep Neural Network (DNN) sequence labelling models - a Bidirectional Long Short-Term Memory (BLSTM) and a Convolutional Neural Network (CNN), to predict the punctuation. The models are trained on the Fisher corpus which includes punctuation annotation. In our experiments, we combine time-aligned and punctuated Fisher corpus transcripts using a sequence alignment algorithm. The neural networks are trained on Common Web Crawl GloVe embedding of the words in Fisher transcripts aligned with conversation side indicators and word time infomation. The CNNs yield a better precision and BLSTMs tend to have better recall. While BLSTMs make fewer mistakes overall, the punctuation predicted by the CNN is more accurate - especially in the case of question marks. Our results constitute significant evidence that the distribution of words in time, as well as pre-trained embeddings, can be useful in the punctuation prediction task.

## BUT OpenSAT 2017 Speech Recognition System

*Martin Karafiát, Murali Karthick Baskar, Igor Szöke, Vladimír Malenovský, Karel Veselý, František Grézl, Lukáš Burget and Jan Černocký*

Brno University of Technology Speech@FIT and IT4I Center of Excellence, Brno, Czechia

**Wed-P-2-4-9, Time: 14:30-16:30**

The paper describes BUT Automatic Speech Recognition (ASR) systems for two domains in OpenSAT evaluations: Low Resourced Languages and Public Safety Communications. The first was challenging due to lack of training data, therefore multilingual approaches for BLSTM training were employed and recently published Residual Memory Networks requiring less training data were used. Combination of both approaches led to superior performance. The second domain was challenging due to recording in extreme conditions: specific channel, speaker under stress, high levels of noise. A data augmentation process was very important to get reasonably good performance.

## Visual Recognition of Continuous Cued Speech Using a Tandem CNN-HMM Approach

*Li Liu, Thomas Hueber, Gang Feng and Denis Beautemps*

Univ. Grenoble Alpes, CNRS, Grenoble INP*, GIPSA-lab, 38000 Grenoble, France

* Institute of Engineering Univ. Grenoble Alpes

**Wed-P-2-4-10, Time: 14:30-16:30**

This study addresses the problem of automatic recognition of Cued Speech (CS), a visual mode of communication for hearing

**Notes**

impaired people in which a complete phonetic repertoire is obtained by combining lip movements with hand cues. In the proposed system, the dynamic of visual features extracted from lip and hand images using convolutional neural networks (CNN) are modeled by a set of hidden Markov models (HMM), for each phonetic context (tandem architecture). CNN-based feature extraction is compared to an unsupervised approach based on the principal component analysis. A novel temporal segmentation of hand streams is used to train CNNs efficiently. Different strategies for combining the extracted visual features within the HMM decoder are investigated. Experimental evaluation is carried on an audiovisual dataset (containing only continuous French sentences) recorded specifically for this study. In its best configuration and without exploiting any dictionary or language model, the proposed tandem CNN-HMM architecture is able to identify correctly more than 73% of the phoneme (62% when considering insertion errors).

### Building Large-vocabulary Speaker-independent Lipreading Systems

*Kwanchiva Thangthai[1,2] and Richard Harvey[1]*

[1]School of Computing Sciences, University of East Anglia, Norwich, UK

[2]National Electronics and Computer Technology Center, Thailand

**Wed-P-2-4-11, Time: 14:30-16:30**

Constructing a viable lipreading system is a challenge because it is claimed that only 30% of information of speech production is visible on the lips. Nevertheless, in small vocabulary tasks, there have been several reports of high accuracies. However, investigation of larger vocabulary tasks is much rarer.

This work examines constructing a large vocabulary lipreading system using an approach based-on Deep Neural Network Hidden Markov Models (DNN-HMMs). We tackle the problem of lipreading an unseen speaker. We investigate the effect of employing several steps to pre-process visual features. Moreover, we examine the contribution of language modelling in a lipreading system where we use longer n-grams to recognise visual speech. Our lipreading system is constructed on the 6000-word vocabulary TCD-TIMIT audiovisual speech corpus. The results show that visual speech recognition can definitely reach 50% word accuracy on large vocabularies. We actually achieved a mean of 53.83% measured via three-fold cross-validation on the speaker independent setting of the TCD-TIMIT corpus using bigrams.

### CRIM's System for the MGB-3 English Multi-Genre Broadcast Media Transcription

*Vishwa Gupta and Gilles Boulianne*

Centre de recherche informatique de Montréal (CRIM)

**Wed-P-2-4-12, Time: 14:30-16:30**

The second English Multi-Genre Broadcast Challenge (MGB-3) is a controlled evaluation of speech recognition and lightly supervised alignment using BBC TV recordings. CRIM is participating in the speech recognition part of the challenge. This paper presents CRIM's contributions to the MGB-3 transcription task. This task is inherently more difficult than the first task as the training audio has been reduced from 1200 hours to 500 hours. CRIM's main contributions are experimentation with bidirectional LSTM models and lattice-free MMI (LF-MMI) trained TDNN models for acoustic modeling, LSTM and DNN models for speech/non-speech detection for input to speaker diarization and LSTM language models for rescoring lattices. We also show that adding senone posteriors to the input of LSTM and DNN models for speech/non-speech detection (VAD) reduces error rate. CRIM's best single decoding WER for the MGB-3 dev17 dev set (with reference segmentation) went down from 27.6% (with our MGB-1 challenge system) to 24.1% for this task using the LF-MMI trained TDNN models. The final WER on dev17 set (after VAD) is 20.9% and on the new dev18 development set is 20.8%.

### Sampling Strategies in Siamese Networks for Unsupervised Speech Representation Learning

*Rachid Riad[1,2], Corentin Dancette[1], Julien Karadayi[1], Neil Zeghidour[1,3], Thomas Schatz[1,4,5] and Emmanuel Dupoux[1,3]*

[1]CoML/ENS/CNRS/EHESS/INRIA/PSL Research University, Paris, France

[2]NPI/ENS/INSERM/UPEC, Créteil, France,

[3]Facebook A.I. Research, Paris, France

[4]Department of Linguistics & UMIACS, University of Maryland, College Park MD, USA

[5]Department of Linguistics, Massachusetts Institute of Technology, Cambridge MA, USA

**Wed-P-2-4-13, Time: 14:30-16:30**

Recent studies have investigated siamese network architectures for learning invariant speech representations using same-different side information at the word level. Here we investigate systematically an often ignored component of siamese networks: the sampling procedure (how pairs of same vs. different tokens are selected). We show that sampling strategies taking into account Zipf's Law, the distribution of speakers and the proportions of same and different pairs of words significantly impact the performance of the network. In particular, we show that word frequency compression improves learning across a large range of variations in the number of training pairs. This effect does not apply to the same extent to the fully unsupervised setting, where the pairs of same-different words are obtained by spoken term discovery. We apply these results to pairs of words discovered using an unsupervised algorithm and show an improvement on the state-of-the-art in unsupervised representation learning using siamese networks.

### Compact Feedforward Sequential Memory Networks for Small-footprint Keyword Spotting

*Mengzhe Chen, ShiLiang Zhang, Ming Lei, Yong Liu, Haitao Yao and Jie Gao*

Alibaba Group, P. R. China

**Wed-P-2-4-14, Time: 14:30-16:30**

Due to limited resource on devices and complicated scenarios, a compact model with high precision, low computational cost and latency is expected for small-footprint keyword spotting tasks. To fulfill these requirements, in this paper, compact Feed-forward Sequential Memory Network (cFSMN) which combines low-rank matrix factorization with conventional FSMN is investigated for a far-field keyword spotting task. The effect of its architecture parameters is analyzed. Towards achieving lower computational cost, multiframe prediction (MFP) is applied to cFSMN. For enhancing the modeling capacity, an advanced MFP is attempted by inserting small DNN layers before output layers. The performance is measured by area under the curve (AUC) for detection error tradeoff (DET) curves. The experiments show that compared with a well-tuned long short-term memory (LSTM) which needs the same latency and twofold computational cost, the cFSMN achieves 18.11% and 29.21% AUC relative decreases on the test sets which are recorded in quiet and noisy environment respectively. After applying advanced MFP, the system gets 0.48% and 20.04% AUC relative decrease over conventional cFSMN on the quiet and noisy test sets respectively, while the computational cost relatively reduces 46.58%.

## Wed-O-3-1: Zero-resource Speech Recognition

Hall 1; 17:00-19:00; Wednesday, 5 September, 2018

Chairs: Florian Metze and Hynek Hermansky

### Multilingual Bottleneck Features for Subword Modeling in Zero-resource Languages

*Enno Hermann and Sharon Goldwater*

ILCC, School of Informatics, University of Edinburgh, UK

**Wed-O-3-1-1, Time: 17:00-17:20**

How can we effectively develop speech technology for languages where no transcribed data is available? Many existing approaches use no annotated resources at all, yet it makes sense to leverage information from

**Notes**

large annotated corpora in other languages, for example in the form of multilingual bottleneck features (BNFs) obtained from a supervised speech recognition system. In this work, we evaluate the benefits of BNFs for subword modeling (feature extraction) in six unseen languages on a word discrimination task. First we establish a strong unsupervised baseline by combining two existing methods: vocal tract length normalisation (VTLN) and the correspondence autoencoder (cAE). We then show that BNFs trained on a single language already beat this baseline; including up to 10 languages results in additional improvements which cannot be matched by just adding more data from a single language. Finally, we show that the cAE can improve further on the BNFs if high-quality same-word pairs are available.

## Exploiting Speaker and Phonetic Diversity of Mismatched Language Resources for Unsupervised Subword Modeling

*Siyuan Feng and Tan Lee*

Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong

**Wed-O-3-1-2, Time: 17:20-17:40**

This study addresses the problem of learning robust frame-level feature representation for unsupervised subword modeling in the zero-resource scenario. Robustness of the learned features is achieved through effective speaker adaptation and exploiting cross-lingual phonetic knowledge. For speaker adaptation, an out-of-domain automatic speech recognition (ASR) system is used to estimate fMLLR features for untranscribed speech of target zero-resource languages. The fMLLR features are applied in multi-task learning of a deep neural network (DNN) to further obtain phonetically discriminative and speaker-invariant bottleneck features (BNFs). Frame-level labels for DNN training can be acquired based on two approaches: Dirichlet process Gaussian mixture model (DPGMM) clustering and out-of-domain ASR decoding. Moreover, system fusion is performed by concatenating BNFs extracted by different DNNs. Our methods are evaluated by ZeroSpeech 2017 Track one, where the performance is evaluated by ABX minimal pair discriminability. Experimental results demonstrate that: (1) Using an out-of-domain ASR system to perform speaker adaptation of zero-resource speech is effective and efficient; (2) Our system achieves highly competitive performance to state of the art; (3) System fusion could improve feature representation capability.

## Unsupervised Word Segmentation from Speech with Attention

*Pierre Godard[1], Marcely Zanon Boito[2], Lucas Ondel[3], Alexandre Berard[2,5], François Yvon[1], Aline Villavicencio[4] and Laurent Besacier[2]*

[1]LIMSI, CNRS, Université Paris-Saclay, Orsay, France

[2]LIG, UGA, G-INP, CNRS, INRIA, Grenoble, France

[3]BUT, Brno, Czech Republic

[4]CSEE, University of Essex, UK

[5]CRISTAL, Université de Lille, France

**Wed-O-3-1-3, Time: 17:40-18:00**

We present a first attempt to perform attentional word segmentation from speech signal, with the final goal of automatically identifying lexical units in a low-resource, unwritten language (UL). Our methodology assumes a pairing between recordings in the UL with translations in a well-resourced language. It uses Acoustic Unit Discovery (AUD) to convert speech into a pseudo-phones sequence that is segmented using neural soft alignments (from a neural machine translation model). Evaluation uses an actual Bantu UL, Mboshi; comparisons to monolingual and bilingual baselines illustrate the potential of attentional word segmentation for language documentation.

## Learning Word Embeddings: Unsupervised Methods for Fixed-size Representations of Variable-length Speech Segments

*Nils Holzenberger[1], Mingxing Du[2], Julien Karadayi[2], Rachid Riad[2,3] and Emmanuel Dupoux[2]*

[1]CLSP, Johns Hopkins University, USA

[2]ENS, EHESS, PSL Research University, CNRS, INRIA, France

[3]INSERM, NPI, UPEC, France

**Wed-O-3-1-4, Time: 18:00-18:20**

Fixed-length embeddings of words are very useful for a variety of tasks in speech and language processing. Here we systematically explore two methods of computing fixed-length embeddings for variable-length sequences. We evaluate their susceptibility to phonetic and speaker-specific variability on English, a high resource language and Xitsonga, a low resource language, using two evaluation metrics: ABX word discrimination and ROC-AUC on same-different phoneme n-grams. We show that a simple downsampling method supplemented with length information can outperform the variable-length input feature representation on both evaluations. Recurrent autoencoders, trained without supervision, can yield even better results at the expense of increased computational complexity.

## Full Bayesian Hidden Markov Model Variational Autoencoder for Acoustic Unit Discovery

*Thomas Glarner, Patrick Hanebrink, Janek Ebbers and Reinhold Haeb-Umbach*

Paderborn University, Germany

**Wed-O-3-1-5, Time: 18:20-18:40**

The invention of the Variational Autoencoder enables the application of Neural Networks to a wide range of tasks in unsupervised learning, including the field of Acoustic Unit Discovery (AUD). The recently proposed Hidden Markov Model Variational Autoencoder (HMMVAE) allows a joint training of a neural network based feature extractor and a structured prior for the latent space given by a Hidden Markov Model.

It has been shown that the HMMVAE significantly outperforms pure GMM-HMM based systems on the AUD task.

However, the HMMVAE cannot autonomously infer the number of acoustic units and thus relies on the GMM-HMM system for initialization.

This paper introduces the Bayesian Hidden Markov Model Variational Autoencoder (BHMMVAE) which solves these issues by embedding the HMMVAE in a Bayesian framework with a Dirichlet Process Prior for the distribution of the acoustic units and diagonal or full-covariance Gaussians as emission distributions.

Experiments on Timit and Xitsonga show that the BHMMVAE is able to autonomously infer a reasonable number of acoustic units, can be initialized without supervision by a GMM-HMM system, achieves computationally efficient stochastic variational inference by using natural gradient descent and, additionally, improves the AUD performance over the HMMVAE.

## Unspeech: Unsupervised Speech Context Embeddings

*Benjamin Milde and Chris Biemann*

Language Technology, Universität Hamburg

**Wed-O-3-1-6, Time: 18:40-19:00**

We introduce "Unspeech" embeddings, which are based on unsupervised learning of context feature representations for spoken language. The embeddings were trained on up to 9500 hours of crawled English speech data without transcriptions or speaker information, by using a straightforward learning objective based on context and non-context discrimination with negative sampling. We use a Siamese convolutional neural network architecture to train Unspeech embeddings and evaluate them on speaker comparison, utterance clustering and as a context feature in TDNN-HMM acoustic models trained on TED-LIUM, comparing it to i-vector baselines. Particularly decoding out-of-domain speech data from the recently released Common Voice corpus shows consistent WER reductions. We release our source code and pre-trained Unspeech models under a permissive open source license.

**Notes**

# Wed-O-3-2: Spatial and Phase Cues for Source Separation and Speech Recognition

Hall 2; 17:00-19:00; Wednesday, 5 September, 2018

Chairs: Jon Barker and Emmanuel Vincent

## Impact of Aliasing on Deep CNN-Based End-to-End Acoustic Models

*Yuan Gong and Christian Poellabauer*

Department of Computer Science and Engineering, University of Notre Dame, IN 46556, USA

**Wed-O-3-2-1, Time: 17:00-17:20**

A recent trend in audio and speech processing is to learn target labels directly from raw waveforms rather than hand-crafted acoustic features. Previous work has shown that deep convolutional neural networks (CNNs) as front-end can learn effective representations from the raw waveform. However, due to the large dimension of raw audio waveforms, pooling layers are usually used aggressively between temporal convolutional layers. In essence, these pooling layers perform operations that are similar to signal downsampling, which may lead to temporal aliasing according to the Nyquist-Shannon sampling theorem. This paper explores, using a series of experiments, if and how this aliasing effect impacts modern deep CNN-based models.

## Keyword Based Speaker Localization: Localizing a Target Speaker in a Multi-speaker Environment

*Sunit Sivasankaran, Emmanuel Vincent and Dominique Fohr*

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

**Wed-O-3-2-2, Time: 17:20-17:40**

Speaker localization is a hard task, especially in adverse environmental conditions involving reverberation and noise. In this work we introduce the new task of localizing the speaker who uttered a given keyword, e.g., the wake-up word of a distant-microphone voice command system, in the presence of overlapping speech. We employ a convolutional neural network based localization system and investigate multiple identifiers as additional inputs to the system in order to characterize this speaker. We conduct experiments using ground truth identifiers which are obtained assuming the availability of clean speech and also in realistic conditions where the identifiers are computed from the corrupted speech. We find that the identifier consisting of the ground truth time-frequency mask corresponding to the target speaker provides the best localization performance and we propose methods to estimate such a mask in adverse reverberant and noisy conditions using the considered keyword.

## End-to-End Speech Separation with Unfolded Iterative Phase Reconstruction

*Zhong-Qiu Wang[1,2], Jonathan Le Roux[1], DeLiang Wang[2,3] and John Hershey[1]*

[1]Mitsubishi Electric Research Laboratories (MERL), USA

[2]Department of Computer Science and Engineering, The Ohio State University, USA

[3]Center for Cognitive and Brain Sciences, The Ohio State University, USA

**Wed-O-3-2-3, Time: 17:40-18:00**

This paper proposes an end-to-end approach for single-channel speaker-independent multi-speaker speech separation, where time-frequency (T-F) masking, the short-time Fourier transform (STFT) and its inverse are represented as layers within a deep network. Previous approaches, rather than computing a loss on the reconstructed signal, used a surrogate loss based on the target STFT magnitudes. This ignores reconstruction error introduced by phase inconsistency. In our approach, the loss function is directly defined on the reconstructed signals, which are optimized for best separation. In addition, we train through unfolded iterations of a phase reconstruction algorithm, represented as a series of STFT and inverse STFT layers. While mask values are typically limited to lie between zero and one for approaches using the mixture phase for reconstruction, this limitation is less relevant if the estimated magnitudes are to be used together with phase reconstruction. We thus propose several novel activation functions for the output layer of the T-F masking, to allow mask values beyond one. On the publicly-available wsj0-2mix dataset, our approach achieves state-of-the-art 12.6 dB scale-invariant signal-to-distortion ratio (SI-SDR) and 13.1 dB SDR, revealing new possibilities for deep learning based phase reconstruction and representing a fundamental progress towards solving the notoriously-hard cocktail party problem.

## PhaseNet: Discretized Phase Modeling with Deep Neural Networks for Audio Source Separation

*Naoya Takahashi[1], Purvi Agrawal[2], Nabarun Goswami[3] and Yuki Mitsufuji[1]*

[1]Sony Corporation

[2]Indian Institute of Science

[3]Sony India Software Centre

**Wed-O-3-2-4, Time: 18:00-18:20**

Previous research on audio source separation based on deep neural networks (DNNs) mainly focuses on estimating the magnitude spectrum of target sources and typically, phase of the mixture signal is combined with the estimated magnitude spectra in an ad-hoc way. Although recovering target phase is assumed to be important for the improvement of separation quality, it can be difficult to handle the periodic nature of the phase with the regression approach. Unwrapping phase is one way to eliminate the phase discontinuity, however, it increases the range of value along with the times of unwrapping, making it difficult for DNNs to model.

To overcome this difficulty, we propose to treat the phase estimation problem as a classification problem by discretizing phase values and assigning class indices to them.

Experimental results show that our classification-based approach 1) successfully recovers the phase of the target source in the discretized domain, 2) improves signal-to-distortion ratio (SDR) over the regression-based approach in both speech enhancement task and music source separation (MSS) task and 3) outperforms state-of-the-art MSS.

## Integrating Spectral and Spatial Features for Multi-Channel Speaker Separation

*Zhong-Qiu Wang[1] and DeLiang Wang[1, 2]*

[1]Department of Computer Science and Engineering, The Ohio State University, USA

[2]Center for Cognitive and Brain Sciences, The Ohio State University, USA

**Wed-O-3-2-5, Time: 18:20-18:40**

This paper tightly integrates spectral and spatial information for deep learning based multi-channel speaker separation. The key idea is to localize individual speakers so that an enhancement network can be used to separate the speaker from an estimated direction and with specific spectral characteristics. To determine the direction of the speaker of interest, we identify time-frequency (T-F) units dominated by that speaker and only use them for direction of arrival (DOA) estimation. The speaker dominance at each T-F unit is determined by a two-channel permutation invariant training network, which combines spectral and interchannel phase patterns at the input feature level. In addition, beamforming is tightly integrated in the proposed system by exploiting the magnitudes and phase pro-duced by T-F masking based beamforming. Strong separation performance has been observed on a spatialized reverberant version of the wsj0-2mix corpus.

## DNN Driven Speaker Independent Audio-Visual Mask Estimation for Speech Separation

*Mandar Gogate[1], Ahsan Adeel[1], Ricard Marxer[2,3], Jon Barker[3] and*

**Notes**

*Amir Hussain*[1]
[1]University of Stirling, UK,
[2]Université de Toulon, Aix Marseille Univ, CNRS, LIS, Marseille, France
[3]University of Sheffield, UK

**Wed-O-3-2-6, Time: 18:40-19:00**

Human auditory cortex excels at selectively suppressing background noise to focus on a target speaker. The process of selective attention in the brain is known to contextually exploit the available audio and visual cues to better focus on target speaker while filtering out other noises. In this study, we propose a novel deep neural network (DNN) based audiovisual (AV) mask estimation model. The proposed AV mask estimation model contextually integrates the temporal dynamics of both audio and noise-immune visual features for improved mask estimation and speech separation. For optimal AV features extraction and ideal binary mask (IBM) estimation, a hybrid DNN architecture is exploited to leverages the complementary strengths of a stacked long short term memory (LSTM) and convolution LSTM network. The comparative simulation results in terms of speech quality and intelligibility demonstrate significant performance improvement of our proposed AV mask estimation model as compared to audio-only and visual-only mask estimation approaches for both speaker dependent and independent scenarios.

# Wed-O-3-3: Dialectal Variation

MR G.01-G.02; 17:00-19:00; Wednesday, 5 September, 2018
Chairs: Elinor Payne and Katherine Earnshaw

## Exploring Temporal Reduction in Dialectal Spanish: A Large-scale Study of Lenition of Voiced Stops and Coda-s

*Ioana Vasilescu*[1], *Nidia Hernandez*[2], *Bianca Vieru*[2] *and Lori Lamel*[1]
[1]LIMSI CNRS, Paris-Saclay University
[2]Vocapia Research

**Wed-O-3-3-1, Time: 17:00-17:20**

Large scale studies of temporal reduction are of interest to obtain linguistic knowledge and this understanding can potentially help improve speech technologies. Automatic speech recognition (ASR) supports the processing of very large corpora and can be used to enrich linguistic studies with models of speech production and perception grounded on observed commonly used pronunciations. In return, ASR can benefit from the linguistic findings by including pronunciation variants that reflect the linguistic variability. This study focuses on two temporal reduction phenomena in Peninsular and Latin American varieties of Spanish: lenition of intervocalic voiced stops V/bdg/V and coda-s. First, the two phenomena are investigated via a study of transcription errors produced by a speech recognition system designed for Peninsular Spanish which can be potentially attributed to lenition. In a second step, automatic forced alignment experiments are conducted using specific pronunciation variants with and without lenition to measure the extent of the phenomenon as a function of geographical and stylistic repartition. The results show that the distribution of pronunciation variants across Peninsular and Latin American Spanish varieties is consistent with trends depicted by classical linguistic studies. The speaking style appears to be the main factor affecting the +/-lenition variation. The findings also suggest that including such variants in ASR system's lexicon may improve performance when processing multiple Spanish varieties.

## Dialect-geographical Acoustic-Tonetics: Five Disyllabic Tone Sandhi Patterns in Cognate Words from the Wu Dialects of ZhèJiāNg Province

*Phil Rose*
Emeritus Faculty, Australian National University, Australia

**Wed-O-3-3-2, Time: 17:20-17:40**

The typology of tone sandhi patterns on disyllabic tonally cognate words in selected sub-groups of the Wu dialects of the east-central Chinese province of Zhejiang is investigated using data from 48 sites collected over 45 years. A new method of extrinsic z-score normalisation is demonstrated which permits comparison of tones across dialects with different pitch ranges. Five different but typical right-dominant word-tone patterns are identified, acoustically quantified and their geographical distribution specified. It is hypothesized that changes in isolation tones and different types of dissimilation of the first tone from the word-final tone, are a possible origin of the observed variation.

## Regional Variation of /r/ in Swiss German Dialects

*Adrian Leemann*[1], *Stephan Schmid*[2], *Dieter Studer-Joho*[2] *and Marie-José Kolly*[3]
[1]Lancaster University, United Kingdom
[2]University of Zurich, Switzerland
[3]Neue Zürcher Zeitung, Switzerland

**Wed-O-3-3-3, Time: 17:40-18:00**

German-speaking Europe is known to feature substantial regional variation in the articulation of /r/. According to historical atlases, this is particularly true for the most southwestern fringe of the region, i.e. German-speaking Switzerland. Large-scale, multilocality studies that show an updated picture of regional variation in this region are lacking, however. To this end, we coded /r/s of almost 3,000 speakers from 438 localities on a predominantly auditory basis, using data crowdsourced through a smartphone app. We report substantial regional variation, with uvular articulations especially dominant in the Northwest and the Northeast and alveolar - particularly tapped - articulations prevalent in the Midlands. We further provide exemplary evidence of an urban ([ʁ]) vs. rural stratification ([ɾ]) in the Northwest. This contribution further discusses (a) issues related to the coding of /r/, given the volatile articulatory and acoustic properties of /r/s and (b) the benefits and pitfalls of the crowdsourcing methodology applied more generally.

## Variation in the FACE Vowel across West Yorkshire: Implications for Forensic Speaker Comparisons

*Kate Earnshaw and Erica Gold*
Department of Linguistics and Modern Languages, University of Huddersfield, United Kingdom

**Wed-O-3-3-4, Time: 18:00-18:20**

In the field of sociophonetics, research is largely focused on the documentation of regional variability. However, the majority of literature in the United Kingdom often reports on variation at a macro-level (e.g. Northern, Yorkshire, West Yorkshire) rather than at a more local level (e.g. West Yorkshire: Bradford, Calderdale, Kirklees, Leeds, Wakefield). Traditionally, for sociophoneticians, examining regional variation at a broader level is adequate for answering research questions related to language change or more general variation. For practical applications (e.g. forensic, speech technology), however, more fine-grained regional analysis is necessary. This paper analyses over 2000 FACE tokens from three metropolitan boroughs (Bradford, Kirklees and Wakefield) within West Yorkshire, in order to determine the extent to which F1~F3 vary across the region. Results suggest that for FACE, these three boroughs within West Yorkshire are more regionally stratified than previously acknowledged. These findings are of particular importance to the forensic speech science community, as experts rely on these regional nuances in order to make important judgments related to strength of the speech evidence in a case. Should decisions be made without the greater understanding of local-level variation, the strength of evidence risks being over- or under-estimated.

## The 'West Yorkshire Regional English Database': Investigations into the Generalizability of Reference Populations for Forensic Speaker Comparison Casework

*Erica Gold, Sula Ross and Kate Earnshaw*
Department of Linguistics and Modern Languages, University of Huddersfield, United Kingdom

**Notes**

**Wed-O-3-3-5, Time: 18:20-18:40**
The West Yorkshire Regional English Database (WYRED) consists of approximately 196 hours of high-quality audio recordings of 180 West Yorkshire (British English) speakers. All participants are male between the ages of 18-30 and are divided evenly (60 per region) across three boroughs within West Yorkshire (Northern England): Bradford, Kirklees and Wakefield. Speakers participated in four spontaneous speaking tasks. The first two tasks relate to a mock crime where the participant speaks to a police officer (Research Assistant 1) followed by an accomplice (Research Assistant 2). Speakers returned a minimum of a week later at which point they were paired with someone from their borough and recorded having a conversation on any topics they wish. The final task is an experimental task in which speakers are asked to leave a voicemail message related to the fictitious crime from the first recording session. In total, each speaker participated in approximately 1 hour of spontaneous speech recordings. This paper details the design of WYRED, in order to introduce forensic speech science research utilizing this data and to promote WYRED's potential application in related research and in forensic speech science casework.

### Studying Vowel Variation in French-Algerian Arabic Code-switched Speech

*Jane Wottawa[1], Amazouz Djegdjiga[1,2], Martine Adda-Decker[2] and Lori Lamel[1]*

[1]LIMSI, CNRS, Université Paris-Saclay, France

[2]LPP, UMR 7018 CNRS - U. Paris 3 / Sorbonne Nouvelle, France

**Wed-O-3-3-6, Time: 18:40-19:00**
Algerian Arabic-French bilinguals show phonetic variation with respect to vowel timber in both their languages. Our study aims to identify vowel variants frequently produced by Algerian Arabic-French bilinguals. To that end, the speech corpus FACST, containing French and Algerian Arabic code-switched speech, was analyzed. A second corpus with native French speakers (NCCFr) was used to provide a reference baseline and to compare vowel variants across the two speaker groups.

Three experiments were carried out: first, the French speech of both corpora was aligned with a French acoustic model including parallel vowel variants in its pronunciation dictionary. Second, the Arabic speech was aligned using using the same acoustic model as in the first experiment including parallel vowel variants in its dictionary. Finally, we tested whether peripheral vowels in Algerian Arabic-French bilinguals are more often centralized than in French native speech by allowing schwa as a competing variant.

The results show that, French natives and Algerian Arabic-French bilinguals have globally a comparable amount of vowel variation in French. However, French natives have stable high vowels whereas bilinguals tend to produce stable low and back vowels. In the centralization experiment, Algerian bilinguals particularly favor the centralization of mid, open and back vowels.

## Wed-O-3-4: Spoken Corpora and Annotation

MR G.03-G.04; 17:00-19:00; Wednesday, 5 September, 2018
Chairs: Joseph Mariani and Alexey Karpov

### Fearless Steps: Apollo-11 Corpus Advancements for Speech Technologies from Earth to the Moon

*John H.L. Hansen, Abhijeet Sangwan, Aditya Joglekar, Ahmet E. Bulut, Lakshmish Kaushik and Chengzhu Yu*

Center for Robust Speech Systems (CRSS), Eric Jonsson School of Engineering, The University of Texas at Dallas (UTD), Richardson, Texas, U.S.A

**Wed-O-3-4-1, Time: 17:00-17:20**
The Apollo Program is one the most significant benchmarks for technology and innovation in human history. The previously introduced UTD-CRSS Apollo initiative resulted in the digitization of the original analog audio tapes recorded during the Apollo Space Missions. This entire speech data is now being made publicly available with the release of the Fearless Steps Corpus. This corpus consists of a cumulative 19,000 hours of conversational speech spanning over thirty time-synchronized channels. With over six hundred speakers, the corpus has a rich collection of information which can be beneficial for research and advancement in the Speech and Language Community. Recent efforts on this data have led to the generation of pipeline diarization transcripts for the entire Speech Corpus. Research has also been done to address speech and natural language tasks such as speech activity detection, speech recognition and sentiment analysis. This paper provides an overview of the Fearless-Steps Corpus as well as a summary of previous research work achieved and highlights the factors that make the processing of this data a challenging problem. To initiate further development of algorithms on this Corpus, five challenge tasks are also organized. We also describe the challenge tasks with their associated transcriptions.

### A Knowledge Driven Structural Segmentation Approach for Play-Talk Classification During Autism Assessment

*Manoj Kumar[1], Pooja Chebolu[1], So Hyun Kim[2], Kassandra Martinez[2], Catherine Lord[2] and Shrikanth Narayanan[1]*

[1]University of Southern California, Los Angeles, United States

[2]Department of Psychiatry, Weill Cornell Medicine, New York, United States

**Wed-O-3-4-2, Time: 17:20-17:40**
Automatically segmenting conversational audio into semantically relevant components has both computational and analytical significance. In this paper, we segment play activities and conversational portions interspersed during clinically administered interactions between a psychologist and a child with autism spectrum disorder (ASD). We show that various acoustic-prosodic and turn-taking features commonly used in the literature differ between these segments and hence can possibly influence further inference tasks. We adopt a two-step approach for the segmentation problem by taking advantage of the structural relation between the two segments. First, we use a supervised machine learning algorithm to estimate class posteriors at frame-level. Next, we use an explicit-duration hidden Markov model (EDHMM) to align the states using the posteriors from the previous step. The durational distributions for both play and talk regions are learnt from training data and modeled using the EDHMM. Our results show that speech features can be used to successfully discriminate between play and talk activities, each providing important insights into the child's condition.

### An Open Source Emotional Speech Corpus for Human Robot Interaction Applications

*Jesin James, Li Tian and Catherine Inez Watson*

Department of Electrical and Computer Engineering, University of Auckland, New Zealand

**Wed-O-3-4-3, Time: 17:40-18:00**
For further understanding the wide array of emotions embedded in human speech, we are introducing a strictly-guided simulated emotional speech corpus. In contrast to existing speech corpora, this was constructed by maintaining an equal distribution of 4 long vowels in New Zealand English. This balance is to facilitate emotion related formant and glottal source feature comparison studies. Also, the corpus has 5 secondary emotions and 5 primary emotions. Secondary emotions are important in Human-Robot Interaction (HRI) to model natural conversations among humans and robots. But there are few existing speech resources to study these emotions, which has motivated the creation of this corpus. A large scale perception test with 120 participants showed that the corpus has approximately 70% and 40% accuracy in the correct classification of primary and secondary emotions respectively. The reasons behind the differences in perception accuracies of the two emotion types is further investigated. A preliminary prosodic analysis of corpus shows significant differences among the emotions. The corpus is made public at: github.com/tli725/JL-Corpus.

**Notes**

### Speech Database and Protocol Validation Using Waveform Entropy

*Itshak Lapidot[1], Héctor Delgado[2], Massimiliano Todisco[2], Nicholas Evans[2] and Jean-François Bonastre[3]*

[1]Afeka Tel-Aviv College of Engineering, ACLP, Israel

[2]EURECOM, Biot France

[3]University of Avignon, LIA, France

**Wed-O-3-4-4, Time: 18:00-18:20**

The assessment of performance for any number of speech processing tasks calls for the use of a suitably large, representative dataset. Dataset design is crucial so as to ensure that any significant variation unrelated to the task in hand is adequately normalised or marginalised. Most datasets are partitioned into training, development and evaluation subsets. Depending on the task, the nature of these three subsets should normally be close to identical. With speech signals being subject to a multitude of different influences, e.g. speaker gender and age, language, dialect, utterance length, etc., the design and validation of speech datasets can become especially challenging. Even if many sources of variation unrelated to the task in hand can easily be marginalised, other sources of more subtle variation can easily be overlooked. Imbalances between training, development and evaluation partitions, can bring into question findings derived from their use. Stringent dataset validation procedures are required. This paper reports a particularly straightforward approach to dataset validation that is based upon waveform entropy.

### A French-Spanish Multimodal Speech Communication Corpus Incorporating Acoustic Data, Facial, Hands and Arms Gestures Information

*Lucas D. Terissi[1], Gonzalo Sad[1], Mauricio Cerda[2], Slim Ouni[3], Rodrigo Galvez[2], Juan C. Gómez[1], Bernard Girau[3] and Nancy Hitschfeld-Kahler[4]*

[1]CIFASIS-CONICET, Universidad Nacional de Rosario, Argentina

[2]SCIAN-Lab, ICBM, Faculty of Medicine, Universidad de Chile, Santiago, Chile

[3]Université de Lorraine, LORIA, UMR 7503, Villers-lés-Nancy, France

[4]Computer Science Department, FCFyM, Universidad de Chile, Santiago, Chile

**Wed-O-3-4-5, Time: 18:20-18:40**

A Bilingual Multimodal Speech Communication Corpus incorporating acoustic data as well as visual data related to face, hands and arms gestures during speech, is presented in this paper. This corpus comprises different speaking modalities, including scripted text speech, natural conversation and free speech. The corpus has been compiled in two different languages, viz., French and Spanish. The experimental setups for the recording of the corpus, the acquisition protocols and the employed equipment are described. Statistics regarding the number and gender of the speakers, number of words, number of sentences and duration of the recording sessions, are also provided. Preliminary results from the analysis of the correlation among speech, head and hand movements during spontaneous speech are also presented in this paper, showing that acoustic prosodic features are related with head and hand gestures.

### L2-ARCTIC: A Non-native English Speech Corpus

*Guanlong Zhao[1], Sinem Sonsaat[2], Alif Silpachai[2], Ivana Lucic[2], Evgeny Chukharev-Hudilainen[2], John Levis[2] and Ricardo Gutierrez-Osuna[1]*

[1]Department of Computer Science and Engineering, Texas A&M University, United States

[2]Department of English, Iowa State University, United States

**Wed-O-3-4-6, Time: 18:40-19:00**

In this paper, we introduce L2-ARCTIC, a speech corpus of non-native English that is intended for research in voice conversion, accent conversion and mispronunciation detection. This initial release includes recordings from ten non-native speakers of English whose first languages (L1s) are Hindi, Korean, Mandarin, Spanish and Arabic, each L1 containing recordings from one male and one female speaker. Each speaker recorded approximately one hour of read speech from the Carnegie Mellon University ARCTIC prompts, from which we generated orthographic and forced-aligned phonetic transcriptions. In addition, we manually annotated 150 utterances per speaker to identify three types of mispronunciation errors: substitutions, deletions and additions, making it a valuable resource not only for research in voice conversion and accent conversion but also in computer-assisted pronunciation training. The corpus is publicly accessible at https://psi.engr.tamu.edu/l2-arctic-corpus/.

## Wed-SS-3-1: Special Session: The First DIHARD Speech Diarization Challenge

MR 1.01-1.02; 17:00-19:00; Wednesday, 5 September, 2018

Chairs: Kenneth Church, Christopher Cieri, Alejandrina Cristia, Jun Du, Sriram Ganapathy, Mark Liberman and Neville Ryant

### ZCU-NTIS Speaker Diarization System for the DIHARD 2018 Challenge

*Zbyněk Zajíc[1], Marie Kunešová[1,2], Jan Zelinka[1,2] and Marek Hrúz[1]*

[1]University of West Bohemia Faculty of Applied Sciences, NTIS - New Technologies for the Information Society, Univerzitní 8, 306 14 Plzeň, Czech Republic

[2]University of West Bohemia Faculty of Applied Sciences, Dept. of Cybernetics, Univerzitní 8, 306 14 Plzeň, Czech Republic

**Wed-SS-3-1-1, Time: 17:00-17:17**

In this paper, we present the system developed by the team from the New Technologies for the Information Society (NTIS) research center of the University of West Bohemia, for the First DIHARD Speech Diarization Challenge. The base of our system follows the currently-standard approach of segmentation, i-vector extraction, clustering and resegmentation. Here, we describe the modifications to the system which allowed us to apply it to data from a range of different domains. The main contribution to our achievement is an ANN-based domain classifier, which categorizes each conversation into one of the ten domains present in the development set. This classification determines the specific system configuration, such as the expected number of speakers and the stopping criterion for the hierarchical clustering. At the time of writing of this abstract, our best submission achieves a DER of 26.90% and an MI of 8.34 bits on the evaluation set (gold speech/nonspeech segmentation).

### Speaker Diarization with Enhancing Speech for the First DIHARD Challenge

*Lei Sun[1], Jun Du[1], Chao Jiang[2], Xueyang Zhang[2], Shan He[2], Bing Yin[2] and Chin-Hui Lee[3]*

[1]University of Science and Technology of China, Hefei, Anhui, P. R. China

[2]iFlytek Research, Hefei, Anhui, P. R. China

[3]Georgia Institute of Technology, Atlanta, GA. USA

**Wed-SS-3-1-2, Time: 17:17-17:34**

We design a novel speaker diarization system for the first DIHARD challenge by integrating several important modules of speech denoising, speech activity detection (SAD), i-vector design and scoring strategy. One main contribution is the proposed long short-term memory (LSTM) based speech denoising model. By fully utilizing the diversified simulated training data and advanced network architecture using progressive multitask learning with dense structure, the denoising model demonstrates the strong generalization capability to realistic noisy environments. The enhanced speech can boost the performance for the subsequent SAD, segmentation

**Notes**

and clustering. To the best of our knowledge, this is the first time we show significant improvements of deep learning based single-channel speech enhancement over state-of-the-art diarization systems in highly mismatch conditions. For the design of i-vector extraction, we adopt a residual convolutional neural network trained on large dataset including more than 30,000 people. Finally, by score fusion of different i-vectors based on all these techniques, our systems yield diarization error rates (DERs) of 24.56% and 36.05% on the evaluation sets of Track1 and Track2, which are both in the second place among 14 and 11 participating teams, respectively.

## BUT System for DIHARD Speech Diarization Challenge 2018

*Mireia Diez, Federico Landini, Lukáš Burget, Johan Rohdin, Anna Silnova, Kateřina Žmolíková, Ondřej Novotný, Karel Veselý, Ondřej Glembek, Oldřich Plchot, Ladislav Mošner and Pavel Matějka*

Brno University of Technology, Speech@FIT, Czechia

**Wed-SS-3-1-3, Time: 17:34-17:51**

This paper presents the approach developed by the BUT team for the first DIHARD speech diarization challenge, which is based on our Bayesian Hidden Markov Model with eigenvoice priors system. Besides the description of the approach, we provide a brief analysis of different techniques and data processing methods tested on the development set. We also introduce a simple attempt for overlapped speech detection that we used for attaining cleaner speaker models and reassigning overlapped speech to multiple speakers.

Finally, we present results obtained on the evaluation set and discuss findings we made during the development phase and with the help of the DIHARD leaderboard feedback.

## Estimation of the Number of Speakers with Variational Bayesian PLDA in the DIHARD Diarization Challenge

*Ignacio Viñals, Pablo Gimeno, Alfonso Ortega, Antonio Miguel and Eduardo Lleida*

ViVoLAB, Aragón Institute for Engineering Research (I3A), University of Zaragoza, Spain

**Wed-SS-3-1-4, Time: 17:51-18:08**

This paper focuses on the estimation of the number of speakers for diarization in the context of the DIHARD Challenge at InterSpeech 2018. This evaluation seeks the improvement of the diarization task in challenging corpora (YouTube videos, meetings, court audios, etc), containing an undetermined number of speakers with different relevance in terms of speech contributions.

Our proposal for the challenge is a system based on the i-vector PLDA paradigm: Given some initial segmentation of the input audio we extract i-vector representations for each acoustic fragment. These i-vectors are clustered with a Fully Bayesian PLDA. This model, a generative model with latent variables as speaker labels, produces the diarization labels by means of Variational Bayes iterations. The number of speakers is decided by comparing multiple hypotheses according to different information criteria. These criteria are developed around the Evidence Lower Bound (ELBO) provided by our PLDA.

## Diarization is Hard: Some Experiences and Lessons Learned for the JHU Team in the Inaugural DIHARD Challenge

*Gregory Sell, David Snyder, Alan McCree, Daniel Garcia-Romero, Jesús Villalba, Matthew Maciejewski, Vimal Manohar, Najim Dehak, Daniel Povey, Shinji Watanabe and Sanjeev Khudanpur*

Center for Language and Speech Processing & Human Language Technology Center of Excellence, Johns Hopkins University, USA

**Wed-SS-3-1-5, Time: 18:08-18:25**

We describe in this paper the experiences of the Johns Hopkins University team during the inaugural DIHARD diarization evaluation. This new task provided microphone recordings in a variety of difficult conditions and challenged researchers to fully consider all speaker activity, without the currently typical practices of unscored collars or ignored overlapping speaker segments. This paper explores several key aspects of currently state-of-the-art diarization methods, such as training data selection, signal bandwidth for feature extraction, representations of speech segments (i-vector versus x-vector) and domain-adaptive processing. In the end, our best system clustered x-vector embeddings trained on wideband microphone data followed by Variational-Bayesian refinement and a speech activity detector specifically trained for this task with in-domain data was found to be the best performing. After presenting these decisions and their final result, we discuss lessons learned and remaining challenges within the lens of this new approach to diarization performance measurement.

## The EURECOM Submission to the First DIHARD Challenge

*Jose Patino, Héctor Delgado and Nicholas Evans*

Department of Digital Security, EURECOM, Sophia Antipolis, France

**Wed-SS-3-1-6, Time: 18:25-18:42**

The first DIHARD challenge aims to promote speaker diarization research and to foster progress in domain robustness. This paper reports EURECOM's submission to the DIHARD challenge. It is based upon a low-resource, domain-robust binary key approach to speaker modelling. New contributions include the use of an infinite impulse response - constant Q Mel-frequency cepstral coefficient (ICMC) front-end, a clustering selection / stopping criterion algorithm based on spectral clustering and a mechanism to detect single-speaker trials. Experimental results obtained using the standard DIHARD database show that the contributions reported in this paper deliver relative improvements of 39% in terms of the diarization error rate over the baseline algorithm. An absolute DER of 29% on the evaluation set compares favourably with those of competing systems, especially given that the binary key system is highly efficient, running 63 times faster than real-time.

## Joint Discriminative Embedding Learning, Speech Activity and Overlap Detection for the DIHARD Speaker Diarization Challenge

*Valter Akira Miasato Filho[1], Diego Augusto Silva[1] and Luis Gustavo Depra Cuozzo[1,2]*

[1]CPqD, Campinas, Sao Paulo, Brazil

[2]Federal University of Technology, Curitiba, Parana, Brazil

**Wed-SS-3-1-7, Time: 18:42-19:00**

The DIHARD is a new, annual speaker diarization challenge focusing on "hard" domains, i.e. datasets in which current state-of-the-art systems are expected to perform poorly. We present our diarization system, which is a neural network jointly optimized for speaker embedding learning, speech activity and overlap detection. We present our network topology and the affinity matrix loss objective function responsible for learning the frame-wise speaker embeddings. The outputs of the network are then clustered with k-means and each frame classified with speech activity is assigned to one or two speakers, depending on the overlap detection. For the training data, we used two well-known meeting corpora - the AMI and the ICSI datasets, together with the provided samples from the DIHARD challenge. To further enhance our system, we present three data augmentation settings: the first is a naive concatenation of isolated speaker utterances from non-diarization datasets, which generates artificial diarization prompts. The second is a simple noise addition with sampled signal-to-noise ratios. The third is using noise suppression over the development data. All training setups are compared in terms of diarization error rate and mutual information in the evaluation set of the challenge.

**Notes**

# Wed-P-3-1: Text Analysis, Multilingual Issues and Evaluation in Speech Synthesis

Hall 4-6: Poster-1; 17:00-19:00; Wednesday, 5 September, 2018
Chair: Javier Latorre

## Multilingual Grapheme-to-Phoneme Conversion with Global Character Vectors

*Jinfu Ni, Yoshinori Shiga and Hisashi Kawai*
Advanced Speech Technology Laboratory, ASTREC, National Institute of Information and Communications Technology, Japan

**Wed-P-3-1-1, Time: 17:00-19:00**

Multilingual grapheme-to-phoneme (G2P) models are useful for multilingual speech synthesis because one model simultaneously copes with multilingual words. We propose a G2P model that combines global character vectors (GCVs) with bidirectional recurrent neural networks (BRNNs) and enables the direct conversion of text (as a sequence of characters) to pronunciation. GCVs are distributional, real-valued representations of characters and their contextual interactions that can be learned from a large-scale text corpus in an unsupervised manner. With the flexibility of learning GCVs from plain text resources, this method has an advantage: it enables monolingual G2P (MoG2P) and multilingual G2P (MuG2P) conversion.

We experiment in four languages (Japanese, Korean, Thai and Chinese) with learning language-dependent (LD) and language-independent (LI) GCVs and then build MoG2P and MuG2P models with two-hidden-layer BRNNs. Our results show that both LD- and LI-GCV-based MoG2P models, whose performances are equivalent, achieved better than 97.7% syllable accuracy, which is a relative improvement from 27% to 90% depending on the language in comparison with Mecab-based models. As for MuG2P, the accuracy is around 98%, which is a slightly degraded performance compared to MoG2P. The proposed method also has the potential of the G2P conversion of non-normalized words, achieving 80% accuracy in Japanese.

## A Hybrid Approach to Grapheme to Phoneme Conversion in Assamese

*Somnath Roy[1] and Shakuntala Mahanta[2]*
[1]Machine Learning Division, Melvault Software Solutions Pvt. Ltd., Hyderabad
[2]Department of Humanities and Social Science, IIT, Guwahati

**Wed-P-3-1-2, Time: 17:00-19:00**

Assamese is one of the low resource Indian languages. This paper implements both rule-based and data-driven grapheme to phoneme (G2P) conversion systems for Assamese. The rule-based system is used as the baseline which yields a word error rate of 35.3%. The data-driven systems are implemented using state-of-the-art sequence learning techniques such as —i) Joint-Sequence Model (JSM), ii) Recurrent Neural Networks with LTSM cell (LSTM-RNN) and iii) bidirectional LSTM (BiLSTM). The BiLSTM yields the lowest WER i.e., 18.7%, which is an absolute 16.6% improvement on the baseline system.

We additionally implement the rules of syllabification for Assamese. The surface output is generated in two forms namely i) phonemic sequence with syllable boundaries and ii) only phonemic sequence. The output of BiLSTM is fed as an input to Hybrid system. The Hybrid system syllabifies the input phonemic sequences to apply the vowel harmony rules. It also applies the rules of schwa-deletion as well as some rules in which the consonants change their form in clusters. The accuracy of the Hybrid system is 17.3% which is an absolute 1.4% improvement over the BiLSTM based G2P.

## Investigation of Using Disentangled and Interpretable Representations for One-shot Cross-lingual Voice Conversion

*Seyed Hamidreza Mohammadi and Taehwan Kim*
ObEN, Inc.

**Wed-P-3-1-3, Time: 17:00-19:00**

We study the problem of cross-lingual voice conversion in non-parallel speech corpora and one-shot learning setting. Most prior work require either parallel speech corpora or enough amount of training data from a target speaker. However, we convert an arbitrary sentences of an arbitrary source speaker to target speaker's given only one target speaker training utterance. To achieve this, we formulate the problem as learning disentangled speaker-specific and context-specific representations and follow the idea of [1] which uses Factorized Hierarchical Variational Autoencoder (FHVAE). After training FHVAE on multi-speaker training data, given arbitrary source and target speakers' utterance, we estimate those latent representations and then reconstruct the desired utterance of converted voice to that of target speaker. We investigate the effectiveness of the approach by conducting voice conversion experiments with varying size of training utterances and it was able to achieve reasonable performance with even just one training utterance. We also examine the speech representation and show that World vocoder outperforms Short-time Fourier Transform (STFT) used in [1]. Finally, in the subjective tests, for one language and cross-lingual voice conversion, our approach achieved significantly better or comparable results compared to VAE-STFT and GMM baselines in speech quality and similarity.

## Using Pupillometry to Measure the Cognitive Load of Synthetic Speech

*Avashna Govender and Simon King*
The Centre for Speech Technology Research, University of Edinburgh, United Kingdom

**Wed-P-3-1-4, Time: 17:00-19:00**

It is common to evaluate synthetic speech using listening tests in which intelligibility is measured by asking listeners to transcribe the words heard and naturalness is measured using Mean Opinion Scores. But, for real-world applications of synthetic speech, the effort (cognitive load) required to understand the synthetic speech may be a more appropriate measure. Cognitive load has been investigated in the past, when rule-based speech synthesizers were popular, but there is little or no recent work using state-of-the-art text-to-speech. Studies on the understanding of natural speech have shown that the pupil dilates when increased mental effort is exerted to perform a task. We use pupillometry to measure the cognitive load of synthetic speech submitted to two of the Blizzard Challenge evaluations. Our results show that pupil dilation is sensitive to the quality of synthetic speech. In all cases, synthetic speech imposes a higher cognitive load than natural speech. Pupillometry is therefore proposed as a sensitive measure that can be used to evaluate synthetic speech.

## Measuring the Cognitive Load of Synthetic Speech Using a Dual Task Paradigm

*Avashna Govender and Simon King*
The Centre for Speech Technology Research, University of Edinburgh, United Kingdom

**Wed-P-3-1-5, Time: 17:00-19:00**

We present a methodology for measuring the cognitive load (listening effort) of synthetic speech using a dual task paradigm. Cognitive load is calculated from changes in a listener's performance on a secondary task (e.g., reaction time to decide if a visually-displayed digit is odd or even). Previous related studies have only found significant differences between the best and worst quality systems but failed to separate the systems that lie in between. A paradigm that is sensitive enough to detect differences between state-of-the-art, high quality speech synthesizers would be very useful for advancing the state of the art. In our work, four speech synthesis systems from a previous Blizzard Challenge and the corresponding natural speech, were compared. Our results show that reaction times slow down

**Notes**

as speech quality reduces, as we expected: lower quality speech imposes a greater cognitive load, taking resources away from the secondary task. However, natural speech did not have the fastest reaction times. This intriguing result might indicate that, as speech synthesizers attain near-perfect intelligibility, this paradigm is measuring something like the listener's level of sustained attention and not listening effort.

## Attentive Sequence-to-Sequence Learning for Diacritic Restoration of YorùBá Language Text

*Iroro Orife*

Niger-Volta Language Technologies Institute, US

**Wed-P-3-1-6, Time: 17:00-19:00**

Yorùbá is a widely spoken West African language with a writing system rich in tonal and orthographic diacritics. With very few exceptions, diacritics are omitted from electronic texts, due to limited device and application support. Diacritics provide morphological information, are crucial for lexical disambiguation, pronunciation and are vital for any Yorùbá text-to-speech (TTS), automatic speech recognition (ASR) and natural language processing (NLP) tasks. Reframing Automatic Diacritic Restoration (ADR) as a machine translation task, we experiment with two different attentive Sequence-to-Sequence neural models to process undiacritized text. On our evaluation dataset, this approach produces diacritization error rates of less than 5%. We have released pre-trained models, datasets and source-code as an open-source project to advance efforts on Yorùbá language technology.

## Gated Convolutional Neural Network for Sentence Matching

*Peixin Chen, Wu Guo, Zhi Chen, Jian Sun and Lanhua You*

National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei, China

**Wed-P-3-1-7, Time: 17:00-19:00**

The recurrent neural networks (RNN) have shown promising results in sentence matching tasks, such as paraphrase identification (PI), natural language inference (NLI) and answer selection (AS). However, the recurrent architecture prevents parallel computation within a sequence and is highly time-consuming. To overcome this limitation, we propose a gated convolutional neural network (GCNN) for sentence matching tasks. In this model, the stacked convolutions encode hierarchical contextaware representations of a sentence, where the gating mechanism optionally controls and stores the convolutional contextual information. Furthermore, the attention mechanism is utilized to obtain interactive matching information between sentences. We evaluate our model on PI and NLI tasks and the experiments demonstrate the advantages of the proposed approach in terms of both speed and accuracy performance.

## On Training and Evaluation of Grapheme-to-Phoneme Mappings with Limited Data

*Dravyansh Sharma*

Google LLC, USA

**Wed-P-3-1-8, Time: 17:00-19:00**

When scaling to low resource languages for speech synthesis or speech recognition in an industrial setting, a common challenge is the absence of a readily available pronunciation lexicon. Common alternatives are handwritten letter-to-sound rules and data-driven grapheme-to-phoneme (G2P) models, but without a pronunciation lexicon it is hard to even determine their quality. We identify properties of a good quality metric and note drawbacks of naive estimates of G2P quality in the domain of small test sets. We demonstrate a novel method for reliable evaluation of G2P accuracy with minimal human effort. We also compare behavior of known state-of-the-art approaches for training with limited data. Finally we evaluate a new active learning approach for training G2P models in the low resource setting.

## The Perception and Analysis of the Likeability and Human Likeness of Synthesized Speech

*Alice Baird[1], Emilia Parada-Cabaleiro[1], Simone Hantke[1,2], Felix Burkhardt[3], Nicholas Cummins[1] and Björn Schuller[1,4]*

[1]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, Univeristy of Augsburg, Germany

[2]Machine Intelligence and Signal Processing Group, Technische Universität München, Germany

[3]Deutsche Telekom, Berlin, Germany

[4]GLAM – Group on Language, Audio and Music, Imperial College London, UK.

**Wed-P-3-1-9, Time: 17:00-19:00**

The synthesized voice has become an ever present aspect of daily life. Heard through our smart-devices and from public announcements, engineers continue in an endeavour to achieve naturalness in such voices. Yet, the degree to which these methods can produce likeable, human like voices, has not been fully evaluated. With recent advancements in synthetic speech technology suggesting that human like imitation is more obtainable, this study asked 25 listeners to evaluate both the likeability and human likeness of a corpus of 13 German male voices, produced via 5 synthesis approaches (from formant to hybrid unit selection, deep neural network systems) and 1 Human control. Results show that unlike visual artificially intelligent elements - as posed by the concept of the Uncanny Valley -likeability consistently improves along with human likeness for the synthesized voice, with recent methods achieving substantially closer results to human speech than older methods. A small scale acoustic analysis shows that the F0 of hybrid systems correlates less closely to human speech with a higher standard deviation for F0. This analysis suggests that limited variance in F0 is linked to a reduction in human likeness, resulting in lower likeability for conventional synthetic speech methods.

## Word Emphasis Prediction for Expressive Text to Speech

*Yosi Mass, Slava Shechtman, Moran Mordechay, Ron Hoory, Oren Sar Shalom, Guy Lev and David Konopnicki*

IBM Research, Haifa, Israel

**Wed-P-3-1-10, Time: 17:00-19:00**

Word emphasis prediction is an important part of expressive prosody generation in modern Text-To-Speech (TTS) systems. We present a method for predicting emphasized words for expressive TTS, based on a Deep Neural Network (DNN). We show that the presented method outperforms machine learning methods based on hand-crafted features in terms of objective metrics such as precision and recall. Using a listening test, we further demonstrate that the contribution of the predicted emphasized words to the expressiveness of the synthesized speech is subjectively perceivable.

## A Comparison of Speaker-based and Utterance-based Data Selection for Text-to-Speech Synthesis

*Kai-Zhan Lee, Erica Cooper and Julia Hirschberg*

Columbia University, USA

**Wed-P-3-1-11, Time: 17:00-19:00**

Building on previous work in subset selection of training data for text-to-speech (TTS), this work compares speaker-level and utterance-level selection of TTS training data, using acoustic features to guide selection. We find that speaker-based selection is more effective than utterance-based selection, regardless of whether selection is guided by a single feature or a combination of features. We use US English telephone data collected for automatic speech recognition to simulate the conditions of TTS training on low-resource languages. Our best voice achieves a human-evaluated WER of 29.0% on semantically-unpredictable sentences. This constitutes a significant improvement over our baseline voice trained on the same amount of randomly selected utterances, which performed at 42.4% WER. In addition to subjective voice evaluations with Amazon Mechanical Turk, we also explored objective voice evaluation using mel-cepstral distortion. We found that this measure correlates strongly with human evaluations of intelligibility, indicating that it may be a useful method to evaluate or pre-select voices in future work.

**Notes**

### Data Requirements, Selection and Augmentation for DNN-based Speech Synthesis from Crowdsourced Data

*Markus Toman, Geoffrey S. Meltzner and Rupal Patel*
VocaliD, Inc.

**Wed-P-3-1-12, Time: 17:00-19:00**

Crowdsourcing speech recordings provides unique opportunities and challenges for personalized speech synthesis as it allows gathering of large quantities of data but with a huge variety in quality. Manual methods for data selection and cleaning quickly become infeasible, especially when producing larger quantities of voices. We present and analyze approaches for data selection and augmentation to cope with this. For differently-sized training sets, we assess speaker adaptation by transfer learning, including layer freezing and sentence selection using maximum likelihood of forced alignment. The methodological framework utilizes statistical parametric speech synthesis based on Deep Neural Networks (DNNs). We compare objective scores for 576 voice models, representing all condition combinations. For a constrained set of conditions we also present results from a subjective listening test. We show that speaker adaptation improves overall quality in nearly all cases, sentence selection helps detecting recording errors and layer freezing proves to be ineffective in our system. We also found that while Mel-Cepstral Distortion (MCD) does not correlate with listener preference across the range of values, the most preferred voices also exhibited the lowest values for MCD. These findings have implications on scalable methods of customized voice building and clinical applications with sparse data.

## Wed-P-3-2: Neural Network Training Strategies for ASR

Hall 4-6: Poster-2; 17:00-19:00; Wednesday, 5 September, 2018
Chair: Martin Karafiat

### Lightly Supervised vs. Semi-supervised Training of Acoustic Model on Luxembourgish for Low-resource Automatic Speech Recognition

*Karel Veselý[1], Carlos Segura[2], Igor Szöke[1], Jordi Luque[2] and Jan Černocký[1]*
[1]BUT, Speech@FIT group and IT4I Center of Excellence, Brno, Czech Republic
[2]Telefonica Research, Barcelona, Spain

**Wed-P-3-2-1, Time: 17:00-19:00**

In this work, we focus on exploiting 'inexpensive' data in order to to improve the DNN acoustic model for ASR. We explore two strategies: The first one uses untranscribed data from the target domain. The second one is related to the proper selection of excerpts from imperfectly transcribed out-of-domain public data, as parliamentary speeches. We found out that both approaches lead to similar results, making them equally beneficial for practical use. The Luxembourgish ASR seed system had a 38.8% WER and it improved by roughly 4% absolute, leading to 34.6% for untranscribed and 34.9% for lightlysupervised data. Adding both databases simultaneously led to 34.4% WER, which is only a small improvement. As a secondary research topic, we experiment with semi-supervised state-level minimum Bayes risk (sMBR) training. Nonetheless, for sMBR we saw no improvement from adding the automatically transcribed target data, despite that similar techniques yield good results in the case of cross-entropy (CE) training.

### Investigation on the Combination of Batch Normalization and Dropout in BLSTM-based Acoustic Modeling for ASR

*Li Wenjie[1,3], Gaofeng Cheng[1,3], Fengpei Ge[1,3], Pengyuan Zhang[1,3] and Yonghong Yan[1,2,3]*
[1]Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics
[2]Xinjiang Laboratory of Minority Speech and Language Information Processing, Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, China
[3]University of Chinese Academy of Sciences

**Wed-P-3-2-2, Time: 17:00-19:00**

The Long Short-Term Memory (LSTM) architecture is a very special kind of recurrent neural network for modeling sequential data like speech. It has been widely used in the large scale acoustic model estimation recently and performs better than many other neural networks. Batch normalization(BN) is a good way to accelerate network training and improve the generalization performance of neural networks. However, applying batch normalization in the LSTM model is more complicated and challenging than in the feed-forward network. In this paper, we explored some novel approaches to add batch normalization to the LSTM model in bidirectional mode. Then we investigated some ways to combine the BN-BLSTM model with dropout, which is a traditional method to alleviate the overfitting problem in neural network training. We evaluated the proposed methods on several speech recognition tasks. Experiments show that the best performance on Switchboard task achieves 9.8% relative reduction on word error rate compared to the baseline, using the total Hub5'2000 evaluation dataset. Additionally, it is easy to implement and brings little extra computation.

### Inference-Invariant Transformation of Batch Normalization for Domain Adaptation of Acoustic Models

*Masayuki Suzuki, Tohru Nagano, Gakuto Kurata and Samuel Thomas*
IBM Research AI

**Wed-P-3-2-3, Time: 17:00-19:00**

Batch normalization, or batchnorm, is a popular technique often used to accelerate and improve training of deep neural networks. When existing models that use this technique via batchnorm layers, are used as initial models for domain adaptation or transfer learning, the novel input feature distributions of the adapted domains, considerably change the batchnorm transformations learnt in the training mode from those which are applied in the inference mode. We empirically find that this mismatch can degrade the performance of domain adaptation for acoustic modeling. To mitigate this degradation, we propose an inference-invariant transformation of batch normalization, a method which reduces the mismatch between training mode and inference mode transformations without changing the inference results. This invariance property is achieved by adjusting the weight and bias terms of the batchnorm to compensate for differences in the mean and variance terms when using the adaptation data. Experimental results show that our proposed method performs the best on several acoustic model adaptation tasks with up to 5% relative improvement in recognition performances in both supervised and unsupervised domain adaptation settings.

### Active Learning for LF-MMI Trained Neural Networks in ASR

*Yanhua Long[1], Hong Ye[1], Yijie Li[2] and Jiaen Liang[2]*
[1]SHNU-Unisound Joint Laboratory of Natural Human-Computer Interaction, Shanghai Normal University, Shanghai, China
[2]Beijing Unisound Information Technology Co., Ltd. Beijing, China

**Wed-P-3-2-4, Time: 17:00-19:00**

This paper investigates how active learning (AL) effects the training of neural network acoustic models based on Lattice-free Maximum Mutual Information (LF-MMI) in automatic speech recognition (ASR). To fully exploit the most informative examples from fresh datasets, different data selection criterions based on the heterogeneous neural networks were studied. In particular, we examined the relationship among the transcription cost of human labeling, example informativeness and data selection criterions for active learning. As a comparison, we tried both semi-supervised training (SST) and active learning to improve the acoustic models. Experiments

**Notes**

were performed for both the small-scale and large-scale ASR systems. Experimental results suggested that, our AL scheme can benefit much more from the fresh data than the SST in reducing the word error rate (WER).The AL yields 6~13% relative WER reduction against the baseline trained on a 4000 hours transcribed dataset, by only selecting 1.2K hrs informative utterances for human labeling via active learning.

## An Investigation of Mixup Training Strategies for Acoustic Models in ASR

*Ivan Medennikov[1,2], Yuri Khokhlov[1], Aleksei Romanenko[2,3], Dmitry Popov[3], Natalia Tomashenko[2,4], Ivan Sorokin[3] and Alexander Zatvornitskiy[1,2,3]*

[1]STC-innovations Ltd, St. Petersburg, Russia

[2]ITMO University, St. Petersburg, Russia

[3]Speech Technology Center Ltd, St. Petersburg, Russia

[4]LIUM, University of Le Mans, France

**Wed-P-3-2-5, Time: 17:00-19:00**

Mixup is a recently proposed technique that creates virtual training examples by combining existing ones. It has been successfully used in various machine learning tasks. This paper focuses on applying mixup to automatic speech recognition (ASR). More specifically, several strategies for acoustic model training are investigated, including both conventional cross-entropy and novel lattice-free MMI models. Considering mixup as a method of data augmentation as well as regularization, we compare it with widely used speed perturbation and dropout techniques. Experiments on Switchboard-1, AMI and TED-LIUM datasets shows consistent improvement of word error rate up to 13% relative. Moreover, mixup is found to be particularly effective on test data mismatched to the training data.

## Comparison of Unsupervised Modulation Filter Learning Methods for ASR

*Purvi Agrawal and Sriram Ganapathy*

Learning and Extraction of Acoustic Patterns Lab (LEAP), Dept. of Electrical Engg., Indian Institute of Science, Bengaluru-560012, India

**Wed-P-3-2-6, Time: 17:00-19:00**

The widespread deployment of automatic speech recognition (ASR) system in consumer centric applications such as voice interaction and voice search demands the need for noise robustness in such systems. One approach to this problem is to achieve the desired robustness in speech representations used in the ASR. Motivated from studies on robust human speech recognition, we analyse the unsupervised data-driven temporal modulation filter learning for robust feature extraction. In this paper, we compare various unsupervised models for data driven filter learning like convolutional autoencoder (CAE), generative adversarial network (GAN) and convolutional restricted Boltzmann machine (CRBM). The unsupervised models are designed to learn a set of filters from long temporal trajectories of speech sub-band energy. The filters learnt from these models are used for modulation filtering of the input spectrogram before the ASR training. The ASR experiments are performed on Wall Street Journal (WSJ) Aurora-4 database with clean and multi condition training setup. The experimental results obtained from the modulation filtered representations shows considerable robustness to noise, channel distortions and reverberant conditions compared to other feature extraction methods. Among the three approaches compared in this paper, the GAN approach provides the most consistent improvements in ASR accuracy in different training scenarios.

## Improved Training for Online End-to-End Speech Recognition Systems

*Suyoun Kim[1], Michael Seltzer[2], Jinyu Li[3] and Rui Zhao[3]*

[1]Carnegie Mellon University

[2]Facebook

[3]Microsoft AI & Research

**Wed-P-3-2-7, Time: 17:00-19:00**

Achieving high accuracy with end-to-end speech recognizers requires careful parameter initialization prior to training. Otherwise, the networks may fail to find a good local optimum. This is particularly true for online networks, such as unidirectional LSTMs. Currently, the best strategy to train such systems is to bootstrap the training from a tied-triphone system. However, this is time consuming and more importantly, is impossible for languages without a high-quality pronunciation lexicon. In this work, we propose an initialization strategy that uses teacher-student learning to transfer knowledge from a large, well-trained, offline end-to-end speech recognition model to an online end-to-end model, eliminating the need for a lexicon or any other linguistic resources. We also explore curriculum learning and label smoothing and show how they can be combined with the proposed teacher-student learning for further improvements. We evaluate our methods on a Microsoft Cortana personal assistant task and show that the proposed method results in a 19% relative improvement in word error rate compared to a randomly-initialized baseline system.

## Combining Natural Gradient with Hessian Free Methods for Sequence Training

*Adnan Haider and Philip Woodland*

Cambridge University Engineering Dept., Trumpington St., Cambridge, CB2 1PZ U.K.

**Wed-P-3-2-8, Time: 17:00-19:00**

This paper presents a new optimisation approach to train Deep Neural Networks (DNNs) with discriminative sequence criteria. At each iteration, the method combines information from the Natural Gradient (NG) direction with local curvature information of the error surface that enables better paths on the parameter manifold to be traversed. The method has been applied within a Hessian Free (HF) style optimisation framework to sequence train both standard fully-connected DNNs and Time Delay Neural Networks as speech recognition acoustic models. The efficacy of the method is shown using experiments on a Multi-Genre Broadcast (MGB) transcription task and neural networks using sigmoid and ReLU activation functions have been investigated. It is shown that for the same number of updates this proposed approach achieves larger reductions in the word error rate (WER) than both NG and HF and also leads to a lower WER than standard stochastic gradient descent.

## Lattice-free State-level Minimum Bayes Risk Training of Acoustic Models

*Naoyuki Kanda, Yusuke Fujita and Kenji Nagamatsu*

Hitachi Ltd., Japan

**Wed-P-3-2-9, Time: 17:00-19:00**

Lattice-free maximum mutual information (LF-MMI) training, which enables MMI-based acoustic model training without any lattice generation procedure, has recently been proposed. Although LF-MMI showed high accuracy in many tasks, its MMI criterion does not necessarily maximize the speech recognition accuracy. In this work, we propose a lattice-free state-level minimum Bayes risk training (LF-sMBR), which maximizes state-level expected accuracy without relying on a lattice generation procedure. As is the case with the LF-MMI, LF-sMBR avoids redundant lattice generation by exploiting forward-backward calculation on phone N-gram space, which enables a much simpler and faster training based on an sMBR criterion. We found that special care for silence phones was essential for improving the accuracy by LF-sMBR. In our experiments on the AMI, CSJ and Librispeech corpora, LF-sMBR achieved small but consistent improvements over LF-MMI AMs, showing state-of-the-art results for each test set.

## A Study of Enhancement, Augmentation and Autoencoder Methods for Domain Adaptation in Distant Speech Recognition

*Hao Tang, Wei-Ning Hsu, François Grondin and James Glass*

Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

**Wed-P-3-2-10, Time: 17:00-19:00**

Speech recognizers trained on close-talking speech do not generalize to distant speech and the word error rate degradation can be as large as 40% absolute. Most studies focus on tackling distant speech recognition as a separate problem, leaving little effort to adapting close-talking speech recognizers to distant speech. In this work, we review several approaches

**Notes**

from a domain adaptation perspective. These approaches, including speech enhancement, multi-condition training, data augmentation and autoencoders, all involve a transformation of the data between domains. We conduct experiments on the AMI data set, where these approaches can be realized under the same controlled setting. These approaches lead to different amounts of improvement under their respective assumptions. The purpose of this paper is to quantify and characterize the performance gap between the two domains, setting up the basis for studying adaptation of speech recognizers from close-talking speech to distant speech. Our results also have implications for improving distant speech recognition.

### Multilingual Deep Neural Network Training Using Cyclical Learning Rate

*Andreas Søeborg Kirkedal and Yeon-Jun Kim*
Interactions LLC, USA

**Wed-P-3-2-11, Time: 17:00-19:00**

Deep Neural Network (DNN) acoustic models are an essential component in automatic speech recognition (ASR). The main sources of accuracy improvements in ASR involve training DNN models that require large amounts of supervised data and computational resources. While the availability of sufficient monolingual data is a challenge for low-resource languages, the computational requirements for resource rich languages increases significantly with the availability of large data sets.

In this work, we provide novel solutions for these two challenges in the context of training a feed-forward DNN acoustic model (AM) for mobile voice search. To address the data-sparsity challenge, we bootstrap our multilingual AM using data from languages in the same language family. To reduce training time, we use cyclical learning rate (CLR) which has demonstrated fast convergence with competitive or better performance when training neural networks on tasks related to text and images.

We reduce training time for our Mandarin Chinese AM with 81.4% token accuracy from 40 to 21.3 hours and increase the word accuracy on three romance languages by 2-5% with multilingual AMs compared to monolingual DNN baselines.

## Wed-P-3-3: Application of ASR in Medical Practice

Hall 4-6: Poster-3; 17:00-19:00; Wednesday, 5 September, 2018
Chair: Phil Green

### Development of the CUHK Dysarthric Speech Recognition System for the UA Speech Corpus

*Jianwei Yu[1], Xurong Xie[2], Shansong Liu[1], Shoukang Hu[1], Max W. Y. Lam[1], Xixin Wu[1], Ka Ho Wong[1], Xunying Liu[1] and Helen Meng[1]*
[1]Department of Systems Engineering and Engineering Management
[2]Department of Electric Engineering
The Chinese University of Hong Kong, Hong Kong SAR, China

**Wed-P-3-3-1, Time: 17:00-19:00**

Dysarthric speech recognition is a highly challenging task. The articulatory motor control problems associated with neuro-motor conditions produces large mismatch against normal speech. In addition, such data is difficult to collect in large quantities. This paper presents an initial attempt at the Chinese University of Hong Kong to develop an automatic speech recognition (ASR) system for the Universal Access Speech (UASpeech) task. A range of deep neural network (DNN) acoustic models and their more advanced variants based on time delayed neural networks (TDNNs) and long short-term memory recurrent neural networks (LSTM-RNNs) were developed. Speaker adaptation by learning hidden unit contributions (LHUC) was used. A semi-supervised complementary auto-encoder system was further constructed to improve the bottleneck feature extraction. Two out-of-domain ASR systems separately trained on broadcast news and switchboard data were cross domain adapted to the UASpeech data and used in system combination. The final combined system gave an overall word accuracy of 69.4% on the 16-speaker test set.

### Automatic Evaluation of Speech Intelligibility Based on I-vectors in the Context of Head and Neck Cancers

*Imed Laaridh[1], Corinne Fredouille[1], Alain Ghio[2], Muriel Lalain[2] and Virginie Woisard[3]*
[1]University of Avignon, LIA, France
[2]Aix-Marseille Univ, CNRS, LPL UMR 7309, Aix-en-Provence, France
[3]Service ORL, CHU Larrey, URI Octogone-Lordat, Toulouse, France

**Wed-P-3-3-2, Time: 17:00-19:00**

In disordered speech context and despite its well-known subjectivity, perceptual evaluation is still the most commonly used method in clinical practice to evaluate the intelligibility level of patients' speech productions.

However and thanks to increasing computing power, automatic speech processing systems have witnessed a democratization in terms of users and application areas including the medical practice.

In this paper, we evaluate an automatic approach for the prediction of cancer patients' speech intelligibility based on the representation of the speech acoustics in the total variability subspace based on the i-vector paradigm. Experimental evaluations of the proposed predictive approach have shown a very high correlation rate with perceptual intelligibility when applied on the French speech corpora C2SI (r=0.84).

They have also demonstrated the robustness of the approach when using a limited amount of disordered speech per patient, which may lead to the redesign and alleviation of the test protocols usually used in disordered speech evaluation context.

### Dysarthric Speech Recognition Using Convolutional LSTM Neural Network

*Myungjong Kim[1], Beiming Cao[1], Kwanghoon An[1] and Jun Wang[1,2]*
[1]Speech Disorders & Technology Lab, Department of Bioengineering
[2]Callier Center for Communication Disorders, University of Texas at Dallas, United States

**Wed-P-3-3-3, Time: 17:00-19:00**

Dysarthria is a motor speech disorder that impedes the physical production of speech. Speech in patients with dysarthria is generally characterized by poor articulation, breathy voice and monotonic intonation. Therefore, modeling the spectral and temporal characteristics of dysarthric speech is critical for better performance in dysarthric speech recognition. Convolutional long short-term memory recurrent neural networks (CLSTM-RNNs) have recently successfully been used in normal speech recognition, but have rarely been used in dysarthric speech recognition. We hypothesized CLSTM-RNNs have the potential to capture the distinct characteristics of dysarthric speech, taking advantage of convolutional neural networks (CNNs) for extracting effective local features and LSTM-RNNs for modeling temporal dependencies of the features. In this paper, we investigate the use of CLSTM-RNNs for dysarthric speech recognition. Experimental evaluation on a database collected from nine dysarthric patients showed that our approach provides substantial improvement over both standard CNN and LSTM-RNN based speech recognizers.

### Perceptual and Automatic Evaluations of the Intelligibility of Speech Degraded by Noise Induced Hearing Loss Simulation

*Imed Laaridh[1], Julien Tardieu[2], Cynthia Magnen[2], Pascal Gaillard[3], Jérôme Farinas[1] and Julien Pinquier[1]*
[1]IRIT, Université de Toulouse, CNRS, Toulouse, France
[2]MSHS-T (USR 3414), Université de Toulouse, CNRS, France

**Notes**

[3]CLLE (UMR 5263) Université de Toulouse, CNRS, France

**Wed-P-3-3-4, Time: 17:00-19:00**

This study aims at comparing perceptual and automatic intelligibility measures on degraded speech. It follows a previous study that designed a novel approach for the automatic prediction of Age-Related Hearing Loss (ARHL) effects on speech intelligibility.

In this work, we adapted this approach to a different type of hearing disorder: the Noise Induced Hearing Loss (NIHL), i.e., hearing loss caused by noise exposure at work. Thus, we created a speech corpus made of both isolated words and short sentences pronounced by three speakers (male, female and child) and we simulated different levels of NIHL. A repetition task has been carried out with 60 participants to collect perceptual intelligibility scores. Then, an Automatic Speech Recognition (ASR) system has been designed to predict the perceptual scores of intelligibility. The perceptual evaluation showed similar effects of NIHL simulation on the male, female and child speakers. In addition, the automatic intelligibility measure, based on automatic speech recognition scores, was proven to well predict the effects of the different severity levels of NIHL. Indeed, high correlation coefficients were obtained between the automatic and perceptual intelligibility measures on both speech repetition tasks: 0.94 for isolated words task and 0.97 for sentences task.

## Articulatory Features for ASR of Pathological Speech

*Emre Yılmaz[1,2], Vikramjit Mitra[3], Chris Bartels[4] and Horacio Franco[4]*

[1]CLS/CLST, Radboud University, Nijmegen, Netherlands

[2]Dept. of Electrical and Computer Engineering, National University of Singapore, Singapore

[3]University of Maryland, College Park, MD, USA

[4]STAR Lab, SRI International, Menlo Park, CA, USA

**Wed-P-3-3-5, Time: 17:00-19:00**

In this work, we investigate the joint use of articulatory and acoustic features for automatic speech recognition (ASR) of pathological speech. Despite long-lasting efforts to build speaker- and text-independent ASR systems for people with dysarthria, the performance of state-of-the-art systems is still considerably lower on this type of speech than on normal speech. The most prominent reason for the inferior performance is the high variability in pathological speech that is characterized by the spectrotemporal deviations caused by articulatory impairments due to various etiologies. To cope with this high variation, we propose to use speech representations which utilize articulatory information together with the acoustic properties. A designated acoustic model, namely a fused-feature-map convolutional neural network (fCNN), which performs frequency convolution on acoustic features and time convolution on articulatory features is trained and tested on a Dutch and a Flemish pathological speech corpus. The ASR performance of fCNN-based ASR system using joint features is compared to other neural network architectures such conventional CNNs and time-frequency convolutional networks (TFCNNs) in several training scenarios.

## Mining Multimodal Repositories for Speech Affecting Diseases

*Joana Correia[1,2], Bhiksha Raj[1], Isabel Trancoso[2] and Francisco Teixeira[2]*

[1]Carnegie Mellon University, USA

[2]INESC-ID / Instituto Superior Técnico, University of Lisbon, Portugal

**Wed-P-3-3-6, Time: 17:00-19:00**

The motivation for this work is to contribute to the collection of large in-the-wild multimodal datasets in which the speech of the subject is affected by certain medical conditions. Our mining effort is focused on video blogs (vlogs) and as a proof-of-concept we have selected three target diseases: Depression, Parkinson's disease and cold.

Given the large scale nature of the online repositories, we take advantage of existing retrieval algorithms to narrow the pool of candidate videos for a given query related with the disease (e.g. depression vlog) and on top of that we apply several filtering techniques. These techniques explore both audio, video, text and metadata cues, in order to retrieve vlogs that include a single speaker which, at some point, admits that he/she is currently affected by a given disease. The use of straightforward NLP techniques on the automatically transcribed data showed that distinguishing between narratives of present and past experiences is harder than distinguishing between narratives of self experiences and of someone else's.

The three resulting speech datasets were tested with neural networks trained with speech data collected in controlled conditions, yielding results only slightly below the ones achieved with the original test datasets.

## Long Distance Voice Channel Diagnosis Using Deep Neural Networks

*Zhen Qin, Tom Ko and Guangjian Tian*

Huawei Noah's Ark Research Lab, Hong Kong, China

**Wed-P-3-3-7, Time: 17:00-19:00**

In long distance telephone network, it is time-consuming to detect and locate the problematic devices. Although hints could be given from the types of distortion in the test calls, it is tedious to manually classify the distortion types from a large number of calls.

In this paper, we present our work on using a deep neural network-based classifier, to automatically detect and identify the type of distortion which often occurs in long distance calls. We verified our approach with data from real telecommunication networks and the results showed that our approach can achieve an average recall rate of 71% in classification. We believe our method can lead to a huge reduction of manpower and time in long distance voice channel troubleshooting.

## Speech Recognition for Medical Conversations

*Chung-Cheng Chiu[1], Anshuman Tripathi[1], Katherine Chou[1], Chris Co[1], Navdeep Jaitly[1], Diana Jaunzeikare[1], Anjuli Kannan[1], Patrick Nguyen[1], Hasim Sak[1], Ananth Sankar[2], Justin Tansuwan[1], Nathan Wan, Yonghui Wu[1] and Xuedong Zhang*

[1]Google

[2]LinkedIn

**Wed-P-3-3-8, Time: 17:00-19:00**

In this paper we document our experiences with developing speech recognition for medical transcription – a system that automatically transcribes doctor-patient conversations. Towards this goal, we built a system along two different methodological lines – a Connectionist Temporal Classification (CTC) phoneme based model and a Listen Attend and Spell (LAS) grapheme based model. To train these models we used a corpus of anonymized conversations representing approximately 14,000 hours of speech. Because of noisy transcripts and alignments in the corpus, a significant amount of effort was invested in data cleaning issues. We describe a two-stage strategy we followed for segmenting the data. The data cleanup and development of a matched language model was essential to the success of the CTC based models. The LAS based models, however were found to be resilient to alignment and transcript noise and did not require the use of language models. CTC models were able to achieve a word error rate of 20.1% and the LAS models were able to achieve 18.3%. Our analysis shows that both models perform well on important medical utterances and therefore can be practical for transcribing medical conversations.

# Wed-P-3-4: Source and Supra-segmentals

Hall 4-6: Poster-4; 17:00-19:00; Wednesday, 5 September, 2018

Chair: T V Sreenivas

## Prosodic Focus Acquisition in French Early Cochlear Implanted Children

**Notes**

*Chadi Farah[1,2,3], Stephane Roman[1] and Mariapaola D'Imperio[2]*
[1]Pediatric Otolaryngology department, La Timone Children's Hospital (APHM), Marseille, France
[2]Aix-Marseille Univ, CNRS, Laboratoire Parole Langage, Aix-en-Provence, France
[3]Faculty of medicine, Saint-Joseph University, Beirut, Lebanon

**Wed-P-3-4-1, Time: 17:00-19:00**

Cochlear implanted (CI) children display an array of speech production and perception problems. No study has evaluated the specific use of prosody regarding information structure in the discourse, in French speaking early CI children.

This study aims to evaluate prosody production in these children, to determine whether they show prosodic effect on word duration.

We conducted a cross-sectional study of 10 prelingually hearing impaired French speaking children (4-7 years old), without comorbidities, CI before the age of 18 months between 2009 and 2012. The speech production task consisted in playing a computer-based semi-structured game, where children interacted with their caregiver. Results were interpreted according to both chronological age and hearing age (HA).

In our series, 6- and 7-year old children (HA<6.2 years) showed stronger lengthening of the focused word in the corrective narrow focus condition than in the contrastive narrow focus which in turn was stronger than in broad focus condition. Only 7-year old children adopted a strategy similar to that of adults, lengthening the end-phrase adjective to preserve the typical phrasing pattern of French.

This study shows for the first time that early CI children are able to acquire important intonation structure features comparable to adult patterns.

## The Role of Temporal Variation in Narrative Organization

*Nassima Fezza*
Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France

**Wed-P-3-4-2, Time: 17:00-19:00**

The aim of this study was to see if temporal variation can be considered a robust cue in the discourse structuring process. If so, at what level (s) of the discursive structure does it operate? In a bottom-up corpus-based approach, we analyze a 58-minute corpus of 60 natural French speech narratives. First, the corpus was segmented at the phonemic, syllabic, lexical and inter-pausal unit levels. Second, a narrative segmentation was applied using the criteria of Labov's evaluative model, which is based on semantic and informational criteria. Duration data was then extracted automatically at each level of granularity. The mapping of discourse segmentation to acoustic-phonetic analyses was made on two structural levels: micro and macro. A positive effect of local temporal variation in discourse structuring was not found, however, the existence of a link between the narrative internal segmentation and speech rate variation was identified. This variation is long-term, progressive and gradual which suggests a manipulation of this feature. In relation to the content, temporal values can be seen as contextual cues: relevant information is presented with a slower speech rate; while minor content is presented with a faster speech rate.

## Interaction Mechanisms between Glottal Source and Vocal Tract in Pitch Glides

*Tiina Murtola[1] and Jarmo Malinen[2]*
[1]Department of Signal Processing and Acoustics, Aalto University, Finland
[2]Department of Mathematics and Systems Analysis, Aalto University, Finland

**Wed-P-3-4-3, Time: 17:00-19:00**

A computational model for vowel production has been used to simulate rising pitch glides in the time domain. Such glides reveal multi-faceted nonlinear system behaviour when the fundamental frequency f_o is near the first vocal tract resonance f_{R1}. There are multiple physical mechanisms for how the acoustic field in the vocal tract can interact with vocal fold dynamics causing this behaviour. The model used in this work includes the direct impact of the acoustic pressure on the transversal plane of the vocal folds and an acoustic perturbation component to the glottal flow. Simulations indicate that both of these mechanisms, when applied separately, cause similar perturbations in phonation parameters when f_o crosses f_{R1}. Enabling both mechanisms simultaneously tends to make the separately emerging features more prominent. In simulated glottal flow waveforms, the tendency towards a formant ripple increases when acoustic feedback to glottal flow is enabled, whereas the phenomenon occurs more rarely as a result of the direct acoustic pressure to vocal folds. In all cases, the formant ripple is more pronounced for frequencies below f_{R1}.

## Relating Articulatory Motions in Different Speaking Rates

*Astha Singh[1], G. Nisha Meenakshi[2] and Prasanta Kumar Ghosh[2]*
[1]Electrical Communication Engineering
[2]Electrical Engineering
Indian Institute of Science, Bangalore-560012, India

**Wed-P-3-4-4, Time: 17:00-19:00**

Movements of articulators (e.g., tongue, lips and jaw) in different speaking rates are related in a complex manner. In this work, we examine the underlying function to transform articulatory movements involved in producing speech at a neutral speaking rate into those at fast and slow speaking rates (N2F and N2S). For this we use articulatory movement data collected from five subjects using an Electromagnetic articulograph at neutral, fast and slow speaking rates. As candidate transformation functions (TF), we use affine transformations with a diagonal matrix and a full matrix and a nonlinear function modeled by a deep neural network (DNN). Since the duration of an utterance in different speaking rates would typically be unequal, it is required to time align the articulatory movement trajectories, which, in turn, affects the TF learnt. Therefore, we propose an iterative algorithm to alternately optimize for the TF and the time alignments. Subject specific experiments reveal that while N2F transformation can be well described by an affine transformation with a full matrix, N2S transformation is better represented by a more complex nonlinear function modeled by a DNN. This could be because subjects exhibit gross articulatory movements during fast speech and hyper-articulate while producing slow speech.

## Estimation of the Asymmetry Parameter of the Glottal Flow Waveform Using the Electroglottographic Signal

*João Cabral*
The ADAPT Research Centre, Trinity College Dublin, Ireland

**Wed-P-3-4-5, Time: 17:00-19:00**

Glottal activity information can be very important in several speech processing applications, such as in speech therapy, voice disorder diagnosis, voice transformation and text-to-speech synthesis. However, the use of algorithms for estimating glottal parameters from the speech signal is very limited in those applications because of problems with robustness and accuracy. For this reason, current research studies of the glottal source are usually constrained to isolated speech sounds or short segments of speech recorded in controlled conditions and methods requiring manual intervention. An alternative way to obtain more accurate and reliable glottal parameter estimates is to use other recording equipment besides the audio microphone. Electroglottography is the most popular non-invasive measurement of vocal fold motion. It has been widely used to estimate the glottal opening and closing instants, but it does not provide direct information about the other important glottal parameters. This paper proposes an automatic method for estimation of the glottal parameters from the electroglottographic signal that permits to measure an additional parameter related to the asymmetry of the glottal flow pulse. This is a very important characteristic correlated with voice quality and widely studied in voice source analysis, commonly represented by the speed quotient parameter.

**Notes**

## Classification of Disorders in Vocal Folds Using Electroglottographic Signal

*Tanumay Mandal, K. Sreenivasa Rao and Sanjay Kumar Gupta*
Indian Institute of Technology Kharagpur, India

**Wed-P-3-4-6, Time: 17:00-19:00**

The main objective of this paper is to accurately classify the pathological voice based on the disorders in vocal folds. For this purpose, we have explored the phase of Electroglottographic (EGG) signal which carries significant information related to characteristics of vocal folds. Four important parameters, namely, close quotient, open quotient, average pitch period and jitter computed from the phase of the EGG signal have been explored for discriminating the patients based on the disorder in their vocal folds. These parameters have been used for classification of three types of vocal folds disorders: vocal nodules, vocal polyps and laryngitis. In this study, we have used the EGG signals of seventy-nine patients having disorders in vocal folds, collected from hospital. The database contains the simultaneous recording of speech and EGG signals of four vowel ('a','e','o','u') utterances. The experimental result shows that the proposed features extracted from phase, performed well in classification of patients according to the disorder in their
vocal folds.

## Automatic Glottis Localization and Segmentation in Stroboscopic Videos Using Deep Neural Network

*Achuth Rao MV[1], Rahul Krishnamurthy[2], Pebbili Gopikishore[3], Veeramani Priyadharshini[3] and Prasanta Kumar Ghosh[1] [1]Electrical Engineering, Indian Institute of Science, Bangalore 560012, India*
*[2]Kasturba medical college, Manipal Academy for Higher Education, Mangalore 575001, India*
*[3]All India Institute of Speech and Hearing, Mysuru, 570006, India*

**Wed-P-3-4-7, Time: 17:00-19:00**

Exact analysis of the glottal vibration patten is vital for assessing voice pathologies. One of the primary steps in this analysis is automatic glottis segmentation, which, in turn, has two main parts, namely, glottis localization and the glottis segmentation. In this paper, we propose a deep neural network (DNN) based automatic glottis localization and segmentation scheme. We pose the problem as a classification problem where colors of each pixel and its neighborhood is classified as belonging to inside or outside the glottis region. We further process the classification result to get the biggest cluster, which is declared as the segmented glottis. The proposed algorithm is evaluated on a dataset comprising of stroboscopic videos from 18 subjects where the glottis region is marked by the three Speech Language Pathologists (SLPs). On average, the proposed DNN based segmentation scheme achieves a localization performance of 65.33% and segmentation DICE score of 0.74 (absolute), which is better than the baseline scheme by 22.66% and 0.09 respectively. We also find that the DICE score obtained by the DNN based segmentation scheme correlates well with the average DICE score computed between annotation provided by any two SLPs suggesting the robustness of the proposed glottis segmentation scheme.

## Respiratory and Respiratory Muscular Control in JL1's and JL2's Text Reading Utilizing 4-RSTs and a Soft Respiratory Mask with a Two-Way Bulb

*Toshiko Isei-Jaakkola[1], Keiko Ochi[2] and Keikichi Hirose[3]*
[1]Chubu University, Japan
[2]Tokyo University of Technology, Japan
[3]National Institute of Informatics, Japan

**Wed-P-3-4-8, Time: 17:00-19:00**

To investigate how respiratory muscles and respiration are controlled during L1 and L2 text readings, experiments were conducted to acquire data on (A) upper- and lower- chest and upper- and lower- abdominal movements, utilizing four respiratory strain-gauge transducers (4-RSTs) and (B) inspiratory and expiratory volumes, utilizing a soft respiratory mask with a two-way bulb. Speech were recorded for (A). In addition, the subjects' pulmonic vital capacities were measured. Five male students read two kinds of text materials in either Japanese (L1) or English (L2) five times. Thus, we acquired 200 data for chest and abdominal movements, 100 data for respiration and 100 data for the recorded voice in addition to his pulmonic data and made quantitative and statistical analyses. Our findings were: (1) text reading speed was more stable in L1 than in L2, (2) the lower abdomen is controllable. This may indicate "stomach respiration," (3) the inspiratory and expiratory controls during speech differed largely from those in quiet breathing, indicating acute active muscular movements, (4) the vital capacity volume did not sufficiently correlate with the subjects' expiratory and inspiratory air volume during speech and (5) a better English pronunciation might be supported by alternative control of inspiration and respiration.

## A Preliminary Study on Tonal Coarticulation in Continuous Speech

*Lixia Hao[1,2], Wei Zhang[1], Yanlu Xie[1] and Jinsong Zhang[1]*
[1]Beijing Advanced Innovation Center for Language Resources, Beijing Language and Culture University, Beijing, China
[2]Dongguan University of Technology, Dongguan, China

**Wed-P-3-4-9, Time: 17:00-19:00**

Tonal variations in continuous speech are complicated in nature and it is a challenge to identify the effect of tonal coarticulation given several influencing factors. To address the issue, the present study proposes a scheme for labeling tonal coarticulation in Mandarin continuous speech by applying Hypo- and Hyper-articulation theory. We assume that the bidirectional tonal coarticulation (both carryover and anticipatory effects) as patterns of Hypo-articulation results from the economical articulatory rule. The effects may partially disappear under the influence of specific stress patterns and become unidirectional (carryover or anticipatory). At a prosodic boundary, the effects of tonal coarticulation may completely disappear and lead to the occurrence of patterns of Hyper-articulation. Based on the scheme, we have labeled the data in the Annotated Speech Corpus of Chinese Discourse. It is shown that: three annotators are consistent at a fairly high level (86.2%) on average and the acoustic parameters of four kinds of tonal coarticulation are significantly different. Therefore, we conclude that the proposal is feasible for investigating tonal coarticulation in Mandarin continuous speech. Though the labeling scheme is language dependent, it may well have cross-linguistic implications.

# Plenary Talk-3

Hall 3; 8:30-9:30; Thursday, 6 September, 2018
Chair: Preeti Rao

## Speech and Language Processing for Learning and Wellbeing

*Helen Meng*
Department of Systems Engineering and Engineering Management, Chinese University of Hong Kong

**Thu-Plenary-3, Time: 08:30-09:30**

Spoken language is a primary form of human communication. Spoken language processing techniques must incorporate knowledge of acoustics, phonetics and linguistics in analyzing speech. While great strides have been made in the community in general speech recognition, reaching human parity in performance, our team has been focusing on the problems of recognizing and analyzing non-native, learners' speech for the purpose of mispronunciation detection and diagnosis in computer-aided pronunciation training. In order to generate personalized, corrective feedback, we have also developed an approach that uses phonetic posterior-grams (PPGs) for personalized, cross-lingual text-to-speech synthesis given arbitrary textual input, based on voice conversion techniques. We have also extended our work to disordered speech, focusing on automated distinctive feature (DF)-based analyses of dysarthric recordings. The analyses are intended to inform intervention strategies. Additionally, voice conversion is further developed to restore disordered speech to normal speech. This talk will present the challenges in these problems, our approaches and solutions, as well as our ongoing work.

**Notes**

# Thu-O-1-1: Distant ASR

Hall 3; 10:00-12:00; Thursday, 6 September, 2018

Chairs: Sri Garimella and Martin Hagmuller

## Far-Field Speech Recognition Using Multivariate Autoregressive Models

*Sriram Ganapathy[1] and Madhumita Harish[2]*

[1]Learning and Extraction of Acoustic Patterns (LEAP) Lab, Indian Institute of Science, Bangalore

[2]Carnegie Mellon University, Pittsburgh, U.S.A.

**Thu-O-1-1-1, Time: 10:00-10:20**

Automatic speech recognition in far-field reverberant environments is challenging even with the state-of-the-art recognition systems. The main issues are artifacts in the signal due to the long-term reverberation that results in temporal smearing. The autoregressive modeling approach to speech feature extraction involves representing the high energy regions of the signal which are less susceptible to noise. In this paper, we propose a novel method of speech feature extraction using multivariate AR modeling (MAR) of temporal envelopes. The sub-band discrete cosine transform coefficients obtained from multiple speech bands are used in a multivariate linear prediction setting to derive features for speech recognition. For single channel far-field speech recognition, the features are derived using multi-band linear prediction. In the case of multi-channel far-field speech recognition, we use the multi-channel data in the MAR framework. We perform several speech recognition experiments in the REVERB Challenge database for single and multi-microphone settings. In these experiments, the proposed feature extraction method provides significant improvements over baseline methods (average relative improvements of 9.7% and 3.9% in single microphone conditions for clean and multi-conditions respectively and 6.3% in multi-microphone conditions). The results with clean training on single microphone conditions further illustrates the effectiveness of the MAR.

## Efficient Implementation of the Room Simulator for Training Deep Neural Network Acoustic Models

*Chanwoo Kim[1], Ehsan Variani[2], Arun Narayanan[2] and Michiel Bacchiani[2]*

[1]Samsung Research

[2]Google Speech

**Thu-O-1-1-2, Time: 10:20-10:40**

In this paper, we describe how to efficiently implement an acoustic room simulator to generate large-scale simulated data for training deep neural networks. Even though Google Room Simulator in [1] was shown to be quite effective in reducing the Word Error Rates (WERs) for far-field applications by generating simulated far-field training sets, it requires a very large number of FFTs. Room Simulator used approximately 80% of CPU usage in our CPU/GPU training architecture [2]. In this work, we implement an efficient OverLap Addition (OLA) based filtering using the open-source FFTW3 library. Further, we investigate the effects of the Room Impulse Response (RIR) lengths. Experimentally, we conclude that we can cut the tail portions of RIRs whose power is less than 20 dB below the maximum power without sacrificing the speech recognition accuracy. However, we observe that cutting RIR tail more than this threshold harms the speech recognition accuracy for rerecorded test sets. Using these approaches, we were able to reduce CPU usage for the room simulator portion down to 9.69% in CPU/GPU training architecture. Profiling result shows that we obtain 22.4 times speed-up on a single machine and 37.3 times speed up on Google's distributed training infrastructure.

## Stream Attention for Distributed Multi-Microphone Speech Recognition

*Xiaofei Wang[1,2], Ruizhi Li[1] and Hynek Hermansky[1]*

[1]Center for Language and Speech Processing, Johns Hopkins University

[2]Institute of Acoustics, Chinese Academy of Sciences

**Thu-O-1-1-3, Time: 10:40-11:00**

Exploiting multiple microphones has been a widely-used strategy for robust automatic speech recognition (ASR). Particularly, in a general hands-free scenario, acquisition of speech usually happens using a set of distributed microphones or arrays simultaneously. Each microphone or array (defined as a stream) carries a different quality of information. The technique of stream fusion is beneficial to provide the best distant recognition performance against the effects of potential disturbances such as noise, reverberation, as well as the speaker movement.
In this work, we propose a stream attention framework to improve the far-field ASR performance in the distributed multi-microphone configuration.

Frame-level attention vectors have been derived by predicting the ASR performance of the acoustic modeling of individual streams using the posterior probabilities from the classifier. They are used to characterize the amount of useful information each stream contributes, for the purpose of an efficient and better-performing decoding scheme.

In this paper, we investigate the ASR performance measures using our proposed stream attention system on real recorded datasets, Mixer-6 and DIRHA-WSJ.

The experimental results show that the proposed framework yields substantial improvements in word error rate (WER) compared to conventional strategies.

## Recognizing Overlapped Speech in Meetings: A Multichannel Separation Approach Using Neural Networks

*Takuya Yoshioka, Hakan Erdogan, Zhuo Chen, Xiong Xiao and Fil Alleva*

Microsoft AI and Research, One Microsoft Way, Redmond, WA, USA

**Thu-O-1-1-4, Time: 11:00-11:20**

The goal of this work is to develop a meeting transcription system that can recognize speech even when utterances of different speakers are overlapped. While speech overlaps have been regarded as a major obstacle in accurately transcribing meetings, a traditional beamformer with a single output has been exclusively used because previously proposed speech separation techniques have critical constraints for application to real meetings. This paper proposes a new signal processing module, called an unmixing transducer and describes its implementation using a windowed BLSTM. The unmixing transducer has a fixed number, say J, of output channels, where J may be different from the number of meeting attendees and transforms an input multi-channel acoustic signal into J time-synchronous audio streams. Each utterance in the meeting is separated and emitted from one of the output channels. Then, each output signal can be simply fed to a speech recognition back-end for segmentation and transcription. Our meeting transcription system using the unmixing transducer outperforms a system based on a state-of-the-art neural mask-based beamformer by 10.8%. Significant improvements are observed in overlapped segments. To the best of our knowledge, this is the first report that applies overlapped speech recognition to unconstrained real meeting audio.

## Integrating Neural Network Based Beamforming and Weighted Prediction Error Dereverberation

*Lukas Drude[1], Christoph Boeddeker[1], Jahn Heymann[1], Reinhold Haeb-Umbach[1], Keisuke Kinoshita[2], Marc Delcroix[2] and Tomohiro Nakatani[2]*

[1]Paderborn University, Department of Communications Engineering, Paderborn, Germany

[2]NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan

**Thu-O-1-1-5, Time: 11:20-11:40**

The weighted prediction error (WPE) algorithm has proven to be a very

**Notes**

successful dereverberation method for the REVERB challenge. Likewise, neural network based mask estimation for beamforming demonstrated very good noise suppression in the CHiME 3 and CHiME 4 challenges. Recently, it has been shown that this estimator can also be trained to perform dereverberation and denoising jointly. However, up to now a comparison of a neural beamformer and WPE is still missing, so is an investigation into a combination of the two. Therefore, we here provide an extensive evaluation of both and consequently propose variants to integrate deep neural network based beamforming with WPE. For these integrated variants we identify a consistent WER reduction on two distinct databases. In particular, our study shows that deep learning based beamforming benefits from a model-based dereverberation technique (i.e. WPE) and vice versa. Our key findings are: (a) Neural beamforming yields the lower WER in comparison to WPE the more channels and noise are present. (b) Integration of WPE and a neural beamformer consistently outperforms all stand-alone systems.

### A Probability Weighted Beamformer for Noise Robust ASR

*Suliang Bu[1], Yunxin Zhao[1], Meiyuh Hwang[2] and Sining Sun[3]*

[1]Dept. of Electrical Engineering and Computer Science, University of Missouri-Columbia, USA

[2]Mobvoi AI Lab, Redmond WA, USA

[3]Sch. of Computer Science, Northwestern Polytechnical University, Xi'an, China

**Thu-O-1-1-6, Time: 11:40-12:00**

We investigate a novel approach to spatial filtering that is adaptive to conditions at different time-frequency (TF) points for noise removal by taking advantage of speech sparsity. Our approach combines a noise reduction beamformer with a minimum variance distortionless response (MVDR) beamformer or Generalized Eigenvalue (GEV) beamformer through TF posterior probabilities of speech presence (PPSP). To estimate PPSP, we study both statistical model-based and neural network based methods, where in the former, we use complex Gaussian mixture modeling (CGMM) on temporally augmented spatial spectral features and in the latter, we use neural network (NN) based TF masks to initialize speech and noise covariance matrices in CGMM. We have conducted experiments on CHiME-3 task. On its real noisy speech test set, our methods of feature augmentation, TF dependent spatial filter and NN-based mask initialization on covariances for CGMM have yielded relative word error rate (WER) reductions cumulatively by 8%, 16% and 25% over the original CGMM based MVDR. On the real test data, the three methods have also produced consistent WER reductions when replacing MVDR by GEV.

## Thu-O-1-2: Expressive Speech Synthesis

Hall 1; 10:00-12:00; Thursday, 6 September, 2018

Chairs: Peter Cahill and Oliver Watts

### Effects of Dimensional Input on Paralinguistic Information Perceived from Synthesized Dialogue Speech with Neural Network

*Masaki Yokoyama, Tomohiro Nagata and Hiroki Mori*

Graduate School of Engineering, Utsunomiya University, Japan

**Thu-O-1-2-1, Time: 10:00-10:20**

A novel method of controlling paralinguistic information in neural network-based dialogue speech synthesis is proposed. Controlling paralinguistic information was achieved by feeding emotion dimensions in continuous values into the input layer of the neural networks. Compared to the method using the multiple regression HMM, the naturalness of synthesized speech was improved. The controllability of paralinguistic information was evaluated by examining the shift of the distribution of synthesized parameters. A subjective evaluation test revealed that the correlation between given and perceived paralinguistic information was moderate, though less apparent compared to the multiple regression HMM-based method.

**Notes**

### Neural MultiVoice Models for Expressing Novel Personalities in Dialog

*Shereen Oraby, Lena Reed, Sharath T.S., Shubhangi Tandon and Marilyn Walker*

Natural Language and Dialog Systems Lab, University of California, Santa Cruz

**Thu-O-1-2-2, Time: 10:20-10:40**

Natural language generators for task-oriented dialog should be able to vary the style of the output utterance while still effectively realizing the system dialog actions and their associated semantics. While the use of neural generation for training the response generation component of conversational agents promises to simplify the process of producing high quality responses in new domains, to our knowledge, there has been very little investigation of neural generators for task-oriented dialog that can vary their response style and we know of no experiments on models that can generate responses that are different in style from those seen during training, while still maintaining semantic fidelity to the input meaning representation. Here, we show that a model that is trained to achieve a single stylistic personality target can produce outputs that combine stylistic targets. We carefully evaluate the multivoice outputs for both semantic fidelity and for similarities to and differences from the linguistic features that characterize the original training style. We show that contrary to our predictions, the learned models do not always simply interpolate model parameters, but rather produce styles that are distinct and novel from the personalities they were trained on.

### Expressive Speech Synthesis Using Sentiment Embeddings

*Igor Jauk[1], Jaime Lorenzo-Trueba[2], Junichi Yamagishi[2] and Antonio Bonafonte[1]*

[1]Universitat Politècnica de Catalunya, Barcelona, Spain

[2]National Institute of Informatics, Tokyo, Japan

**Thu-O-1-2-3, Time: 10:40-11:00**

In this paper we present a DNN based speech synthesis system trained on an audiobook including sentiment features predicted by the Stanford sentiment parser.

The baseline system uses DNN to predict acoustic parameters based on conventional linguistic features, as they have been used in statistical parametric speech synthesis.

The predicted parameters are transformed into speech using a conventional high-quality vocoder. In the proposed system the conventional linguistic features are enriched using sentiment features. Different sentiment representations have been considered, combining sentiment probabilities with hierarchical distance and context.

After preliminary analysis a listening experiment is conducted, where participants evaluate the different systems. The results show the usefulness of the proposed features and reveal differences between expert and non-expert TTS user.

### Expressive Speech Synthesis via Modeling Expressions with Variational Autoencoder

*Kei Akuzawa, Yusuke Iwasawa and Yutaka Matsuo*

Graduate School of Engineering, The University of Tokyo, Japan

**Thu-O-1-2-4, Time: 11:00-11:20**

Recent advances in neural autoregressive models have improve the performance of speech synthesis (SS). However, as they lack the ability to model global characteristics of speech (such as speaker individualities or speaking styles), particularly when these characteristics have not been labeled, making neural autoregressive SS systems more expressive is still an open issue. In this paper, we propose to combine VoiceLoop, an autoregressive SS model, with Variational Autoencoder (VAE). This approach, unlike traditional autoregressive SS systems, uses VAE to model the global characteristics explicitly, enabling the expressiveness of the synthesized speech to be controlled in an unsupervised manner. Experiments using the VCTK and Blizzard2012 datasets show the VAE helps VoiceLoop to generate

higher quality speech and to control the experssions in its synthesized speech by incorporating global characteristics into the speech generating process.

### Rapid Style Adaptation Using Residual Error Embedding for Expressive Speech Synthesis

*Xixin Wu[1], Yuewen Cao[1], Mu Wang[2], Songxiang Liu[1], Shiyin Kang[3], Zhiyong Wu[1,2], Xunying Liu[1], Dan Su[3], Dong Yu[3] and Helen Meng[1,2]*

[1]Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, China

[2]Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University, Shenzhen, China

[3]Tencent AI Lab, Tencent, Shenzhen, China

**Thu-O-1-2-5, Time: 11:20-11:40**

Synthesizing expressive speech with appropriate prosodic variations, e.g., various styles, still has much room for improvement. Previous methods have explored to use manual annotations as conditioning attributes to provide variation information. However, the related training data are expensive to obtain and the annotated style codes can be ambiguous and unreliable. In this paper, we explore utilizing the residual error as conditioning attributes. The residual error is the difference between the prediction of a trained average model and the ground truth. We encode the residual error into a style embedding via a neural network-based error encoder. The embedding is then fed to the target synthesis model to provide information for modeling various style distributions more accurately. The average model and the error encoder are jointly optimized with the target synthesis model. Our proposed method has two advantages: 1) the embedding is automatically learned with no need of manual annotations, which helps overcome data sparsity and ambiguity limitations; 2) For any unseen audio utterance, the style embedding can be efficiently generated. This enables rapid adaptation to the desired style to be achieved with only one adaptation utterance. Experimental results show that our method outperforms the baseline in speech quality and style similarity.

### EMPHASIS: An Emotional Phoneme-based Acoustic Model for Speech Synthesis System

*Hao Li, Yongguo Kang and Zhenyu Wang*

Baidu Speech Department

Baidu Inc. Baidu Technology Park, Beijing, 100193, China

**Thu-O-1-2-6, Time: 11:40-12:00**

We present EMPHASIS, an emotional phoneme-based acoustic model for speech synthesis system. EMPHASIS includes a phoneme duration prediction model and an acoustic parameter prediction model. It uses a CBHG-based regression network to model the dependencies between linguistic features and acoustic features. We modify the input and output layer structures of the network to improve the performance. For the linguistic features, we apply a feature grouping strategy to enhance emotional and prosodic features. The acoustic parameters are designed to be suitable for the regression task and waveform reconstruction. EMPHASIS can synthesize speech in real-time and generate expressive interrogative and exclamatory speech with high audio quality. EMPHASIS is designed to be a multi-lingual model and can synthesize Mandarin-English speech for now. In the experiment of emotional speech synthesis, it achieves better subjective results than other real-time speech synthesis systems.

## Thu-O-1-3: Representation Learning for Emotion

Hall 2; 10:00-12:00; Thursday, 6 September, 2018

Chairs: Nicholas Cummins and Julien Epps

### Bags in Bag: Generating Context-Aware Bags for Tracking Emotions from Speech

*Jing Han[1], Zixing Zhang[2], Maximilian Schmitt[1], Zhao Ren[1], Fabien Ringeval[3] and Björn Schuller[1,2]*

[1]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany

[2]GLAM – Group on Language, Audio & Music, Imperial College London, UK

[3]Laboratoire d'Informatique de Grenoble, Université Grenoble Alpes, France

**Thu-O-1-3-1, Time: 10:00-10:20**

Whereas systems based on deep learning have been proposed to learn efficient representations of emotional speech data, methods such as Bag-of-Audio-Words (BoAW) have yielded similar or even better performance while providing understandable representations of the data. In those representations, however, context information is overlooked as the BoAW include only local information. In this paper, we propose to learn a novel representation 'Bag-of-Context-Aware-Words' that encapsulates the context with neighbouring frames of BoAW; segment-level BoAW are extracted in the first layer which are then utilised to create a final instance-level bag. Such a hierarchical structure of BoAW enables the system to learn representations with context information. To evaluate the effectiveness of the method, we perform extensive experiments on a time- and value-continuous spontaneous emotion database: RECOLA. The results show that, the best segment length for valence is twice of that for arousal, suggesting that the prediction of the emotional valence requires more context information than the prediction of arousal and the performance obtained on RECOLA with the proposed Bag-of-Context-Aware-Words outperforms all previously reported results.

### An Attention Pooling Based Representation Learning Method for Speech Emotion Recognition

*Pengcheng Li[1], Yan Song[1], Ian McLoughlin[2], Wu Guo[1] and Lirong Dai[1]*

[1]National Engineering Laboratory of Speech and Language Information Processing, University of Science and Technology of China, Hefei, China

[2]School of Computing, University of Kent, Medway, UK

**Thu-O-1-3-2, Time: 10:20-10:40**

This paper proposes an attention pooling based representation learning method for speech emotion recognition (SER). The emotional representation is learned in an end-to-end fashion by applying a deep convolutional neural network (CNN) directly to spectrograms extracted from speech utterances. Motivated by the success of GoogLeNet, two groups of filters with different shapes are designed to capture both temporal and frequency domain context information from the input spectrogram. The learned features are concatenated and fed into the subsequent convolutional layers. To learn the final emotional representation, a novel attention pooling method is further proposed. Compared with the existing pooling methods, such as max-pooling and average-pooling, the proposed attention pooling can effectively incorporate class-agnostic bottom-up and class-specific top-down, attention maps. We conduct extensive evaluations on benchmark IEMOCAP data to assess the effectiveness of the proposed representation. Results demonstrate a recognition performance of 71.8% weighted accuracy (WA) and 68% unweighted accuracy (UA) over four emotions, which outperforms the state-of-the-art method by about 3% absolute for WA and 4% for UA.

**Notes**

## Predicting Arousal and Valence from Waveforms and Spectrograms Using Deep Neural Networks

*Zixiaofan Yang and Julia Hirschberg*

Columbia University

**Thu-O-1-3-3, Time: 10:40-11:00**

Automatic recognition of spontaneous emotion in conversational speech is an important yet challenging problem. In this paper, we propose a deep neural network model to track continuous emotion changes in the arousal-valence two-dimensional space by combining inputs from raw waveform signals and spectrograms, both of which have been shown to be useful in the emotion recognition task. The neural network architecture contains a set of convolutional neural network (CNN) layers and bidirectional long short-term memory (BLSTM) layers to account for both temporal and spectral variation and model contextual content. Experimental results of predicting valence and arousal on the SEMAINE database and the RECOLA database show that the proposed model significantly outperforms model using hand-engineered features, by exploiting waveforms and spectrograms as input. We also compare the effects of waveforms vs. spectrograms and find that waveforms are better at capturing arousal, while spectrograms are better at capturing valence. Moreover, combining information from both inputs provides further improvement to the performance.

## Emotion Identification from Raw Speech Signals Using DNNs

*Mousmita Sarma[4], Pegah Ghahremani[1], Daniel Povey[1,2], Nagendra Kumar Goel[3], Kandarpa Kumar Sarma[4] and Najim Dehak[1]*

[1]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA

[2]Human Language Technology Center of Excellence, Johns Hopkins University, Baltimore, MD, USA

[3]Go-Vivace Inc., McLean, VA, USA

[4]Department of Electronics and Communication Engineering, Gauhati University, Guwahati, Assam, India

**Thu-O-1-3-4, Time: 11:00-11:20**

We investigate a number of Deep Neural Network (DNN) architectures for emotion identification with the IEMOCAP database. First we compare different feature extraction front-ends: we compare high-dimensional MFCC input (equivalent to filterbanks), versus frequency-domain and time-domain approaches to learning filters as part of the network. We obtain the best results with the time-domain filter-learning approach. Next we investigated different ways to aggregate information over the duration of an utterance. We tried approaches with a single label per utterance with time aggregation inside the network; and approaches where the label is repeated for each frame. Having a separate label per frame seemed to work best and the best architecture that we tried interleaves TDNN-LSTM withtime-restricted self-attention, achieving a weighted accuracy of 70.6%, versus 61.8% for the best previously published system which used 257-dimensional Fourier log-energies as input.

## Encoding Individual Acoustic Features Using Dyad-Augmented Deep Variational Representations for Dialog-level Emotion Recognition

*Jeng-Lin Li[1,2] and Chi-Chun Lee[1,2]*

[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan

[2]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan

**Thu-O-1-3-5, Time: 11:20-11:40**

Face-to-face dyadic spoken dialog is a fundamental unit of human interaction. Despite numerous empirical evidences in demonstrating interlocutor's behavior dependency in dyadic interactions, few technical works exist in leveraging the unique pattern of dynamics in task of advancing emotion recognition during face-to-face settings. In this work, we propose a framework of encoding an individual's acoustic features with dyad-augmented deep networks. The dyad-augmented deep networks includes a general variational deep Gaussian Mixture embedding network and a dyad-specific fine-tuned network. Our framework utilizes the augmented dyad-specific feature space to incorporate the unique behavior pattern emerged when two people interact. We perform dialog-level emotion regression tasks in both the CreativeIT and the NNIME databases. We obtain affect regression accuracy of 0.544 and 0.387 for activation and valence in the CreativeIT database (a relative improvement of 4.41% and 4.03% compared to using features without augmenting the dyad-specific representation) and we obtain 0.700 and 0.604 (4.48% and 4.14% relative improvement) for regressing activation and valence in the NNIME database.

## Variational Autoencoders for Learning Latent Representations of Speech Emotion: A Preliminary Study

*Siddique Latif[1], Rajib Rana[2], Junaid Qadir[1] and Julien Epps[3]*

[1]Information Technology University (ITU)-Punjab, Pakistan

[2]University of Southern Queensland, Australia

[3]University of New South Wales, Sydney, Australia

**Thu-O-1-3-6, Time: 11:40-12:00**

Learning the latent representation of data in unsupervised fashion is a very interesting process that provides relevant features for enhancing the performance of a classifier. For speech emotion recognition tasks, generating effective features is crucial. Currently, handcrafted features are mostly used for speech emotion recognition, however, features learned automatically using deep learning have shown strong success in many problems, especially in image processing. In particular, deep generative models such as Variational Autoencoders (VAEs) have gained enormous success in generating features for natural images. Inspired by this, we propose VAEs for deriving the latent representation of speech signals and use this representation to classify emotions. To the best of our knowledge, we are the first to propose VAEs for speech emotion classification. Evaluations on the IEMOCAP dataset demonstrate that features learned by VAEs can produce state-of-the-art results for speech emotion classification.

# Thu-O-1-4: Articulatory Information, Modeling and Inversion

MR G.01-G.02; 10:00-12:00; Thursday, 6 September, 2018

Chairs: Chandra Sekhar Seelamantula and P.C. Pandey

## Phoneme-to-Articulatory Mapping Using Bidirectional Gated RNN

*Théo Biasutto-Lervat and Slim Ouni*

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

**Thu-O-1-4-1, Time: 10:00-10:20**

Deriving articulatory dynamics from the acoustic speech signal has been addressed in several speech production studies. In this paper, we investigate whether it is possible to predict articulatory dynamics from phonetic information without having the acoustic speech signal. The input data may be considered as not sufficiently rich acoustically, as probably there is no explicit coarticulation information but we expect that the phonetic sequence provides compact yet rich knowledge.

Motivated by the recent success of deep learning techniques used in the acoustic-to-articulatory inversion, we have experimented around the bidirectional gated recurrent neural network architectures. We trained these models with an EMA corpus and have obtained good performances similar to the state-of-the-art articulatory inversion from LSF features, but using only the phoneme labels and durations.

## Tongue Segmentation with Geometrically Constrained Snake Model

*Zhihua Su[1], Jianguo Wei[1], Qiang Fang[2], Jianrong Wang[3] and Kiyoshi Honda[3]*

[1]School of Computer Software, Tianjin University, Tianjin, China

**Notes**

[2]Institute of Linguistics, Chinese Academy of Social Sciences, Beijing, China
[3]School of Computer Science, Tianjin University, Tianjin, China

**Thu-O-1-4-2, Time: 10:20-10:40**

Articulatory visualization aims at providing precise visual information of the speech organs (tongue, lips and velum) that accompany with speech signals. It is often critical in fundamental studies and certain applications. To construct an articulatory visualization system, the profile of the speech organs must be segmented from images acquired by various types of medical equipments. In this paper, a geometrically constrained snake model is proposed to segment tongue profiles from mid-sagittal MRI to deal with the situation in which the tongue contacts with the surrounding structures and the target object with inhomogeneity nature. The result indicates that the proposed method improves segmentation performance significantly compared with the traditional snake model.

### Low Resource Acoustic-to-articulatory Inversion Using Bi-directional Long Short Term Memory

*Aravind Illa and Prasanta Kumar Ghosh*
Electrical Engineering, Indian Institute of Science (IISc), Bangalore-560012, India

**Thu-O-1-4-3, Time: 10:40-11:00**

Estimating articulatory movements from speech acoustic features is known as acoustic-to-articulatory inversion (AAI). Large amount of parallel data from speech and articulatory motion is required for training an AAI model in a subject dependent manner, referred to as subject dependent AAI (SD-AAI). Electromagnetic articulograph (EMA) is a promising technology to record such parallel data, but it is expensive, time consuming and tiring for a subject. In order to reduce the demand for parallel acoustic-articulatory data in the AAI task for a subject, we, in this work, propose a subject-adaptative AAI method (SA-AAI) from an existing AAI model which is trained using large amount of parallel data from a fixed set of subjects. Experiments are performed with 30 subjects' acoustic-articulatory data and AAI is trained using BLSTM network to examine the amount of data needed from a new target subject for the SA-AAI to achieve an AAI performance equivalent to that of SD-AAI. Experimental results reveal that the proposed SA-AAI performs similar to that of the SD-AAI with ∼62.5% less training data. Among different articulators, the SA-AAI performance for tongue articulators matches with the corresponding SD-AAI performance with only ∼12.5% of the data used for SD-AAI training.

### Automatic Visual Augmentation for Concatenation Based Synthesized Articulatory Videos from Real-time MRI Data for Spoken Language Training

*Chandana S, Chiranjeevi Yarra, Ritu Aggarwal, Sanjeev Kumar Mittal, Kausthubha N K, Raseena K T, Astha Singh and Prasanta Kumar Ghosh*
Electrical Engineering, Indian Institute of Science (IISc), Bangalore-560012, India

**Thu-O-1-4-4, Time: 11:00-11:20**

For the benefit of spoken language training, concatenation based articulatory video synthesis has been proposed in the past to overcome the limitation in the articulatory data recording. For this, real time magnetic resonance imaging (rt-MRI) video image-frames (IFs) containing articulatory movements have been used. These IFs require a visual augmentation for better understanding. We, in this work, propose an augmentation method using pixel intensities in the regions enclosed by the articulatory boundaries obtained from air-tissue boundaries (ATBs). Since, the pixel intensities reflect the muscle movements in the articulators, the augmented IFs could provide realistic articulatory movements, when we color them accordingly. However, the ATB manual annotation is time consuming; hence, we propose to synthesize ATBs using the ATBs from a few selected frames that have been used in synthesizing the articulatory videos. We augment a set of synthesized articulatory videos for 50 words obtained from the MRI-TIMIT database. Subjective evaluation on the quality of the augmented videos using twenty-one subjects suggests that the videos are visually more appealing than the respective synthesized rt-MRI videos with a rating of 3.75 out of 5, where a score of 5 (1) indicates that the augmented video quality is excellent (poor).

**Notes**

### Air-Tissue Boundary Segmentation in Real-Time Magnetic Resonance Imaging Video Using Semantic Segmentation with Fully Convolutional Networks

*Valliappan CA, Renuka Mannem and Prasanta Kumar Ghosh*
Electrical Engineering, Indian Institute of Science, Bangalore-560012, India

**Thu-O-1-4-5, Time: 11:20-11:40**

In this paper, we propose a new technique for the segmentation of the Air-Tissue Boundaries (ATBs) in the vocal tract from the real-time magnetic resonance imaging (rtMRI) videos of the upper airway in the midsagittal plane. The proposed technique uses the approach of semantic segmentation using the Deep learning architecture called Fully Convolutional Networks (FCN). The architecture takes an input image and produces images of the same size with air and tissue class labels at each pixel. These output images are post processed using morphological filling and image smoothing to predict realistic ATBs. The performance of the predicted contours is evaluated using Dynamic Time Warping (DTW) distance between the manually annotated ground truth contours and the predicted contours. Four fold experiments with four subjects from USC-TIMIT corpus (with ∼2900 training images in every fold) demonstrate that the proposed FCN based approach has 8.87% and 9.65% lesser average error than the baseline Maeda Grid based scheme, for the lower and upper ATBs respectively. In addition, the proposed FCN based rtMRI segmentation achieves an average pixel classification accuracy of 99.05% across all subjects.

### Noise Robust Acoustic to Articulatory Speech Inversion

*Nadee Seneviratne, Ganesh Sivaraman, Vikramjit Mitra and Carol Espy-Wilson*
University of Maryland College Park

**Thu-O-1-4-6, Time: 11:40-12:00**

In previous work, we have shown that using articulatory features derived from a speech inversion system trained using synthetic data can significantly improve the robustness of an automatic speech recognition (ASR) system. This paper presents results from the first of two steps needed for exploring if the same will hold true for a speech inversion system trained with natural speech. Specifically, we developed a noise robust multi-speaker acoustic to articulatory speech inversion system. A feed forward neural network was trained using contextualized mel-frequency cepstral coefficients (MFCC) as the input acoustic features and six tract-variable (TV) trajectories as the output articulatory features. Experiments were performed on the U. Wisc. X-ray Microbeam (XRMB) database with 8 noise types artificially added at 5 different SNRs. Performance of the system was measured by computing the correlation between estimated and actual TVs. The performance of the multi-condition trained system was compared to the clean-speech trained system. The effect of speech enhancement on TV estimation was also evaluated. Experiments showed a 10% relative improvement in correlation over the baseline clean-speech trained system.

## Thu-SS-1-1: Special Session: Novel Paradigms for Direct Synthesis Based on Speech-Related Biosignals

MR G.03-G.04; 10:00-12:00; Thursday, 6 September, 2018
Chairs: Lorenz Diener, José Gonzale and Tanja Schultz

### Introduction

**Thu-SS-1-1-1, Time: 10:00-10:05**

## Designing a Pneumatic Bionic Voice Prosthesis - A Statistical Approach for Source Excitation Generation

*Farzaneh Ahmadi[1] and Tomoki Toda[2]*

[1]MARCS Institute, Western Sydney University

[2]Information Technology Center, Nagoya University

**Thu-SS-1-1-2, Time: 10:05-10:20**

This study follows up on our pioneering work in designing a Pneumatic Bionic Voice (PBV) prosthesis for larynx amputees. PBV prostheses are electronic adaptations of the traditional Pneumatic Artificial Larynx (PAL) device. The PAL is a non-invasive mechanical voice source, driven exclusively by respiration and with an exceptionally high voice quality. Following the PAL design closely, the PBV prosthesis is anticipated to substitute the medical gold standard of voice prostheses by generating a similar voice quality while remaining non-invasive and non-surgical. This paper describes a statistical approach to estimate the excitation waveform of the PBV source using the PAL as a reference. A Gaussian mixture model of the joint probability density of respiration and PAL voice features is implemented to estimate the excitation waveform of the PBV. The evaluation on a database of more than two hours of continuous speech shows a close match between f0 pattern and mel-cepstra of the estimated PBV source and the PAL. When used to re-synthesize the original speech, the intelligibility of the PBV speech remains high and is scored 7.1±0.4 compared to 7.9±0.15 of the original PAL source.

## A Neural Model to Predict Parameters for a Generalized Command Response Model of Intonation

*Bastian Schnell [1,2] and Philip N. Garner [1]*

[1]Idiap Research Institute, Martigny, Switzerland

[2]Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

**Thu-SS-1-1-3, Time: 10:20-10:35**

The Generalised Command Response (GCR) model is a time-local model of intonation that has been shown to lend itself to (cross-language) transfer of emphasis. In order to generalise the model to longer prosodic sequences, we show that it can be driven by a recurrent neural network emulating a spiking neural network. We show that a loss function for error backpropagation can be formulated analogously to that of the Spike Pattern Association Neuron (SPAN) method for spiking networks. The resulting system is able to generate prosody comparable to a state-of-the-art deep neural network implementation, but potentially retaining the transfer capabilities of the GCR model.

## Articulation-to-Speech Synthesis Using Articulatory Flesh Point Sensors' Orientation Information

*Beiming Cao[1], Myungjong Kim[1], Jun R. Wang[1], Jan van Santen[3], Ted Mau[4] and Jun Wang[1,2]*

[1]Speech Disorders & Technology Lab, Department of Bioengineering

[2]Callier Center for Communication Disorders, University of Texas at Dallas, United States

[3]Center for Spoken Language Understanding, Oregon Health & Science University, United States

[4]Department of Otolaryngology - Head and Neck Surgery University of Texas Southwestern Medical Center, United States

**Thu-SS-1-1-4, Time: 10:35-10:50**

Articulation-to-speech (ATS) synthesis generates audio waveform directly from articulatory information. Current works in ATS used articulatory movement information (spatial coordinates) only. The orientation information of articulatory flesh points has rarely been used, although some devices (e.g., electromagnetic articulography) provide that. Previous work indicated that orientation information contains significant information for speech production. In this paper, we explored the performance of applying orientation information of flesh points on articulators (i.e., tongue, lips and jaw) in ATS. Experiments using articulators' movement information with or without orientation information were conducted using standard deep neural networks (DNNs) and long-short term memory-recurrent neural networks (LSTM-RNNs). Both objective and subjective evaluations indicated that adding orientation information of flesh points on articulators in addition to movement information generated higher quality speech output than using movement information only.

## Effectiveness of Generative Adversarial Network for Non-Audible Murmur-to-Whisper Speech Conversion

*Neil Shah, Nirmesh Shah and Hemant Patil*

Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, India

**Thu-SS-1-1-5, Time: 10:50-11:05**

The murmur produced by the speaker and captured by the NonAudible Murmur (NAM)-one of the Silent Speech Interface (SSI) technique, suffers from the speech quality degradation. This is due to the lack of radiation effect at the lips and lowpass nature of the soft tissue, which attenuates the high frequency related information. In this work, a novel method for NAM-toWhisper (NAM2WHSP) speech conversion incorporating Generative Adversarial Network (GAN) is proposed. The GAN minimizes the distributional divergence between the whispered speech and the generated speech parameters (through adversarial optimization). The objective and subjective evaluation performed on the proposed system, justifies the ability of adversarial optimization over Maximum Likelihood (ML)-based optimization networks, such as a Deep Neural Network (DNN), in preserving and improving the speech quality and intelligibility. The adversarial optimization learns the mapping function with 54.2% relative improvement in MOS and 29.83% absolute reduction in % WER w.r.t. the state-of-the-art mapping techniques. Furthermore, we evaluated the proposed framework by analyzing the level of contextual information and the number of training utterances required for optimizing the network parameters, for the given task and database.

## Investigating Objective Intelligibility in Real-Time EMG-to-Speech Conversion

*Lorenz Diener and Tanja Schultz*

Cognitive Systems Lab, University of Bremen, Germany

**Thu-SS-1-1-6, Time: 11:05-11:20**

This paper presents an analysis of the influence of various system parameters on the output quality of our neural network based real-time EMG-to-Speech conversion system. This EMG-to-Speech system allows for the direct conversion of facial surface electromyographic signals into audible speech in real time, allowing for a closed-loop setup where users get direct audio feedback. Such a setup opens new avenues for research and applications through co-adaptation approaches.

In this paper, we evaluate the influence of several parameters on the output quality, such as time context, EMG-Audio delay, network-, training data- and Mel spectrogram size. The resulting output quality is evaluated based on the objective output quality measure STOI.

## Domain-Adversarial Training for Session Independent EMG-based Speech Recognition

*Michael Wand[1], Tanja Schultz[2] and Jürgen Schmidhuber[1]*

[1]Istituto Dalle Molle di studi sull'Intelligenza Artificiale (IDSIA), USI & SUPSI, Manno-Lugano, Switzerland

[2]University of Bremen, Bremen, Germany

**Thu-SS-1-1-7, Time: 11:20-11:35**

We present our research on continuous speech recognition based on Surface Electromyography (EMG), where speech information is captured by electrodes attached to the speaker's face. This method allows speech processing without requiring that an acoustic signal is present; however, reattachment of the EMG electrodes causes subtle changes in the recorded signal, which degrades the recognition accuracy and thus poses a major challenge for practical application of the system. Based on the growing body of recent work in domain-adversarial training of neural networks, we present a system which adapts the neural network frontend of our recognizer to data from a new recording session, without requiring supervised enrollment.

**Notes**

### Multi-Task Learning of Speech Recognition and Speech Synthesis Parameters for Ultrasound-based Silent Speech Interfaces

*László Tóth[1], Gábor Gosztolya[2], Tamás Grósz[1,2], Alexandra Markó[3,4] and Tamás Gábor Csapó[4,5]*

[1]Institute of Informatics, University of Szeged, Hungary

[2]MTA-SZTE Research Group on Artificial Intelligence, Szeged, Hungary

[3]Department of Phonetics, Eötvös Loránd University, Budapest, Hungary

[4]MTA-ELTE Lendület Lingual Articulation Research Group, Budapest, Hungary

[5]Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Budapest, Hungary

**Thu-SS-1-1-8, Time: 11:35-11:50**

Silent Speech Interface systems apply two different strategies to solve the articulatory-to-acoustic conversion task. The recognition-and-synthesis approach applies speech recognition techniques to map the articulatory data to a textual transcript, which is then converted to speech by a conventional text-to-speech system. The direct synthesis approach seeks to convert the articulatory information directly to speech synthesis (vocoder) parameters. In both cases, deep neural networks are an evident and popular choice to learn the mapping task. Recognizing that the learning of speech recognition and speech synthesis targets (acoustic model states vs. vocoder parameters) are two closely related tasks over the same ultrasound tongue image input, here we experiment with the multi-task training of deep neural networks, which seeks to solve the two tasks simultaneously.

Our results show that the parallel learning of the two types of targets is indeed beneficial for both tasks. Moreover, we obtained further improvements by using multi-task training as a weight initialization step before task-specific training. Overall, we report a relative error rate reduction of about 7% in both the speech recognition and the speech synthesis tasks.

## Discussion and Closing

**Thu-SS-1-1-9, Time: 11:50-12:00**

.....

# Thu-SS-1-2: Special Session: Low Resource Speech Recognition Challenge for Indian Languages

MR 1.01-1.02; 10:00-12:00; Thursday, 6 September, 2018

Chairs: Kalika Bali, Krishna Doss Mohan, Rupesh Kumar Mehta, Niranjan Nayak, Sunayana Sitaram and Radhakrishnan Srikanth

.....

## Introduction

**Thu-SS-1-2-1, Time: 10:00-10:15**

### Transcription Correction for Indian Languages Using Acoustic Signatures

*Jeena JPrakash[1], Golda Brunet Rajan[2] and Hema Murthy[1]*

[1]Indian Institute of Technology, Madras, India

[2]Government College of Engineering, Salem, India

**Thu-SS-1-2-2, Time: 10:15-10:30**

Accurate phonetic transcription of the speech corpus has a significant impact on the performance of speech processing applications especially for low resource languages. Mismatches between the transcriptions and their utterances occur often at phoneme level due to insertion/deletion/substitution errors. This is very common in Indian languages owing to schwa deletion in the context of vowels and agglutination in the context of consonants.

An attempt is made in this paper to use acoustic cues at the syllable level to remove vowels from the transcription when they are poorly articulated or absent. Hidden Markov model (HMM) based forced Viterbi alignment (FVA) and group delay (GD) based signal processing are employed in tandem to achieve this task. Disagreement between FVA (which produces vowel boundaries based on transcription) and GD boundaries (which uses signal processing cues for syllables) are used to correct the transcription. An increase in likelihood of 0.3% is observed across 3 Indian languages, namely, Gujarati, Telugu and Tamil.

### BUT System for Low Resource Indian Language ASR

*Bhargav Pulugundla[1,2], Murali Karthick Baskar[1], Santosh Kesiraju[1,3], Ekaterina Egorova[1], Martin Karafiát[1], Lukáš Burget[1] and Jan Černocký[1]*

[1]Brno University of Technology, Speech@FIT and IT4I Center of Excellence, Czechia

[2]Phonexia s.r.o., Czechia

[3]IIIT Hyderabad, India

**Thu-SS-1-2-3, Time: 10:30-10:45**

This paper describes the BUT 'Jilebi' team's speech recognition systems created for the 2018 low resource speech recognition challenge for Indian languages. We investigate modifications of multilingual time-delay neural network (TDNN) architectures with transfer learning and compare them to bi-directional residual memory networks (BRMN) and bi-directional LSTM. Our best submission based on system combination achieved word error rates of 13.92% (Tamil), 14.71% (Telugu) and 14.06% (Gujarati). We present the details of submitted systems and also the post-evaluation analysis done for lexicon discovery using unsupervised word segmentation.

### DA-IICT/IIITV System for Low Resource Speech Recognition Challenge 2018

*Hardik B. Sailor[1], Maddala Venkata Siva Krishna[2], Diksha Chhabra[2], Ankur T. Patil[1], Madhu Kamble[1] and Hemant Patil[1]*

[1]Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar-382007, Gujarat, India

[2]Indian Institute of Information Technology (IIIT), Vadodara, Gujarat, India

**Thu-SS-1-2-4, Time: 10:45-11:00**

This paper presents an Automatic Speech Recognition (ASR) system, in the Gujarati language, developed for Low Resource Speech Recognition Challenge for Indian Languages in INTERSPEECH 2018. For front-end, Amplitude Modulation (AM) features are extracted using the standard and data-driven auditory filterbanks. Recurrent Neural Network Language Models (RNNLM) are used for this task. There is a relative improvement of 36.18% and 40.95% in perplexity on the test and blind test sets, respectively, compared to 3-gram LM. TimeDelay Neural Network (TDNN) and TDNN-Long Short-Term Memory (LSTM) models are employed for acoustic modeling.

The statistical significance of proposed approaches is justified using a bootstrap-based % Probability of Improvement (POI) measure. RNNLM rescoring with 3-gram LM gave an absolute reduction of 0.69-1.29% in Word Error Rate (WER) for various feature sets. AM features extracted using the gammatone filterbank (AM-GTFB) performed well on the blind test set compared to the FBANK baseline (POI>70%). The combination of ASR systems further increased the performance with an absolute reduction of 1.89 and 2.24% in WER for test and blind test sets, respectively (100% POI).

**Notes**

### An Exploration towards Joint Acoustic Modeling for Indian Languages: IIIT-H Submission for Low Resource Speech Recognition Challenge for Indian Languages, INTERSPEECH 2018

*Hari Krishna, Krishna Gurugubelli, Vishnu Vidyadhara Raju V and Anil Kumar Vuppala*

Speech Processing Laboratory, KCIS

International Institute of Information Technology, Hyderabad, India

**Thu-SS-1-2-5, Time: 11:00-11:15**

India being a multilingual society, a multilingual automatic speech recognition system (ASR) is widely appreciated. Despite different orthographies, Indian languages share same phonetic space. To exploit this property, a joint acoustic model has been trained for developing multilingual ASR system using a common phone-set. Three Indian languages namely Telugu, Tamil and, Gujarati are considered for the study. This work studies the amenability of two different acoustic modeling approaches for training a joint acoustic model using common phone-set. Sub-space Gaussian mixture models (SGMM) and recurrent neural networks (RNN) trained with connectionist temporal classification (CTC) objective function are explored for training joint acoustic models. From the experimental results, it can be observed that the joint acoustic models trained with RNN-CTC have performed better than SGMM system even on 120 hours of data (approx 40 hrs per language). The joint acoustic model trained with RNN-CTC has performed better than monolingual models, due to an efficient data sharing across the languages. Conditioning the joint model with language identity had a minimal advantage. Sub-sampling the features by a factor of 2 while training RNN-CTC models has reduced the training times and has performed better.

### TDNN-based Multilingual Speech Recognition System for Low Resource Indian Languages

*Noor Fathima, Tanvina Patel, Mahima C and Anuroop Iyengar*

Cogknit Semantics, Bangalore, Karnataka, India

**Thu-SS-1-2-6, Time: 11:15-11:30**

India is a diverse and multilingual country. It has vast linguistic variations, spoken across its billion plus population. Lack of resources in terms of transcribed speech data, phonetic pronunciation dictionary or lexicon and text collection has hindered the development and improvement of the ASR systems for Indic languages. With the Interspeech 2018 Special Session: Low Resource Speech Recognition Challenge for Indian Languages, efforts have been made to solve this issue to an extent. In this paper, we explore the fact that the shared phonetic properties of the languages are essential for improved ASR performance. We build a multilingual Time Delay Neural Network (TDNN) system that uses combined acoustic modeling and language-specific information to decode the input test sequences. Using this approach, for Tamil, Telugu and Gujarati language we obtain a Word Error Rate (WER) of 16.07%, 17.14%, 17.69%, respectively, which was the second best system at the challenge.

### Articulatory and Stacked Bottleneck Features for Low Resource Speech Recognition

*Vishwas M. Shetty, Rini A Sharon, Basil Abraham, Tejaswi Seeram, Anusha Prakash, Nithya Ravi and S. Umesh*

Indian Institute of Technology-Madras, India

**Thu-SS-1-2-7, Time: 11:30-11:45**

In this paper, we discuss the benefits of using articulatory and stacked bottleneck features (SBF) for low resource speech recognition. Articulatory features (AF) which capture the underlying attributes of speech production are found to be robust to channel and speaker variations. However, building an efficient articulatory classifier to extract AF requires an enormous amount of data. In low resource acoustic modeling, we propose to train the bidirectional long short-term memory (BLSTM) articulatory classifier by pooling data from the available low resource Indian languages, namely, Gujarati, Tamil and Telugu. This is done in the context of Microsoft Indian Language challenge. Similarly, we train a multilingual bottleneck feature extractor and an SBF extractor using the pooled data. To bias, the SBF network towards the target language, a second network in the stacked architecture was trained using the target language alone. The performance of ASR system trained with stand-alone AF is observed to be at par with the multilingual bottleneck features. When the AF and the biased SBF are appended, they are found to outperform the conventional filterbank features in the multilingual deep neural network (DNN) framework and the high-resolution Mel frequency cepstral coefficient (MFCC) features in the time-delayed neural network(TDNN) framework.

### ISI ASR System for the Low Resource Speech Recognition Challenge for Indian Languages

*Jayadev Billa*

Information Sciences Institute, University of Southern California, Marina del Rey, CA 90292, USA

**Thu-SS-1-2-8, Time: 11:45-12:00**

This paper describes the ISI ASR system used to generate ISI's submissions across Gujarati, Tamil and Telugu speech recognition tasks as part of the Low Resource Speech Recognition Challenge for Indian Languages. The key constraints on this task were limited training data and the restriction that no external data be used. The ISI ASR system leverages our earlier work on data augmentation and dropout approaches and current work on multilingual training within a Eesen based end-to-end Long Short Term Memory (LSTM) based automatic speech recognition (ASR) system trained with the Connectionist Temporal Classification (CTC) loss criterion and demonstrates, to the best of our knowledge, one of the first times such systems have been applied to low resource languages with performance comparable and some cases better than hybrid DNN systems.

Our best monolingual systems show between 6.5% to 25.5% relative reduction in word error rate (WER) compared to the challenge organizer's Time Delay Neural Network (TDNN) based baseline WERs. We further extend these systems with multilingual training approaches that lead to an additional 4.5% to 11.1% relative reduction in WER as measured on the development set.

## Thu-S&T-1-1: Show and Tell 7

MR G.05-G.06; 10:00-12:00; Thursday, 6 September, 2018

Chairs: Mahesh M and Jitendra Vepa

### An Automated Assistant for Medical Scribes

*Gregory Finley, Erik Edwards, Amanda Robinson, Najmeh Sadoughi, James Fone, Mark Miller, David Suendermann-Oeft, Michael Brenndoerfer[1] and Nico Axtmann[2]*

EMR.AI Inc., San Francisco, CA, USA

[1]University of California, Berkeley, USA

[2]DHBW, Karlsruhe, Germany

**Thu-S&T-1-1-1, Time: 10:00-12:00**

Healthcare professionals spend a significant amount of their time on administrative tasks rather than direct patient care. One effective way to alleviate some of this burden is to employ a medical scribe, who charts patient-physician encounters in real time. We present a complete implementation of an automated medical scribe assistant, which listens to the encounter and produces a draft of report text and the information used to create it, which can dramatically streamline the scribe's task. This system is, to our knowledge, the first automated scribe ever presented and relies on multiple speech and language technologies, including speaker diarization, medical speech recognition, knowledge extraction and natural language generation.

**Notes**

## AGROASSAM: A Web Based Assamese Speech Recognition Application for Retrieving Agricultural Commodity Price and Weather Information

*Abhishek Dey[1], Abhash Deka[2], Siddika Imani[1], Barsha Deka[1], Rohit Sinha[2], S R Mahadeva Prasanna[2,3], Priyankoo Sarmah[2], K Samudravijaya[2] and Nirmala S.R.[1]*

[1]GUIST, Gauhati University, Guwahati-781014, India

[2]Indian Institute of Technology Guwahati, Guwahati-781039, India

[3]Indian Institute of Technology Dharwad, Dharwad-580011, India

**Thu-S&T-1-1-2, Time: 10:00-12:00**

This paper presents a speech-based web application developed for retrieving the price of agricultural commodities and weather related information in Assamese language. The price of agricultural commodities are retrieved from AGMARKNET website while the weather related informations are extracted from the IMD website. Both these websites are updated on a daily basis by the Government of India. The back-end of the application consists of automatic speech recognition (ASR) modules developed using state-of-the-art acoustic modeling approaches. Word error rates (WERs) of 7.79% and 4.98% are achieved for commodity and district names respectively.

## Voice-powered Solutions with Cloud AI

*Dan Aharon*

Google, Inc.

**Thu-S&T-1-1-3, Time: 10:00-12:00**

Google has been a leader in research into speech recognition, speech synthesis and natural language and much of that innovation has powered Google products like Search, Google Assistant, Google Maps android and others.

Google Cloud AI aims to help enterprises, startups, students and any other developer use AI to build great end user experiences and benefiting from investments Google has made in this space.
Google Cloud Speech-to-Text (formerly known as Cloud Speech API) was first released in Beta in 2016, was followed by Dialogflow Enterprise Edition in 2017 and Cloud Text-to-Speech in 2018. Each of these products is backed by cutting-edge research that Google has conducted in these spaces.

This set of products is expected to be generally available before this conference starts and we're excited to see how participants would use them to build new speech-powered experiences including mobile apps, connected devices (cars, TVs, speakers), robots, etc.
We will showcase multiple demos at the show, that show how to build your own personal shopper or personal assistant using Cloud AI technology.

We will also show how easy it is to connect this to a phone line and turn this into a conversational IVR that feels like a personal agent.

## Speech Synthesis in the Wild

*Ganesh Sivaraman, Parav Nagarsheth and Elie Khoury*

Pindrop, Atlanta USA

**Thu-S&T-1-1-4, Time: 10:00-12:00**

Speech synthesis has wide range of applications in modern artificial intelligence technologies. Most state-of-the-art speech synthesis systems usually require high quality recordings of large amounts of speech data of the target speaker. We focus on low-budget speech synthesis. Our software deals with methods to perform statistical parametric speech synthesis using unlabeled and mixed quality speech data sourced from the internet. An average voice model trained using DNN is adapted to a target speaker using different speaker adaptation strategies. Preprocessing methods like speech enhancement, diarization and segmentation are applied to the sourced data. Utterance selection based on Mean cepstral distortion and forced alignment confidence are applied to prune the noisy and mis-aligned data. The mixed quality data thus pre-processed is then used to adapt the average voice model and duration models to the target speaker.

The software to be demonstrated automates the whole procedure from preprocessing to synthesis. The software will be demonstrated by performing live synthesis using audio sourced from Youtube.

# Thu-P-1-1: Deep Enhancement

Hall 4-6: Poster-1; 10:00-12:00; Thursday, 6 September, 2018

Chair: Shinji Watanabe

## Deep Noise Tracking Network: A Hybrid Signal Processing/Deep Learning Approach to Speech Enhancement

*Shuai Nie[1,3], Shan Liang[1], Bin Liu[1,3], Yaping Zhang[1,3], Wenju Liu[1] and Jianhua Tao1[,2,3]*

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences

[2]CAS Center for Excellence in Brain Science and Intelligence Technology

[3]School of Artificial Intelligence, University of Chinese Academy of Sciences

**Thu-P-1-1-1, Time: 10:00-12:00**

Noise statistics and speech spectrum characteristics are the essential information for the single channel speech enhancement. The signal processing-based methods mainly rely on noise statistics estimation. They perform very well for stationary noise, but have remained difficult to cope with non-stationary noise. While the deep learning-based methods mainly focus on the perception on the spectrum characteristics of speech and have a capacity in dealing with non-stationary noise. However, the performance would degrade dramatically for the unseen noise types, which could be due to the over-reliance on data and the ignorance to domain knowledge of signal process. Obviously, the hybrid signal processing/deep learning scheme may be a smart alternative. In this paper, we incorporate the powerful perceptual capabilities of deep learning in the conventional speech enhancement framework. Deep learning is used to estimate the speech presence probability and the update factor of noise statistics, which are then integrated into the Wiener filter-based speech enhancement structure to enhance the desired speech. All components are jointly optimized by a spectrum approximation objective. Systematic experiments on CHiME-4 and NOISEX-92 demonstrate the proposed hybrid signal processing/deep learning approach to noise suppression in noise-unmatched and noise-matched conditions.

## A Deep Neural Network Based Harmonic Noise Model for Speech Enhancement

*Zhiheng Ouyang[1], Hongjiang Yu[1], Wei-Ping Zhu[1] and Benoit Champagne[2]*

[1]Dept. of Electrical and Computer Engineering, Concordia University, Montreal, Canada

[2]Dept. of Electrical and Computer Engineering, McGill University, Montreal, Canada

**Thu-P-1-1-2, Time: 10:00-12:00**

In this paper, we present a novel deep neural network (DNN) based speech enhancement method that uses a harmonic noise model (HNM) to estimate the clean speech. By utilizing HNM to model the clean speech in the short-time Fourier transform domain and extracting some time-frequency features of noisy speech for the DNN training, the new method predicts the harmonic and residual amplitudes of clean speech from a set of noisy speech features. In order to emphasize the importance of the harmonic component and reduce the effect caused by the residual, a scaling factor is also introduced and applied to the residual amplitude. The enhanced speech is reconstructed with the estimated clean speech amplitude and the noisy phase of HNM. Experimental results demonstrate that our proposed HNM-

**Notes**

DNN method outperforms two existing DNN based speech enhancement methods in terms of both speech quality and intelligibility.

## A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement

*Ke Tan[1] and DeLiang Wang[1,2]*

[1]Department of Computer Science and Engineering, The Ohio State University, USA

[2]Center for Cognitive and Brain Sciences, The Ohio State University

**Thu-P-1-1-3, Time: 10:00-12:00**

Many real-world applications of speech enhancement, such as hearing aids and cochlear implants, desire real-time processing, with no or low latency. In this paper, we propose a novel convolutional recurrent network (CRN) to address real-time monaural speech enhancement. We incorporate a convolutional encoder-decoder (CED) and long short-term memory (LSTM) into the CRN architecture, which leads to a causal system that is naturally suitable for real-time processing. Moreover, the proposed model is noise- and speaker-independent, i.e. noise types and speakers can be different between training and test. Our experiments suggest that the CRN leads to consistently better objective intelligibility and perceptual quality than an existing LSTM based model. Moreover, the CRN has much fewer trainable parameters.

## All-Neural Multi-Channel Speech Enhancement

*Zhong-Qiu Wang[1] and DeLiang Wang[1,2]*

[1]Department of Computer Science and Engineering, The Ohio State University, USA

[2]Center for Cognitive and Brain Sciences, The Ohio State University, USA

**Thu-P-1-1-4, Time: 10:00-12:00**

This study proposes a novel all-neural approach for multi-channel speech enhancement, where robust speaker localization, acoustic beamforming, post-filtering and spatial filtering are all done using deep learning based time-frequency (T-F) masking. Our system first performs monaural speech enhancement on each microphone signal to obtain the estimated ideal ratio masks for beamforming and robust time delay of arrival (TDOA) estimation. Then with the estimated TDOA, directional features indicating whether each T-F unit is dominated by the signal coming from the estimated target direction are computed. Next, the directional features are combined with the spectral features extracted from the beamformed signal to achieve further enhancement. Experiments on a two-microphone setup in reverberant environments with strong diffuse babble noise demonstrate the effectiveness of the proposed approach for multi-channel speech enhancement.

## Deep Learning for Acoustic Echo Cancellation in Noisy and Double-Talk Scenarios

*Hao Zhang[1] and DeLiang Wang[1,2,3]*

[1]Department of Computer Science and Engineering, The Ohio State University, USA

[2]Center for Cognitive and Brain Sciences, The Ohio State University, USA

[3]Center of Intelligent Acoustics and Immersive Communications, Northwestern Polytechnical University, China

**Thu-P-1-1-5, Time: 10:00-12:00**

Traditional acoustic echo cancellation (AEC) works by identifying an acoustic impulse response using adaptive algorithms. We formulate AEC as a supervised speech separation problem, which separates the loudspeaker signal and the near-end signal so that only the latter is transmitted to the far end. A recurrent neural network with bidirectional long short-term memory (BLSTM) is trained to estimate the ideal ratio mask from features extracted from the mixtures of near-end and far-end signals. A BLSTM estimated mask is then applied to separate and suppress the far-end signal, hence removing the echo. Experimental results show the effectiveness of the proposed method for echo removal in double-talk, background noise and nonlinear distortion scenarios. In addition, the proposed method can be generalized to untrained speakers.

## The Conversation: Deep Audio-Visual Speech Enhancement

*Triantafyllos Afouras, Joon Son Chung and Andrew Zisserman*

Visual Geometry Group, Department of Engineering Science, University of Oxford, UK

**Thu-P-1-1-6, Time: 10:00-12:00**

Our goal is to isolate individual speakers from multi-talker simultaneous speech in videos. Existing works in this area have focussed on trying to separate utterances from known speakers in controlled environments. In this paper, we propose a deep audio-visual speech enhancement network that is able to separate a speaker's voice given lip regions in the corresponding video, by predicting both the magnitude and the phase of the target signal. The method is applicable to speakers unheard and unseen during training and for unconstrained environments. We demonstrate strong quantitative and qualitative results, isolating extremely challenging real-world examples.

## Student-Teacher Learning for BLSTM Mask-based Speech Enhancement

*Aswin Shanmugam Subramanian, Szu-Jui Chen and Shinji Watanabe*

Center for Language and Speech Processing, Johns Hopkins University

**Thu-P-1-1-7, Time: 10:00-12:00**

Spectral mask estimation using bidirectional long short-term memory (BLSTM) neural networks has been widely used in various speech enhancement applications and it has achieved great success when it is applied to multichannel enhancement techniques with a mask-based beamformer. However, when these masks are used for single channel speech enhancement they severely distort the speech signal and make them unsuitable for speech recognition. This paper proposes a student-teacher learning paradigm for single channel speech enhancement. The beamformed signal from multichannel enhancement is given as input to the teacher network to obtain soft masks. An additional cross-entropy loss term with the soft mask target is combined with the original loss, so that the student network with single-channel input is trained to mimic the soft mask obtained with multichannel input through beamforming. Experiments with the CHiME-4 challenge single channel track data shows improvement in ASR performance.

## Speech Enhancement Using Deep Mixture of Experts Based on Hard Expectation Maximization

*Pavan Karjol and Prasanta Kumar Ghosh*

Electrical Engineering, Indian Institute of Science, Bengaluru 560012, India

**Thu-P-1-1-8, Time: 10:00-12:00**

We consider the problem of deep mixture of experts based speech enhancement. The deep mixture of experts, where experts are considered as deep neural network (DNN), is difficult to train due to the network structure. In this work, we propose a pre-training method for individual DNN in deep mixture of experts. We use hard expectation maximization (EM) to pre-train the individual DNNs. After pre-training, we take a weighted combination of outputs of individual DNN experts and jointly train the whole system. We compare the proposed method with single DNN based speech enhancement scheme. Speech enhancement experiments, in four SNR conditions, show the superiority of proposed method over the baseline scheme. The average improvements obtained for four seen noise cases over single DNN scheme are 0.08, 0.59 dB and 0.015 in terms of objective measures viz perceptual evaluation of speech quality (PESQ), segmental signal to noise ratio (seg SNR) and short time objective intelligibility (STOI) respectively.

## Adversarial Feature-Mapping for Speech Enhancement

*Zhong Meng[1,2], Jinyu Li[1], Yifan Gong[1] and Biing-Hwang (Fred)*

**Notes**

*Juang*[2]

[1]Microsoft AI and Research, Redmond, WA, USA

[2]Georgia Institute of Technology, Atlanta, GA, USA

**Thu-P-1-1-9, Time: 10:00-12:00**

Feature-mapping with deep neural networks is commonly used for single-channel speech enhancement, in which a feature-mapping network directly transforms the noisy features to the corresponding enhanced ones and is trained to minimize the mean square errors between the enhanced and clean features. In this paper, we propose an adversarial feature-mapping (AFM) method for speech enhancement which advances the feature-mapping approach with adversarial learning. An additional discriminator network is introduced to distinguish the enhanced features from the real clean ones. The two networks are jointly optimized to minimize the feature-mapping loss and simultaneously mini-maximize the discrimination loss. The distribution of the enhanced features is further pushed towards that of the clean features through this adversarial multi-task training. To achieve better performance on ASR task, senone-aware (SA) AFM is further proposed in which an acoustic model network is jointly trained with the feature-mapping and discriminator networks to optimize the senone classification loss in addition to the AFM losses. Evaluated on the CHiME-3 dataset, the proposed AFM achieves 16.95% and 5.27% relative word error rate (WER) improvements over the real noisy data and the feature-mapping baseline respectively and the SA-AFM achieves 9.85% relative WER improvement over the multi-conditional acoustic model.

### Biophysically-inspired Features Improve the Generalizability of Neural Network-based Speech Enhancement Systems

*Deepak Baby and Sarah Verhulst*

Dept. of Information Technology, Ghent University, Belgium

**Thu-P-1-1-10, Time: 10:00-12:00**

Recent advances in neural network (NN)-based speech enhancement schemes are shown to outperform most conventional techniques. However, the performance of such systems in adverse listening conditions such as negative signal-to-noise ratios and unseen noises is still far from that of humans. Motivated by the remarkable performance of humans under these challenging conditions, this paper investigates whether biophysically-inspired features can mitigate the poor generalization capabilities of NN-based speech enhancement systems. We make use of features derived from several human auditory periphery models for training a speech enhancement system that employs long short-term memory (LSTM) and evaluate them on a variety of mismatched testing conditions. The results reveal that biophysically-inspired auditory models such as nonlinear transmission line models improve the generalizability of LSTM-based noise suppression systems in terms of various objective quality measures, suggesting that such features lead to robust speech representations that are less sensitive to the noise type.

### Error Modeling via Asymmetric Laplace Distribution for Deep Neural Network Based Single-Channel Speech Enhancement

*Li Chai*[1]*, Jun Du*[1] *and Chin-Hui Lee*[2]

[1]University of Science and Technology of China, Hefei, Anhui, P. R. China

[2]Georgia Institute of Technology, Atlanta, GA. USA

**Thu-P-1-1-11, Time: 10:00-12:00**

The minimum mean squared error (MMSE) as a conventional training criterion for deep neural network (DNN) based speech enhancement has been found many problems. In our recent work, a maximum likelihood (ML) approach to parameter learning by modeling the prediction error vector as a Gaussian density was proposed. In this study, our preliminary statistical analysis reveals the super-Gaussianity and asymmetricity of the prediction error distribution. Consequently, we adopt the asymmetric Laplace distribution (ALD) instead of the Gaussian distribution (GD) to model the prediction error vectors. Then the new derivation for optimizing the the proposed ML-ALD-DNN with both DNN and ALD parameters is presented. Moreover, we can well interpret the asymmetry parameter of ALD as the balance control between noise reduction and speech preservation from both formulations and experiments. This implies that the customization of DNN models for the different noise types and levels is possible by the setting of the asymmetry parameter. Finally, our ML-ALD-DNN approach achieves better STOI and SSNR measures over both MMSE-DNN and ML-GD-DNN approaches.

### A Priori SNR Estimation Based on a Recurrent Neural Network for Robust Speech Enhancement

*Yangyang Xia*[1] *and Richard Stern*[1,2]

[1]Department of Electrical and Computer Engineering, Carnegie Mellon University

[2]Language Technologies Institute, Carnegie Mellon University

**Thu-P-1-1-12, Time: 10:00-12:00**

Speech enhancement under highly non-stationary noise conditions remains a challenging problem. Classical methods typically attempt to identify a frequency-domain optimal gain function that suppresses noise in noisy speech. These algorithms typically produce artifacts such as "musical noise" that are detrimental to machine and human understanding, largely due to inaccurate estimation of noise power spectra. The optimal gain function is commonly referred to as the ideal ratio mask (IRM) in neural-network-based systems and the goal becomes estimation of the IRM from the short-time Fourier transform amplitude of degraded speech. While these data-driven techniques are able to enhance speech quality with reduced artifacts, they are frequently not robust to types of noise that they had not been exposed to in the training process. In this paper, we propose a novel recurrent neural network (RNN) that bridges the gap between classical and neural-network-based methods. By reformulating the classical decision-directed approach, the a priori and a posteriori SNRs become latent variables in the RNN, from which the frequency-dependent estimated likelihood of speech presence is used to update recursively the latent variables. The proposed method provides substantial enhancement of speech quality and objective accuracy in machine interpretation of speech.

## Thu-P-1-2: Acoustic Scenes and Rare Events

Hall 4-6: Poster-2; 10:00-12:00; Thursday, 6 September, 2018

Chair: Vinay Namboodiri

### Multiple Instance Deep Learning for Weakly Supervised Small-Footprint Audio Event Detection

*Shao-Yen Tseng*[2]*, Juncheng Li*[1,3]*, Yun Wang*[3]*, Florian Metze*[3]*, Joseph Szurley*[1] *and Samarjit Das*[1]

[1]Robert Bosch LLC, Research and Technology Center, USA

[2]University of Southern California, Department of Electrical Engineering, USA

[3]Carnegie Mellon University, Language Technology Institute, USA

**Thu-P-1-2-1, Time: 10:00-12:00**

State-of-the-art audio event detection (AED) systems rely on supervised learning using strongly labeled data. However, this dependence severely limits scalability to large-scale datasets where fine resolution annotations are too expensive to obtain. In this paper, we propose a small-footprint multiple instance learning (MIL) framework for multi-class AED using weakly annotated labels. The proposed MIL framework uses audio embeddings extracted from a pre-trained convolutional neural network as input features. We show that by using audio embeddings the MIL framework can be implemented using a simple DNN with performance comparable to recurrent neural networks.

We evaluate our approach by training an audio tagging system using a subset of AudioSet, which is a large collection of weakly labeled YouTube video excerpts. Combined with a late-fusion approach, we improve the

**Notes**

F1 score of a baseline audio tagging system by 17%. We show that audio embeddings extracted by the convolutional neural networks significantly boost the performance of all MIL models. This framework reduces the model complexity of the AED system and is suitable for applications where computational resources are limited.

## Unsupervised Temporal Feature Learning Based on Sparse Coding Embedded BoAW for Acoustic Event Recognition

*Liwen Zhang[1], Jiqing Han[1] and Shiwen Deng[2]*

[1]Harbin Institute of Technology, China

[2]Harbin Normal University, China

**Thu-P-1-2-2, Time: 10:00-12:00**

The performance of an Acoustic Event Recognition (AER) system highly depends on the statistical information and the temporal dynamics in the audio signals. Although the traditional Bag of Audio Words (BoAW) and the Gaussian Mixture Models (GMM) approaches can obtain more statistics information by aggregating multiple frame-level descriptors of an audio segment compared with the frame-level feature learning methods, its temporal information is unreserved. Recently, more and more Deep Neural Networks (DNN) based AER methods have been proposed to effectively capture the temporal information in audio signals and achieved better performance, however, these methods usually required the manually annotated labels and fixed-length input during feature learning process. In this paper, we proposed a novel unsupervised temporal feature learning method, which can effectively capture the temporal dynamics for an entire audio signal with arbitrary duration by building direct connections between the BoAW histograms sequence and its time indexes using a non-linear Support Vector Regression (SVR) model. Furthermore, to make the feature representation have a better signal reconstruction ability, we embedded the sparse coding approach in the conventional BoAW framework. Compared with the BoAW and Convolutional Neural Network (CNN) baselines, experimental results showed our method brings improvements of 9.7% and 4.1% respectively.

## Data Independent Sequence Augmentation Method for Acoustic Scene Classification

*Zhang Teng, Kailai Zhang and Ji Wu*

Multimedia Signal and Intelligent Information Processing Lab, Department of Electronic Engineering, Tsinghua University, Beijing, P.R.China

**Thu-P-1-2-3, Time: 10:00-12:00**

Augmenting datasets by transforming inputs in a way such as vocal tract length perturbation (VTLP) is a crucial ingredient of the state of the art methods for speech recognition tasks. In contrast to speech, sounds coming from realistic environments have no speaker to speaker variations. Thus VTLP is invalid for acoustic scene classification tasks. This paper investigates a novel sequence augmentation method for long short-term memory (LSTM) acoustic modeling to deal with data sparsity in acoustic scene classification tasks. The audio sequences are randomly rearranged and concatenated during training, but at test time, a prediction is made by the original audio sequence. The rearrangement is well-designed to adapt to the long short-term dependency in LSTM models. Experiments on acoustic scene classification task show performance improvements of the proposed methods. The classification errors in LITIS ROUEN dataset and DCASE2016 dataset are reduced by 18.1% and 6.4% relatively.

## A Compact and Discriminative Feature Based on Auditory Summary Statistics for Acoustic Scene Classification

*Hongwei Song[1], Jiqing Han[1] and Shiwen Deng[2]*

[1]Harbin Institute of Technology, China

[2]Harbin Normal University, China

**Thu-P-1-2-4, Time: 10:00-12:00**

One of the biggest challenges of acoustic scene classification (ASC) is to find proper features to better represent and characterize environmental sounds. Environmental sounds generally involve more sound sources while exhibiting less structure in temporal spectral representations. However, the background of an acoustic scene exhibits temporal homogeneity in acoustic properties, suggesting it could be characterized by distribution statistics rather than temporal details. In this work, we investigated using auditory summary statistics as the feature for ASC tasks. The inspiration comes from a recent neuroscience study, which shows the human auditory system tends to perceive sound textures through time-averaged statistics. Based on these statistics, we further proposed to use linear discriminant analysis to eliminate redundancies among these statistics while keeping the discriminative information, providing an extreme compact representation for acoustic scenes. Experimental results show the outstanding performance of the proposed feature over the conventional handcrafted features.

## ASe: Acoustic Scene Embedding Using Deep Archetypal Analysis and GMM

*Pulkit Sharma[1], Vinayak Abrol[2] and Anshul Thakur[1]*

[1]IIT Mandi, India

[2]Idiap Research Institute, Martigny, Switzerland

**Thu-P-1-2-5, Time: 10:00-12:00**

In this paper, we propose a deep learning framework which combines the generalizability of Gaussian mixture models (GMM) and discriminative power of deep matrix factorization to learn acoustic scene embedding (ASe) for the acoustic scene classification task. The proposed approach first builds a Gaussian mixture model-universal background model (GMM-UBM) using frame-wise spectral representations. This UBM is adapted to a waveform and the likelihood for each spectral frame representation is stored as a feature matrix. This matrix is fed to a deep matrix factorization pipeline (with audio recording level max-pooling) to compute a sparse-convex discriminative representation. The proposed deep factorization model is based on archetypal analysis, a form of convex NMF, which has been shown to be well suited for audio analysis. Finally, the obtained representation is mapped to a class label using a dictionary based auto-encoder consisting of linear and symmetric encoder and decoder with an efficient learning algorithm. The encoder projects the ASe of a waveform to the label space, while the decoder ensures that the feature can be reconstructed, resulting in better generalization on the test data.

## Deep Convolutional Neural Network with Scalogram for Audio Scene Modeling

*Hangting Chen[1,2], Pengyuan Zhang[1,2], Haichuan Bai[1,2], Qingsheng Yuan[3], Xiuguo Bao[3] and Yonghong Yan[1,2]*

[1]Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China

[2]University of Chinese Academy of Sciences, Beijing 100049, China

[3]National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing, 100029 China

**Thu-P-1-2-6, Time: 10:00-12:00**

Deep learning has improved the performance of acoustic scene classification recently. However, learning is usually based on short-time Fourier transform and hand-tailored filters. Learning directly from raw signals has remained a big challenge. In this paper, we proposed an approach to learning audio scene patterns from scalogram, which is extracted from raw signal with simple wavelet transforms. The experiments were conducted on DCASE2016 dataset. We compared scalogram with classical Mel energy, which showed that multi-scale feature led to an obvious accuracy increase. The convolutional neural network integrated with maximum-average downsampled scalogram achieved an accuracy of 90.5% in the evaluation step in DCASE2016.

## Time Aggregation Operators for Multi-label Audio Event Detection

*Pankaj Joshi, Digvijaysingh Gautam, Ganesh Ramakrishnan and Preethi Jyothi*

Dept. of Computer Science and Engineering, Indian Institute of Technology Bombay

**Notes**

**Thu-P-1-2-7, Time: 10:00-12:00**
In this paper, we present a state-of-the-art system for audio event detection. The labels on the training (and evaluation) data specify the set of events occurring in each audio clip, but neither the time spans nor the order in which they occur. Specifically, our task of weakly supervised learning is the "Detection and Classification of Acoustic Scenes and Events (DCASE) 2017" challenge. We use the winning entry in this challenge given by Xu et al. as our starting point and identify several important modifications that allow us to improve on their results significantly. Our techniques pertain to aggregation and consolidation over time and frequency signals over a (temporal) sequence before decoding the labels. In general, our work is also relevant to other tasks involving learning from weak labeling of sequential data.

### Early Detection of Continuous and Partial Audio Events Using CNN

*Ian McLoughlin[1,2], Yan Song[2], Lam Dang Pham[1], Ramaswamy Palaniappan[1], Huy Phan[3] and Yue Lang[4]*

[1]The University of Kent, School of Computing, Medway, UK

[2]The University of Science and Technology of China, Hefei, PRC

[3]University of Oxford, Department of Engineering Science, Oxford, UK

[4]Huawei European Research Center, Munich, Germany

**Thu-P-1-2-8, Time: 10:00-12:00**
Sound event detection is an extension of the static auditory classification task into continuous environments, where performance depends jointly upon the detection of overlapping events and their correct classification.

Several approaches have been published to date which either develop novel classifiers or employ well-trained static classifiers with a detection front-end.

This paper takes the latter approach, by combining a proven CNN classifier acting on spectrogram image features, with time-frequency shaped energy detection that identifies seed regions within the spectrogram that are characteristic of auditory energy events. Furthermore, the shape detector is optimised to allow early detection of events as they are developing. Since some sound events naturally have longer durations than others, waiting until completion of entire events before classification may not be practical in a deployed system. The early detection capability of the system is thus evaluated for the classification of partial events.

Performance for continuous event detection is shown to be good, with accuracy being maintained well when detecting partial events.

### Robust Acoustic Event Classification Using Bag-of-Visual-Words

*Manjunath Mulimani and Shashidhar G Koolagudi*

National Institute of Technology Karnataka, Surathkal, India

**Thu-P-1-2-9, Time: 10:00-12:00**
This paper presents a novel Bag-of-Visual-Words (BoVW) approach, to represent the grayscale spectrograms of acoustic events. Such, BoVW representations are referred as histograms of visual features, used for Acoustic Event Classification (AEC). Further, Chi-square distance between histograms of visual features evaluated, which generates kernel to Support Vector Machines (Chi-square SVM) classifier. Evaluation of the proposed histograms of visual features together with Chi-square SVM classifier is conducted on different categories of acoustic events from UPC-TALP corpora in clean and different noise conditions. Results show that proposed approach is more robust to noise and achieves improved recognition accuracy compared to other methods.

### Wavelet Transform Based Mel-scaled Features for Acoustic Scene Classification

*Shefali Waldekar and Goutam Saha*

Dept of Electronics and Electrical Communication Engineering, IIT Kharagpur, Kharagpur, India

**Thu-P-1-2-10, Time: 10:00-12:00**
Acoustic scene classification (ASC) is an audio signal processing task where mel-scaled spectral features are widely used by researchers. These features, considered de facto baseline in speech processing, traditionally employ Fourier based transforms. Unlike speech, environmental audio spans a larger range of audible frequency and might contain short high-frequency transients and continuous low-frequency background noise, simultaneously. Wavelets, with a better time-frequency localization capacity, can be considered more suitable for dealing with such signals. This paper attempts ASC by a novel use of wavelet transform based mel-scaled features. The proposed features are shown to possess better discriminative properties than other spectral features while using a similar classification framework. The experiments are performed on two datasets, similar in scene classes but differing by dataset size and length of the audio samples. When compared with two benchmark systems, one based on mel-frequency cepstral coefficients and Gaussian mixture models and the other based on log mel-band energies and multi-layer perceptron, the proposed system performed considerably better on the test data.

### Multi-modal Attention Mechanisms in LSTM and Its Application to Acoustic Scene Classification

*Teng Zhang, Kailai Zhang and Ji Wu*

Multimedia Signal and Intelligent Information Processing Lab, Department of Electronic Engineering, Tsinghua University, Beijing, P.R.China

**Thu-P-1-2-11, Time: 10:00-12:00**
Neural network architectures such as long short-term memory (LSTM) have been proven to be powerful models for processing sequences including text, audio and video. On the basis of vanilla LSTM, multi-modal attention mechanisms are proposed in this paper to synthesize the time and semantic information of input sequences. First, we reconstruct the forget and input gates of the LSTM unit from the perspective of attention model in the temporal dimension. Then the memory content of the LSTM unit is recalculated using a cluster-based attention mechanism in semantic space. Experiments on acoustic scene classification tasks show performance improvements of the proposed methods when compared with vanilla LSTM. The classification errors on LITIS ROUEN dataset and DCASE2016 dataset are reduced by 16.5% and 7.7% relatively. We get a second place in the Kaggle's YouTube-8M video understanding challenge and multi-modal attention based LSTM model is one of our best-performing single systems.

## Thu-P-1-3: Language Modeling

Hall 4-6: Poster-3; 10:00-12:00; Thursday, 6 September, 2018

Chair: Mikko Kurimo

### Contextual Language Model Adaptation for Conversational Agents

*Anirudh Raju[1], Behnam Hedayatnia[1], Linda Liu[2], Ankur Gandhe[1], Chandra Khatri[1], Angeliki Metallinou[1], Anu Venkatesh[1] and Ariya Rastrow[1]*

[1]Amazon Alexa Machine Learning

[2]University of Rochester

**Thu-P-1-3-1, Time: 10:00-12:00**
Statistical language models (LM) play a key role in Automatic Speech Recognition (ASR) systems used by conversational agents. These ASR systems should provide a high accuracy under a variety of speaking styles, domains, vocabulary and argots. In this paper, we present a DNN-based method to adapt the LM to each user-agent interaction based on generalized contextual information, by predicting an optimal, context-dependent set of LM interpolation weights. We show that this framework for contextual adaptation provides accuracy improvements under different possible mixture LM partitions that are relevant for both (1) Goal-oriented

**Notes**

conversational agents where it's natural to partition the data by the requested application and for (2) Non-goal oriented conversational agents where the data can be partitioned using topic labels that come from predictions of a topic classifier. We obtain a relative WER reduction of 3% with a 1-pass decoding strategy and 6% in a 2-pass decoding framework, over an unadapted model. We also show up to a 15% relative WER reduction in recognizing named entities which is of significant value for conversational ASR systems.

## Active Memory Networks for Language Modeling

*Oscar Chen, Anton Ragni, Mark Gales and Xie Chen*

Cambridge University Engineering Dept., Cambridge, U.K

**Thu-P-1-3-2, Time: 10:00-12:00**

Making predictions of the following word given the back history of words may be challenging without meta-information such as the topic. Standard neural network language models have an implicit representation of the topic via the back history of words. In this work a more explicit form of topic representation is used via an attention mechanism. Though this makes use of the same information as the standard model, it allows parameters of the network to focus on different aspects of the task. The attention model provides a form of topic representation that is automatically learned from the data. Whereas the recurrent model deals with the (conditional) history representation. The combined model is expected to reduce the stress on the standard model to handle multiple aspects. Experiments were conducted on the Penn Tree Bank and BBC Multi-Genre Broadcast News (MGB) corpora, where the proposed approach outperforms standard forms of recurrent models in perplexity. Finally, N-best list rescoring for speech recognition in the MGB3 task shows word error rate improvements over comparable standard form of recurrent models.

## Unsupervised and Efficient Vocabulary Expansion for Recurrent Neural Network Language Models in ASR

*Yerbolat Khassanov and Eng Siong Chng*

Rolls-Royce@NTU Corporate Lab, Nanyang Technological University, Singapore

**Thu-P-1-3-3, Time: 10:00-12:00**

In automatic speech recognition (ASR) systems, recurrent neural network language models (RNNLM) are used to rescore a word lattice or N-best hypotheses list. Due to the expensive training, the RNNLM's vocabulary set accommodates only small shortlist of most frequent words. This leads to suboptimal performance if an input speech contains many out-of-shortlist (OOS) words.

An effective solution is to increase the shortlist size and retrain the entire network which is highly inefficient. Therefore, we propose an efficient method to expand the shortlist set of a pretrained RNNLM without incurring expensive retraining and using additional training data. Our method exploits the structure of RNNLM which can be decoupled into three parts: input projection layer, middle layers and output projection layer. Specifically, our method expands the word embedding matrices in projection layers and keeps the middle layers unchanged. In this approach, the functionality of the pretrained RNNLM will be correctly maintained as long as OOS words are properly modeled in two embedding spaces. We propose to model the OOS words by borrowing linguistic knowledge from appropriate in-shortlist words. Additionally, we propose to generate the list of OOS words to expand vocabulary in unsupervised manner by automatically extracting them from ASR output.

## Improving Language Modeling with an Adversarial Critic for Automatic Speech Recognition

*Yike Zhang[1,2], Pengyuan Zhang[1,2] and Yonghong Yan[1,2,3]*

[1]Institute of Acoustics, Chinese Academy of Sciences, China

[2]University of Chinese Academy of Sciences, China

[3]Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences

**Thu-P-1-3-4, Time: 10:00-12:00**

Recurrent neural network language models (RNN LMs) trained via the maximum likelihood principle suffer from the exposure bias problem in the inference stage. Therefore, potential recognition errors limit their performance on re-scoring N-best lists of the speech recognition outputs. Inspired by the generative adversarial net (GAN), this paper proposes a novel approach to alleviate this problem. We regard the RNN LM as a generative model in the training stage. And an auxiliary neural critic is used to encourage the RNN LM to learn long-term dependencies from corrupted contexts by forcing it generating valid sentences. Since the vanilla GAN has limitations when generating discrete sequences, the proposed framework is optimized though the policy gradient algorithm. Experiments were conducted on two mandarin speech recognition tasks. Results show the proposed method achieved lower character error rates on both datasets compared with the maximum likelihood method, whereas it increased perplexities slightly. Finally, we visualised the sentences generated from the RNN LM. Results demonstrate the proposed method really helps the RNN LM to learn long-term dependencies and alleviates the exposure bias problem.

## Training Recurrent Neural Network through Moment Matching for NLP Applications

*Yue Deng[1], Yilin Shen[1], KaWai Chen[1,2] and Hongxia Jin[1]*

[1]AI Center, Samsung Research America, Mountain View, CA, USA

[2]University of California, San Diego, La Jolla, CA, USA

**Thu-P-1-3-5, Time: 10:00-12:00**

Recurrent neural network (RNN) is conventionally trained in the supervised mode but used in the free-running mode for inferences on testing samples. The supervised mode takes ground truth token values as RNN inputs but the free-running mode can only use self-predicted token values as surrogating inputs. Such inconsistency inevitably results in poor generalizations of RNN on out-of-sample data. We propose a moment matching (MM) training strategy to alleviate such inconsistency by simultaneously taking these two distinct modes and their corresponding dynamics into consideration. Our MM-RNN shows significant performance improvements over existing approaches when tested on practical NLP applications including logic form generation and image captioning.

## Investigation on LSTM Recurrent N-gram Language Models for Speech Recognition

*Zoltán Tüske[1,2], Ralf Schlüter[1] and Hermann Ney[1]*

[1]Human Language Technology and Pattern Recognition, Computer Science Department, RWTH Aachen University, 52056 Aachen, Germany

[2]IBM Research, Thomas J. Watson Research Center, P.O. Box 704, Yorktown Heights, NY 10598

**Thu-P-1-3-6, Time: 10:00-12:00**

Recurrent neural networks (NN) with long short-term memory (LSTM) are the current state of the art to model long term dependencies. However, recent studies indicate that NN language models (LM) need only limited length of history to achieve excellent performance. In this paper, we extend the previous investigation on LSTM network based n-gram modeling to the domain of automatic speech recognition (ASR). First, applying recent optimization techniques and up to 6-layer LSTM networks, we improve LM perplexities by nearly 50% relative compared to classic count models on three different domains. Then, we demonstrate by experimental results that perplexities improve significantly only up to 40-grams when limiting the LM history. Nevertheless, the ASR performance saturates already around 20-grams despite across sentence modeling. Analysis indicates that the performance gain of LSTM NNLM over count models results only partially from the longer context and cross sentence modeling capabilities. Using equal context, we show that deep 4-gram LSTM can significantly outperform large interpolated count models by performing the backing off and smoothing significantly better. This observation also underlines the decreasing importance to combine state-of-the-art deep NNLM with count based model.

## Online Incremental Learning for Speaker-Adaptive Language Models

*Chih Chi Hu, Bing Liu, John Shen and Ian Lane*

Electrical and Computer Engineering, Carnegie Mellon University,

**Notes**

USA

**Thu-P-1-3-7, Time: 10:00-12:00**

In this study, we present a computational framework to participate in the Self-Assessed Affect Sub-Challenge in the INTERSPEECH 2018 Computation Paralinguistics Challenge. The goal of this sub-challenge is to classify the valence scores given by the speaker themselves into three different levels, i.e., low, medium and high. We explore fusion of Bi-directional LSTM with baseline SVM models to improve the recognition accuracy. In specifics, we extract frame-level acoustic LLDs as input to the BLSTM with a modified attention mechanism and separate SVMs are trained using the standard ComParE_16 baseline feature sets with minority class upsampling. These diverse prediction results are then further fused using a decision-level score fusion scheme to integrate all of the developed models. Our proposed approach achieves a 62.94% and 67.04% unweighted average recall (UAR), which is an 6.24% and 1.04% absolute improvement over the best baseline provided by the challenge organizer. We further provide a detailed comparison analyVoice control is a prominent interaction method on personal computing devices. While automatic speech recognition (ASR) systems are readily applicable for large audiences, there is room for further adaptation at the edge, ie. locally on devices, targeted for individual users. In this work, we explore improving ASR systems over time through a user's own interactions. Our online learning approach for speaker-adaptive language modeling leverages a user's most recent utterances to enhance the speaker dependent features and traits. We experiment with the Large-Vocabulary Continuous Speech Recognition corpus Tedlium v2 and demonstrate an average reduction in perplexity (PPL) of 19.18% and average relative reduction in word error rate (WER) of 2.80% compared to a state-of-the-art baseline on Tedlium v2.sis between different models.

### Efficient Language Model Adaptation with Noise Contrastive Estimation and Kullback-Leibler Regularization

*Jesús Andrés-Ferrer, Nathan Bodenstab and Paul Vozila*

Nuance Communications

**Thu-P-1-3-8, Time: 10:00-12:00**

Many language modeling (LM) tasks have limited in-domain data for training. Exploiting out-of-domain data while retaining the relevant in-domain statistics is a desired property in these scenarios.

Kullback-Leibler Divergence (KLD) regularization is a popular method for acoustic model (AM) adaptation. KLD regularization assumes that the last layer is a softmax that fully activates the targets of both in-domain and out-of-domain models.

Unfortunately, this softmax activation is computationally prohibitive for language modeling where the number of output classes is large, typically 50k to 100K, but may even exceed 800k in some cases.
The computational bottleneck of the softmax during LM training can be reduced by an order of magnitude using techniques such as noise contrastive estimation (NCE), which replaces the cross-entropy loss function with a binary classification problem between the target output and random noise samples.

In this work we combine NCE and KLD regularization and offer a fast domain adaptation method for LM training, while also retaining important attributes of the original NCE, such as self-normalization.
We show on a medical domain-adaptation task that our method improves perplexity by 10.1% relative to a strong LSTM baseline.

### Recurrent Neural Network Language Model Adaptation for Conversational Speech Recognition

*Ke Li[1], Hainan Xu[1], Yiming Wang[1], Daniel Povey[1,2] and Sanjeev Khudanpur[1,2]*

[1]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA

[2]Human Language Technology Center of Excellence, Johns Hopkins University, Baltimore, MD, USA

**Thu-P-1-3-9, Time: 10:00-12:00**

We propose two adaptation models for recurrent neural network language models (RNNLMs) to capture topic effects and long-distance triggers for conversational automatic speech recognition (ASR). We use a fast marginal adaptation (FMA) framework to adapt a RNNLM. Our first model is effectively a cache model - the word frequencies are estimated by counting words in a conversation (with utterance-level hold-one-out) from 1st-pass decoded word lattices and then is interpolated with a background unigram distribution. In the second model, we train a deep neural network (DNN) on conversational transcriptions to predict word frequencies given word frequencies from 1st-pass decoded word lattices. The second model can in principle model trigger and topic effects but is harder to train. Experiments on three conversational corpora show modest WER and perplexity reductions with both adaptation models.

### What to Expect from Expected Kneser-Ney Smoothing

*Michael Levit, Sarangarajan Parthasarathy and Shuangyu Chang*

Microsoft, USA

**Thu-P-1-3-10, Time: 10:00-12:00**

Kneser-Ney smoothng on expected counts was proposed recently. In this paper we revisit this technique and suggest a number of optimizations and extensions. We then analyse its performance in several practical speech recognition scenarios that depend on fractional sample counts, such as training on uncertain data, language model adaptation and Word-Phrase-Entity models. We show that the proposed approach to smoothing outperforms known alternatives by a significant margin.

### i-Vectors in Language Modeling: An Efficient Way of Domain Adaptation for Feed-Forward Models

*Karel Beneš[1], Santosh Kesiraju[1,2] and Lukáš Burget[1]*

[1]Brno University of Technology, Speech@FIT and IT4I Center of Excellence, Czechia

[2]IIIT - Hyderabad, India

**Thu-P-1-3-11, Time: 10:00-12:00**

We show an effective way of adding context information to shallow neural language models. We propose to use Subspace Multinomial Model (SMM) for context modeling and we add the extracted i-vectors in a computationally efficient way. By adding this information, we shrink the gap between shallow feed-forward network and an LSTM from 65 to 31 points of perplexity on the Wikitext-2 corpus (in the case of neural 5-gram model). Furthermore, we show that SMM i-vectors are suitable for domain adaptation and a very small amount of adaptation data (e.g. endmost 5% of a Wikipedia article) brings asubstantial improvement. Our proposed changes are compatible with most optimization techniques used for shallow feedforward LMs.

.....

## Thu-P-1-4: Speech Pathology, Depression and Medical Applications

Hall 4-6: Poster-4; 10:00-12:00; Thursday, 6 September, 2018

Chair: Elmar Noeth

.....

### How Did You like 2017? Detection of Language Markers of Depression and Narcissism in Personal Narratives

*Eva-Maria Rathner [1], Julia Djamali [1], Yannik Terhorst [1], Björn Schuller [2,3], Nicholas Cummins [2], Gudrun Salamon [4], Christina Hunger-Schoppe [5] and Harald Baumeister [1]*

[1]Clinical Psychology and Psychotherapy, University of Ulm, Germany

[2]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany

**Notes**

[3]GLAM – Group on Language, Audio & Music, Imperial College London, UK
[4]Sigmund Freud University, Vienna, Austria
[5]Institute of Medical Psychology, University of Heidelberg, Germany

**Thu-P-1-4-1, Time: 10:00-12:00**

Language analyses reveals crucial information about an individual's current state of mind. Maladaptive psychological functioning appears in cognition, emotional experience and behaviour. In the time of the internet of things, a vast number of text and speech is available; subsequently, the interest in the automated detection of psychological functioning via language is rising. The current study indicates that depression and narcissism can be predicted through word use in personal narratives. Both conditions are characterised by an altered word count regarding anxiety and we (LIWC-based). While depressive individuals use less social words and more anxiety-related words, narcissists do the opposite. This might reflect the verbal correlate of the cognitive triad in depression. In contrast, narcissists' word use mirrors their excommunicated anxiety of being an undesired self and their inability to reach long-term goals due to a lack of impulse control. The automated recognition of mental state through word use could improve early detection of mental disease, monitoring of disease course, delivery of tailored interventions and evaluation of therapy outcome.

## Depression Detection from Short Utterances via Diverse Smartphones in Natural Environmental Conditions

*Zhaocheng Huang[1], Julien Epps[1], Dale Joachim[2] and Michael Chen[2]*
[1]School of Electrical Engineering and Telecommunications, UNSW Sydney, Australia
[2]Sonde Health, Boston MA, USA

**Thu-P-1-4-2, Time: 10:00-12:00**

Depression is a leading cause of disease burden worldwide, however there is an unmet need for screening and diagnostic measures that can be widely deployed in real-world environments. Voice-based diagnostic methods are convenient, non-invasive to elicit and can be collected and processed in near real-time using modern smartphones, smart speakers and other devices. Studies in voice-based depression detection to date have primarily focused on laboratory-collected voice samples, which are not representative of typical user environments or devices. This paper conducts the first investigation of voice-based depression assessment techniques on real-world data from 887 speakers, recorded using a variety of different smartphones. Evaluations on 16 hours of speech show that conservative segment selection strategies using highly thresholded voice activity detection, coupled with tailored normalization approaches are effective for mitigating smartphone channel variability and background environmental noise. Together, these strategies can achieve F1 scores comparable with or better than those from a combination of clean recordings, a single recording environment and long utterances. The scalability of speech elicitation via smartphone allows detailed models dependent on gender, smartphone manufacturer and/or elicitation task. Interestingly, results herein suggest that normalization based on these criteria may be more effective than tailored models for detecting depressed speech.

## Multi-Lingual Depression-Level Assessment from Conversational Speech Using Acoustic and Text Features

*Yasin Özkanca[1], Cenk Demiroglu[1], Aslı Besirli[2] and Selime Celik[2]*
[1]Ozyegin University, Turkey
[2]Sisli Etfal Hospital, Turkey

**Thu-P-1-4-3, Time: 10:00-12:00**

Depression is a common mental health problem around the world with a large burden on economies, well-being, hence productivity, of individuals. Its early diagnosis and treatment are critical to reduce the costs and even save lives. One key aspect to achieve that goal is to use voice technologies and monitor depression remotely and relatively inexpensively using automated agents.

Although there has been efforts to automatically assess depression levels from audiovisual features, use of transcriptions along with the acoustic features has emerged as a more recent research venue. Moreover, difficulty in data collection and the limited amounts of data available for research are also challenges that are hampering the success of the algorithms. One of the novel contributions in this paper is to exploit the databases from multiple languages for feature selection. Since a large number of features can be extracted from speech and given the small amounts of training data available, effective data selection is critical for success. Our proposed multi-lingual method was effective at selecting better features and significantly improved the depression assessment accuracy. We also use text-based features for assessment and propose a novel strategy to fuse the text- and speech-based classifiers which further boosted the performance.

## Dysarthric Speech Classification Using Glottal Features Computed from Non-words, Words and Sentences

*Narendra N P and Paavo Alku*
Aalto University, Department of Signal Processing and Acoustics, Espoo, Finland

**Thu-P-1-4-4, Time: 10:00-12:00**

Dysarthria is a neuro-motor disorder resulting from the disruption of normal activity in speech production leading to slow, slurred and imprecise (low intelligible) speech. Automatic classification of dysarthria from speech can be used as a potential clinical tool in medical treatment. This paper examines the effectiveness of glottal source parameters in dysarthric speech classification from three categories of speech signals, namely non-words, words and sentences. In addition to the glottal parameters, two sets of acoustic parameters extracted by the openSMILE toolkit are used as baseline features. A dysarthric speech classification system is proposed by training support vector machines (SVMs) using features extracted from speech utterances and their labels indicating dysarthria/healthy. Classification accuracy results indicate that the glottal parameters contain discriminating information required for the identification of dysarthria. Additionally, the complementary nature of the glottal parameters is demonstrated when these parameters, in combination with the openSMILE-based acoustic features, result in improved classification accuracy. Analysis of classification accuracies of the glottal and openSMILE features for non-words, words and sentences is carried out. Results indicate that in terms of classification accuracy the word level is best suited in identifying the presence of dysarthria.

## Identifying Schizophrenia Based on Temporal Parameters in Spontaneous Speech

*Gábor Gosztolya[1], Anita Bagi[2,5], Szilvia Szalóki[3,5], István Szendi[3,5] and Ildikó Hoffmann[2,4,5]*
[1]MTA-SZTE Research Group on Artificial Intelligence, Szeged, Hungary
[2]University of Szeged, Department of Hungarian Linguistics, Szeged, Hungary
[3]Department of Psychiatry, University of Szeged, Hungary
[4]Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, Hungary
[5]Prevention of Mental Illnesses Interdisciplinary Research Group, University of Szeged, Hungary

**Thu-P-1-4-5, Time: 10:00-12:00**

Schizophrenia is a neurodegenerative disease with spectrum disorder, consisting of groups of different deficits. It is, among other symptoms, characterized by reduced information processing speed and deficits in verbal fluency. In this study we focus on the speech production fluency of patients with schizophrenia compared to healthy controls. Our aim is to show that a temporal speech parameter set consisting of articulation tempo, speech tempo and various pause-related indicators, originally defined for the sake of early detection of various dementia types such as Mild Cognitive Impairment and early Alzheimer's Disease, is able to capture specific differences in the spontaneous speech of the two groups. We tested the applicability of the temporal indicators by machine learning (i.e. by using Support-Vector Machines). Our results show that members of the

**Notes**

two speaker groups could be identified with classification accuracy scores of between 70-80% and F-measure scores between 81% and 87%. Our detailed examination revealed that, among the pause-related temporal parameters, the most useful for distinguishing the two speaker groups were those which took into account both the silent and filled pauses.

## Using Prosodic and Lexical Information for Learning Utterance-level Behaviors in Psychotherapy

*Karan Singla[1], Zhuohao Chen[1], Nikolaos Flemotomos[1], James Gibson[1], Dogan Can[1], David Atkins[2] and Shrikanth Narayanan[1]*

[1]Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, CA, USA

[2]Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA

**Thu-P-1-4-6, Time: 10:00-12:00**

In this paper, we present an approach for predicting utterance level behaviors in psychotherapy sessions using both speech and lexical features. We train long short term memory (LSTM) networks with an attention mechanism using words, both manually and automatically transcribed and prosodic features, at the word level, to predict the annotated behaviors. We demonstrate that prosodic features provide discriminative information relevant to the behavior task and show that they improve prediction when fused with automatically derived lexical features. Additionally, we investigate the weights of the attention mechanism to determine words and prosodic patterns which are of importance to the behavior prediction task.

## Automatic Speech Assessment for People with Aphasia Using TDNN-BLSTM with Multi-Task Learning

*Ying Qin[1], Tan Lee[1], Siyuan Feng[1] and Anthony Pak Hin Kong[2]*

[1]Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong

[2]Department of Communication Sciences and Disorders, University of Central Florida, USA

**Thu-P-1-4-7, Time: 10:00-12:00**

This paper describes an investigation on automatic speech assessment for people with aphasia (PWA) using a DNN based automatic speech recognition (ASR) system. The main problems being addressed are the lack of training speech in the intended application domain and the relevant degradation of ASR performance for impaired speech of PWA. We adopt the TDNN-BLSTM structure for acoustic modeling and apply the technique of multi-task learning with large amount of domain-mismatched data. This leads to a significant improvement on the recognition accuracy, as compared with a conventional single-task learning DNN system. To facilitate the extraction of robust text features for quantifying language impairment in PWA speech, we propose to incorporate N-best hypotheses and confusion network representation of the ASR output. The severity of impairment is predicted from text features and supra-segmental duration features using different regression models. Experimental results show a high correlation of 0.842 between the predicted severity level and the subjective Aphasia Quotient score.

## Towards an Unsupervised Entrainment Distance in Conversational Speech Using Deep Neural Networks

*Md Nasir[1], Brian Baucom[2], Shrikanth Narayanan[1] and Panayiotis Georgiou[1]*

[1]University of Southern California, Los Angeles, CA, USA

[2]University of Utah, Salt Lake City, UT, USA

**Thu-P-1-4-8, Time: 10:00-12:00**

Entrainment is a known adaptation mechanism that causes interaction participants to adapt or synchronize their acoustic characteristics. Understanding how interlocutors tend to adapt to each other's speaking style through entrainment involves measuring a range of acoustic features and comparing those via multiple signal comparison methods. In this work, we present a turn-level distance measure obtained in an unsupervised manner using a Deep Neural Network (DNN) model, which we call Neural Entrainment Distance (NED). This metric establishes a framework that learns an embedding from the population-wide entrainment in an unlabeled training corpus. We use the framework for a set of acoustic features and validate the measure experimentally by showing its efficacy in distinguishing real conversations from fake ones created by randomly shuffling speaker turns. Moreover, we show real world evidence of the validity of the proposed measure. We find that high value of NED is associated with high ratings of emotional bond in suicide assessment interviews, which is consistent with prior studies.

## Patient Privacy in Paralinguistic Tasks

*Francisco Teixeira, Alberto Abad and Isabel Trancoso*

INESC-ID / Instituto Superior Técnico, University of Lisbon, Portugal

**Thu-P-1-4-9, Time: 10:00-12:00**

Recent developments in cryptography and, in particular in Fully Homomorphic Encryption (FHE), have allowed for the development of new privacy preserving machine learning schemes. In this paper, we show how these schemes can be applied to the automatic assessment of speech affected by medical conditions, allowing for patient privacy in diagnosis and monitoring scenarios. More specifically, we present results for the assessment of the degree of Parkinson's Disease, the detection of a Cold and both the detection and assessment of the degree of Depression.

To this end, we use a neural network in which all operations are performed in an FHE context. This implies replacing the activation functions by linear and second degree polynomials, as only additions and multiplications are viable. Furthermore, to guarantee that the inputs of these activation functions fall within the convergence interval of the approximation, a batch normalization layer is introduced before each activation function. After training the network with unencrypted data, the resulting model is then employed in an encrypted version of the network, to produce encrypted predictions.

Our tests show that the use of this framework yields results with little to no performance degradation, in comparison to the baselines produced for the same datasets.

## A Lightly Supervised Approach to Detect Stuttering in Children's Speech

*Sadeen Alharbi[1], Madina Hasan[1], Anthony J H Simons[1], Shelagh Brumfitt[2] and Phil Green[1]*

[1]Computer Science Department, The University of Sheffield,

[2]Human Communication Sciences Department, The University of Sheffield, Sheffield, United Kingdom

**Thu-P-1-4-10, Time: 10:00-12:00**

In speech pathology, new assistive technologies using ASR and machine learning approaches are being developed for detecting speech disorder events. Classically-trained ASR model tends to remove disfluencies from spoken utterances, due to its focus on producing clean and readable text output. However, diagnostic systems need to be able to track speech disfluencies, such as stuttering events, in order to determine the severity level of stuttering. To achieve this, ASR systems must be adapted to recognise full verbatim utterances, including pseudo-words and non-meaningful part-words. This work proposes a training regime to address this problem and preserve a full verbatim output of stuttering speech. We use a lightly-supervised approach using task-oriented lattices to recognise the stuttering speech of children performing a standard reading task. This approach improved the WER by 27.8% relative to a baseline that uses word-lattices generated from the original prompt. The improved results preserved 63% of stuttering events (including sound, word, part-word and phrase repetition and revision). This work also proposes a separate correction layer on top of the ASR that detects prolongation events (which are poorly recognised by the ASR). This increases the percentage of preserved stuttering events to 70%.

**Notes**

### Learning Conditional Acoustic Latent Representation with Gender and Age Attributes for Automatic Pain Level Recognition

*Jeng-Lin Li[1,5], Yi-Ming Weng[2,3,4], Chip-Jin Ng[2] and Chi-Chun Lee[1,5]*
[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan
[2]Department of Emergency Medicine, Chang Gung Memorial Hospital, Taiwan
[3]Department of Emergency Medicine, Tao-Yuan General Hospital, Taiwan
[4]Faculty of Medicine, National Yang-Ming University, Taiwan
[5]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan

**Thu-P-1-4-11, Time: 10:00-12:00**

Pain is an unpleasant internal sensation caused by bodily damages or physical illnesses with varied expressions conditioned on personal attributes. In this work, we propose an age-gender embedded latent acoustic representation learned using conditional maximum mean discrepancy variational autoencoder (MMD-CVAE). The learned MMD-CVAE embeds personal attributes information directly in the latent space. Our method achieves a 70.7% in extreme set classification (severe versus mild) and 47.7% in three-class recognition (severe, moderate and mild) by using these MMD-CVAE encoded features on a large-scale real patients pain database. Our method improves a relative of 11.34% and 17.51% compared to using acoustic representation without age-gender conditioning in the extreme set and the three-class recognition respectively. Further analyses reveal under severe pain, females have higher maximum of jitter and lower harmonic energy ratio between F0, H1 and H2 compared to males and the minimum value of jitter and shimmer are higher in the elderly compared to the non-elder group.

## Perspective Talk-4

Hall 3; 12:00-12:30; Thursday, 6 September, 2018
Chair: Srikanth Madikeri

### Speaker and Language Recognition -- From Laboratory Technologies to the Wild

*Sriram Ganapathy*
Department of Electrical Engineering, Indian Institute of Science, Bangalore, India

**Thu-Perspective-4, Time: 12:00-12:30**

Detecting the paralinguistic components of speech like speaker and language is of substantial interest for many commercial, surveillance and security applications. The problem is at least three decades old with some of the early techniques based on simple Gaussian mixture models. A significant advancement in this area came about a decade ago with the advent of joint factor analysis and i-vector models. The last couple of years have seen further breakthroughs with deep embeddings and end-to-end models based on deep learning. With these improvements in modeling speaker and language, the application of the technology has also moved from clean controlled speech data to telephone channel recordings, far-field microphones and more recently to multi-speaker conversations in the wild. In the talk, I will provide a prospective view of the broad research directions in the field of speaker and language recognition. I will also highlight some of the recent advancements from our work on hierarchical end-to-end approaches with relevance modeling.

## Industry Presentation-7

Hall 3; 12:30-13:00; Thursday, 6 September, 2018
Chairs: Raghavendra Bilgi and Priyankoo Sarmah

**Presenter:** Vikram Vij
**Company:** Samsung

## Industry Presentation-8

Hall 1; 12:30-13:00; Thursday, 6 September, 2018
Chairs: Krishna Doss Mohan and Samudravijaya K

**Presenter:** Liang Gao
**Company:** Baidu

## Industry Presentation-9

Hall 2; 12:30-13:00; Thursday, 6 September, 2018
Chairs: Kalika Bali and Deepu Vijayasenan

**Presenter:** Ryan Leary
**Company:** Nvidia

## Thu-O-2-1: Spoken Language Understanding

Hall 3; 14:30-16:30; Thursday, 6 September, 2018
Chairs: Dilek Hakkani-Tür and Giuseppe Riccardi

### A Deep Reinforcement Learning Based Multimodal Coaching Model (DCM) for Slot Filling in Spoken Language Understanding(SLU)

*Yu Wang, Abhishek Patel, Yilin Shen and Hongxia Jin*
Samsung Research America

**Thu-O-2-1-1, Time: 14:30-14:50**

In this paper, a deep reinforcement learning(DRL) based multimodal coaching model (DCM) for slot filling in SLU is proposed. This new model functions as a

**Notes**

coach to help an RNN based tagger to learn the wrong labeled slots, hence may further improve the performance of an SLU system. Besides, users can also coach the model by correcting its mistakes and help it progress further. The performance of DCM is evaluated on two datasets: one is the benchmark ATIS corpus dataset, another is our in-house dataset with three different domains. It shows that the new system gives a better performance than the current state-of-the-art results on ATIS by using DCM. Furthermore, we build a demo app to further explain how user's input can also be used as a real-time coach to improve model's performance even more.

### Is ATIS Too Shallow to Go Deeper for Benchmarking Spoken Language Understanding Models?

*Frédéric Béchet[1] and Christian Raymond[2]*

[1]Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France

[2]INSA Rennes INRIA/IRISA, Rennes, France

**Thu-O-2-1-2, Time: 14:50-15:10**

The ATIS (Air Travel Information Service) corpus will be soon celebrating its 30th birthday. Designed originally to benchmark spoken dialog systems, it still represents the most well-known corpus for benchmarking Spoken Language Understanding (SLU) systems. In 2010, in a paper titled "What is left to be understood in ATIS? " Tur et al. discussed the relevance of this corpus after more than 10 years of research on statistical models for performing SLU tasks. Nowadays, in the Deep Neural Network (DNN) era, ATIS is still used as the main benchmark corpus for evaluating all kinds of DNN models, leading to further improvements, although rather limited, in SLU accuracy compared to previous state-of-the-art models. We propose in this paper to investigate these results obtained on ATIS from a qualitative point of view rather than just a quantitative point of view and answer the two following questions: what kind of qualitative improvement brought DNN models to SLU on the ATIS corpus? Is there anything left, from a qualitative point of view, in the remaining 5% of errors made by current state-of the-art models?

### Robust Spoken Language Understanding via Paraphrasing

*Avik Ray, Yilin Shen and Hongxia Jin*

Samsung Research America, Mountain View, California, USA

**Thu-O-2-1-3, Time: 15:10-15:30**

Learning intents and slot labels from user utterances is a fundamental step in all spoken language understanding (SLU) and dialog systems. State-of-the-art neural network based methods, after deployment, often suffer from performance degradation on encountering paraphrased utterances and out-of-vocabulary words, rarely observed in their training set. We address this challenging problem by introducing a novel paraphrasing based SLU model which can be integrated with any existing SLU model in order to improve their overall performance. We propose two new paraphrase generators using RNN and sequence-to-sequence based neural networks, which are suitable for our application. Our experiments on existing benchmark and in house datasets demonstrate the robustness of our models to rare and complex paraphrased utterances, even under adversarial test distributions.

### Spoken SQuAD: A Study of Mitigating the Impact of Speech Recognition Errors on Listening Comprehension

*Chia-Hsuan Lee, Szu-Lin Wu, Chi-Liang Liu and Hung-yi Lee*

College of Electrical Engineering and Computer Science, National Taiwan University, Taiwan

**Thu-O-2-1-4, Time: 15:30-15:50**

Reading comprehension has been widely studied. One of the most representative reading comprehension tasks is Stanford Question Answering Dataset (SQuAD), on which machine is already comparable with human. On the other hand, accessing large collections of multimedia or spoken content is much more difficult and time-consuming than plain text content for humans. It's therefore highly attractive to develop machines which can automatically understand spoken content. In this paper, we propose a new listening comprehension task – Spoken SQuAD. On the new task, we found that speech recognition errors have catastrophic impact on machine comprehension and several approaches are proposed to mitigate the impact.

### User Information Augmented Semantic Frame Parsing Using Progressive Neural Networks

*Yilin Shen, Xiangyu Zeng, Yu Wang and Hongxia Jin*

Samsung Research America, Mountain View, CA, USA

**Thu-O-2-1-5, Time: 15:50-16:10**

Semantic frame parsing is a crucial component in spoken language understanding (SLU) to build spoken dialog systems. It has two main tasks: intent detection and slot filling. Although state-of-the-art approaches showed good results, they require large annotated training data and long training time. In this paper, we aim to alleviate these drawbacks for semantic frame parsing by utilizing the ubiquitous user information. We design a novel progressive deep neural network model to incorporate prior knowledge of user information intermediately to better and quickly train a semantic frame parser. Due to the lack of benchmark dataset with real user information, we synthesize the simplest type of user information (location and time) on ATIS benchmark data. The results show that our approach leverages such simple user information to outperform state-of-the-art approaches by 0.25% for intent detection and 0.31% for slot filling using standard training data. When using smaller training data, the performance improvement on intent detection and slot filling reaches up to 1.35% and 1.20% respectively. We also show that our approach can achieve similar performance as state-of-the-art approaches by using less than 80% annotated training data. Moreover, the training time to achieve the similar performance is also reduced by over 60%.

### An Efficient Approach to Encoding Context for Spoken Language Understanding

*Raghav Gupta, Abhinav Rastogi and Dilek Hakkani-Tür*

Google AI, Mountain View

**Thu-O-2-1-6, Time: 16:10-16:30**

In task-oriented dialogue systems, spoken language understanding, or SLU, refers to the task of parsing natural language user utterances into semantic frames. Making use of context from prior dialogue history holds the key to more effective SLU. State of the art approaches to SLU use memory networks to encode context by processing multiple utterances from the dialogue at each turn, resulting in significant trade-offs between accuracy and computational efficiency. On the other hand, downstream components like the dialogue state tracker (DST) already keep track of the dialogue state, which can serve as a summary of the dialogue history. In this work, we propose an efficient approach to encoding context from prior utterances for SLU. More specifically, our architecture includes a separate recurrent neural network (RNN) based encoding module that accumulates dialogue context to guide the frame parsing sub-tasks and can be shared between SLU and DST. In our experiments, we demonstrate the effectiveness of our approach on dialogues from two domains.

## Thu-O-2-2: Source Separation from Monaural Input

Hall 1; 14:30-16:30; Thursday, 6 September, 2018

Chairs: Pravesh Biyani and Michael Mandel

### Deep Speech Denoising with Vector Space Projections

*Jeffrey Hetherly, Paul Gamble, Maria Alejandra Barrios, Cory Stephenson and Karl Ni*

Lab41, In-Q-Tel

**Thu-O-2-2-1, Time: 14:30-14:50**

We propose an algorithm to denoise speakers from a single microphone in the presence of non-stationary and dynamic noise. Our approach is inspired

**Notes**

by the recent success of neural network models separating speakers from other speakers and singers from instrumental accompaniment. Unlike prior art, we leverage embedding spaces produced with source-contrastive estimation, a technique derived from negative sampling techniques in natural language processing, while simultaneously obtaining a continuous inference mask. Our embedding space directly optimizes for the discrimination of speaker and noise by jointly modeling their characteristics. This space is generalizable in that it is not speaker or noise specific and is capable of denoising speech even if the model has not seen the speaker in the training set. Parameters are trained with dual objectives: one that promotes a selective bandpass filter that eliminates noise at time-frequency positions that exceed signal power and another that proportionally splits time-frequency content between signal and noise. We compare to state of the art algorithms as well as traditional sparse non-negative matrix factorization solutions. The resulting algorithm avoids severe computational burden by providing a more intuitive and easily optimized approach, while achieving competitive accuracy.

## A Shifted Delta Coefficient Objective for Monaural Speech Separation Using Multi-task Learning

*Chenglin Xu[1,2], Wei Rao[2], Eng Siong Chng[1,2] and Haizhou Li[2,3]*

[1]School of Computer Science and Engineering, Nanyang Technological University, Singapore

[2]Temasek Laboratories@NTU, Nanyang Technological University, Singapore

[3]Department of Electrical and Computer Engineering, National University of Singapore, Singapore

**Thu-O-2-2-2, Time: 14:50-15:10**

This paper addresses the problem of monaural speech separation for simultaneous speakers. Recent studies such as uPIT, cuPIT-Grid LSTM and their variants have advanced the state-of-the-art separation models. Delta and acceleration coefficients are typically used in the objective function to capture short time dynamics. We consider that such coefficients don't benefit from the temporal information over a long range such as phoneme and syllable. In this paper, we propose a shifted delta coefficient (SDC) objective to explore the temporal information over a long range of the spectral dynamics. The SDC ensures the temporal continuity of output frames within the same speaker. In addition, we propose a novel multi-task learning framework, that we call SDC-MTL, by extending the SDC objective with a subtask of predicting the time-frequency labels ({silence, single, overlapped}) of the mixture. The experimental results show 11.7% and 3.9% relative improvements on WSJ0-2mix dataset under open conditions over the uPIT and cuPIT-Grid LSTM baselines. A further analysis shows 17.8% and 6.2% relative improvements with speakers of same gender.

## A Two-Stage Approach to Noisy Cochannel Speech Separation with Gated Residual Networks

*Ke Tan[1] and DeLiang Wang[1,2]*

[1]Department of Computer Science and Engineering, The Ohio State University, USA

[2]Center for Cognitive and Brain Sciences, The Ohio State University, USA

**Thu-O-2-2-3, Time: 15:10-15:30**

Cochannel speech separation is the task of separating two speech signals from a single mixture. The task becomes even more challenging if the speech mixture is further corrupted by background noise. In this study, we focus on a gender-dependent scenario, where target speech is from a male speaker and interfering speech from a female speaker. We propose a two-stage separation strategy to address this problem in a noise-independent way. In the proposed system, denoising and cochannel separation are performed successively by two modules, which are based on a newly-introduced convolutional neural network for speech separation. The evaluation results demonstrate that the proposed system substantially outperforms one-stage baselines in terms of objective intelligibility and perceptual quality.

## Monoaural Audio Source Separation Using Variational Autoencoders

*Laxmi Pandey, Anurendra Kumar and Vinay Namboodiri*

Indian Institute of Technology Kanpur

**Thu-O-2-2-4, Time: 15:30-15:50**

We introduce a monaural audio source separation framework using a latent generative model. Traditionally, discriminative training for source separation is proposed using deep neural networks or non-negative matrix factorization. In this paper, we propose a principled generative approach using variational autoencoders (VAE) for audio source separation. VAE computes efficient Bayesian inference which leads to a continuous latent representation of the input data(spectrogram). It contains a probabilistic encoder which projects an input data to latent space and a probabilistic decoder which projects data from latent space back to input space. This allows us to learn a robust latent representation of sources corrupted with noise and other sources. The latent representation is then fed to the decoder to yield the separated source. Both encoder and decoder are implemented via multilayer perceptron (MLP). In contrast to prevalent techniques, we argue that VAE is a more principled approach to source separation. Experimentally, we find that the proposed framework yields reasonable improvements when compared to baseline methods available in the literature i.e. DNN and RNN with different masking functions and autoencoders. We show that our method performs better than best of the relevant methods with ∼2dB improvement in the source to distortion ratio.

## Towards Automated Single Channel Source Separation Using Neural Networks

*Arpita Gang[1], Pravesh Biyani[1] and Akshay Soni[2]*

[1]IIIT Delhi, India

[2]Oath Research, USA

**Thu-O-2-2-5, Time: 15:50-16:10**

Many applications of single channel source separation (SCSS) including automatic speech recognition (ASR), hearing aids etc. require an estimation of only one source from a mixture of many sources. Treating this special case as a regular SCSS problem where in all constituent sources are given equal priority in terms of reconstruction may result in a suboptimal separation performance. In this paper, we tackle the one source separation problem by suitably modifying the orthodox SCSS framework and focus only on one source at a time. The proposed approach is a generic framework that can be applied to any existing SCSS algorithm, improves performance and scales well when there are more than two sources in the mixture unlike most existing SCSS methods. Additionally, existing SCSS algorithms rely on fine hyper-parameter tuning hence making them difficult to use in practice. Our framework takes a step towards automatic tuning of the hyper-parameters thereby making our method better suited for the mixture to be separated and thus practically more useful. We test our framework on a neural network based algorithm and the results show an improved performance in terms of SDR and SAR.

## Investigations on Data Augmentation and Loss Functions for Deep Learning Based Speech-Background Separation

*Hakan Erdogan and Takuya Yoshioka*

Microsoft AI and Research, One Microsoft Way, Redmond, WA, USA

**Thu-O-2-2-6, Time: 16:10-16:30**

A successful deep learning-based method for separation of a speech signal from an interfering background audio signal is based on neural network prediction of time-frequency masks which multiply noisy signal's short-time Fourier transform (STFT) to yield the STFT of an enhanced signal. In this paper, we investigate training strategies for mask-prediction-based speech-background separation systems. First, we examine the impact of mixing speech and noise files on the fly during training, which enables models to be trained on virtually infinite amount of data. We also investigate the effect of using a novel signal-to-noise ratio related loss function, instead of mean-squared error which is prone to scaling differences among utterances. We evaluate bi-directional long-short term memory (BLSTM) networks as well as a combination of convolutional and BLSTM (CNN+BLSTM) networks for mask prediction and compare performances of real and complex-valued

**Notes**

mask prediction. Data-augmented training combined with a novel loss function yields significant improvements in signal to distortion ratio (SDR) and perceptual evaluation of speech quality (PESQ) as compared to the best published result on CHiME-2 medium vocabulary data set when using a CNN+BLSTM network.

# Thu-O-2-3: Multimodal Systems

Hall 2; 14:30-16:30; Thursday, 6 September, 2018

Chairs: Björn Hoffmeister and Dinesh Jayagopi

## Annotator Trustability-based Cooperative Learning Solutions for Intelligent Audio Analysis

*Simone Hantke[1,2], Christoph Stemp[1,3] and Björn Schuller[1,4]*

[1]Chair for Embedded Intelligence on Health Care and Wellbeing, University of Augsburg, Germany

[2]Machine Intelligence & Signal Processing Group, Technische Universität München, Germany

[3]Chair of Complex & Intelligent Systems, University of Passau, Germany

[4]GLAM – Group on Language, Audio & Music, Imperial College London, UK

**Thu-O-2-3-1, Time: 14:30-14:50**

A broad range of artificially intelligent applications are nowadays available resulting in a need for masses of labelled data for the underlying machine learning models. This annotated data, however, is scarce and expensive to obtain from expert-like annotators. Crowdsourcing has been shown as a viable alternative, but it has to be carried out with adequate quality control to obtain reliable labels. Whilst crowdsourcing allows for the rapid collection of large-scale annotations, another technique called Cooperative Learning, aims at reducing the overall annotation costs, by learning to select only the most important instances for manual annotation. In this regard, we investigate the advantages of this approach and combine crowdsourcing with different iterative cooperative learning paradigms for audio data annotation, incorporating an annotator trustability score to reduce the labelling effort needed and, at the same time, to achieve better classification results. Key experimental results on an emotion recognition task show a considerable relative annotation reduction compared to a 'non-intelligent' approach of up to 85.3%. Moreover, the proposed trustability-based methods reach an unweighted average recall of 74.8%, while the baseline approach peaks at 61.2%. Therefore, the proposed trustability-based approaches efficiently reduce the manual annotation load, as well as improving the model.

## Semi-supervised Cross-domain Visual Feature Learning for Audio-Visual Broadcast Speech Transcription

*Rongfeng Su[1,2,3], Xunying Liu[1,3] and Lan Wang[1,3]*

[1]CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

[2]Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences

[3]The Chinese University of Hong Kong, Hong Kong, China

**Thu-O-2-3-2, Time: 14:50-15:10**

Visual information can be incorporated into automatic speech recognition (ASR) systems to improve their robustness in adverse acoustic conditions. Conventional audio-visual speech recognition (AVSR) systems require highly specialized audio-visual (AV) data in both system training and evaluation. For many real-world speech recognition applications, only audio information is available. This presents a major challenge to a wider application of AVSR systems. In order to address this challenge, this paper proposes a semi-supervised visual feature learning approach for developing AVSR systems on a DARPA GALE Mandarin broadcast transcription task. Audio to visual feature inversion long short-term memory neural networks (LSTMs) were initially constructed using limited amounts of out of domain AV data. The acoustic features domain mismatch against the broadcast data was further reduced using multi-level domain adaptive deep networks. Visual features were then automatically generated from the broadcast speech audio and used in both AVSR system training and testing time. Experimental results suggest a CNN based AVSR system using the proposed semi-supervised cross-domain audio-to-visual feature generation technique outperformed the baseline audio only CNN ASR system by an average CER reduction of 6.8% relative. In particular, on the most difficult Phoenix TV subset, a CER reduction of 1.32% absolute (8.34% relative) was obtained.

## Deep Lip Reading: A Comparison of Models and an Online Application

*Triantafyllos Afouras, Joon Son Chung and Andrew Zisserman*

Visual Geometry Group, Department of Engineering Science, University of Oxford, UK

**Thu-O-2-3-3, Time: 15:10-15:30**

The goal of this paper is to develop state-of-the-art models for lip reading -visual speech recognition. We develop three architectures and compare their accuracy and training times: (i) a recurrent model using LSTMs; (ii) a fully convolutional model; and (iii) the recently proposed transformer model. The recurrent and fully convolutional models are trained with a Connectionist Temporal Classification loss and use an explicit language model for decoding, the transformer is a sequence-to-sequence model. Our best performing model improves the state-of-the-art word error rate on the challenging BBC-Oxford Lip Reading Sentences 2 (LRS2) benchmark dataset by over 20 percent. As a further contribution we investigate the fully convolutional model when used for online (real time) lip reading of continuous speech and show that it achieves high performance with low latency.

## Iterative Learning of Speech Recognition Models for Air Traffic Control

*Ajay Srinivasamurthy[1], Petr Motlicek[1], Mittul Singh[2], Youssef Oualil[2], Matthias Kleinert[3], Heiko Ehr[3] and Hartmut Helmke[3]*

[1]Idiap Research Institute, Martigny, Switzerland

[2]Saarland Informatics Campus, Saarbrücken, Germany

[3]German Aerospace Center, Braunschweig, Germany

**Thu-O-2-3-4, Time: 15:30-15:50**

Automatic Speech Recognition (ASR) has recently proved to be a useful tool to reduce the workload of air traffic controllers leading to significant gains in operational efficiency. Air Traffic Control (ATC) systems in operation rooms around the world generate large amounts of untranscribed speech and radar data each day, which can be utilized to build and improve ASR models. In this paper, we propose an iterative approach that utilizes increasing amounts of untranscribed data to incrementally build the necessary ASR models for an ATC operational area. Our approach uses a semi-supervised learning framework to combine speech and radar data to iteratively update the acoustic model, language model and command prediction model (i.e. prediction of possible commands from radar data for a given air traffic situation) of an ASR system. Starting with seed models built with a limited amount of manually transcribed data, we simulate an operational scenario to adapt and improve the models through semi-supervised learning. Experiments on two independent ATC areas (Vienna and Prague) demonstrate the utility of our proposed methodology that can scale to operational environments with minimal manual effort for learning and adaptation.

## Speaker Adaptive Audio-Visual Fusion for the Open-Vocabulary Section of AVICAR

*Leda Sari[1,2], Mark Hasegawa-Johnson[1,2], Kumaran S[3], Georg Stemmer[3] and Krishnakumar N Nair[3]*

[1]Department of Electrical and Computer Engineering, University of

**Notes**

Illinois Urbana-Champaign, US
[2]Beckman Institute, University of Illinois Urbana-Champaign, US
[3]Intel Corporation

**Thu-O-2-3-5, Time: 15:50-16:10**

This experimental study establishes the first audio-visual speech recognition baseline for the TIMIT sentence portion of the AVICAR dataset, a dataset recorded in a real, noisy car environment. We use an automatic speech recognizer trained on a larger dataset to generate an audio-only recognition baseline for AVICAR. We utilize the forced alignment of the audio modality of AVICAR to get training targets for the convolutional neural network based visual front end. Based on our observation that there is a great amount of variation between visual features of different speakers, we apply feature space maximum likelihood linear regression (fMMLR) based speaker adaptation to the visual features. We find that the quality of fMLLR is sensitive to the quality of the alignment probabilities used to compute it; experimental tests compare the quality of fMLLR trained using audio-visual versus audio-only alignment probabilities. We report the first audio-visual results for TIMIT subset of AVICAR and show that the word error rate of the proposed audio-visual system is significantly better than that of the audio-only system.

### Multimodal Name Recognition in Live TV Subtitling

*Marek Hrúz[1], Aleš Pražák[1] and Michal Bušta[2]*
[1]University of West Bohemia, Faculty of Applied Sciences, NTIS - New Technologies for the Information Society, Univerzitní 8, 306 14 Plzeň, Czech Republic
[2]Czech Technical University, Faculty of Electrical Engineering, Center for Machine Perception, Department of Cybernetics, Karlovo namesti 13, 121 35 Prague, Czech Republic

**Thu-O-2-3-6, Time: 16:10-16:30**

In this paper, we present a method of combining a visual text reader with a system of automatic speech recognition to suppress errors when encountering out-of-vocabulary words - specifically names. The visual text reader outputs detected words that are mapped into a large list of names via the Levenshtein distance. The detected names are inserted into the class-based language model on the fly which improves recognition results. To demonstrate the effect on the real speech recognition task we use data from sports TV broadcasting where a lot of names are present in both the audio and video streams. We superseded manual vocabulary editing in live TV subtitling through respeaking by an automated online process. Further, we show that automatically adding the names to the recognition vocabulary online and with forgetting lowers the WER relatively by 39% in comparison with the case when names of all sportsmen are added to the vocabulary beforehand and by 15% when only the relevant names are added beforehand.

## Thu-O-2-4: Coding

MR G.01-G.02; 14:30-16:30; Thursday, 6 September, 2018
Chairs: Isabel Trancoso and Gernot Kubin

### Dithered Quantization for Frequency-Domain Speech and Audio Coding

*Tom Bäckström[1], Johannes Fischer[2] and Sneha Das[1]*
[1]Aalto University, Department of Signal Processing and Acoustics, Finland
[2]International Audio Laboratories Erlangen, Friedrich-Alexander University (FAU), Germany

**Thu-O-2-4-1, Time: 14:30-14:50**

A common issue in coding speech and audio in the frequency domain, which appears with decreasing bitrate, is that quantization levels become increasingly sparse. With low accuracy, high-frequency components are typically quantized to zero, which leads to a muffled output signal and musical noise. Band-width extension and noise-filling methods attempt to treat the problem by inserting noise of similar energy as the original signal, at the cost of low signal to noise ratio. Dithering methods however provide an alternative approach, where both accuracy and energy are retained. We propose a hybrid coding approach where low-energy samples are quantized using dithering, instead of the conventional uniform quantizer. For dithering, we apply 1 bit quantization in a randomized sub-space. We further show that the output energy can be adjusted to the desired level using a scaling parameter. Objective measurements and listening tests demonstrate the advantages of the proposed methods.

### Postfiltering with Complex Spectral Correlations for Speech and Audio Coding

*Sneha Das and Tom Bäckström*
Department of Signal Processing and Acoustics, Aalto University, Finland

**Thu-O-2-4-2, Time: 14:50-15:10**

State-of-the-art speech codecs achieve a good compromise between quality, bitrate and complexity. However, retaining performance outside the target bitrate range remains challenging. To improve performance, many codecs use pre- and post-filtering techniques to reduce the perceptual effect of quantization-noise. In this paper, we propose a postfiltering method to attenuate quantization noise which uses the complex spectral correlations of speech signals. Since conventional speech codecs cannot transmit information with temporal dependencies as transmission errors could result in severe error propagation, we model the correlation offline and employ them at the decoder, hence removing the need to transmit any side information. Objective evaluation indicates an average 4 dB improvement in the perceptual SNR of signals using the context-based post-filter, with respect to the noisy signal and an average 2 dB improvement relative to the conventional Wiener filter. These results are confirmed by an improvement of up to 30 MUSHRA points in a subjective listening test.

### Postfiltering Using Log-Magnitude Spectrum for Speech and Audio Coding

*Sneha Das and Tom Bäckström*
Department of Signal Processing and Acoustics, Aalto University, Finland

**Thu-O-2-4-3, Time: 15:10-15:30**

Advanced coding algorithms yield high quality signals with good coding efficiency within their target bit-rate ranges, but their performance suffer outside the target range. At lower bitrates, the degradation in performance is because the decoded signals are sparse, which gives a perceptually muffled and distorted characteristic to the signal. Standard codecs reduce such distortions by applying noise filling and post-filtering methods. In this paper, we propose a post-processing method based on modeling the inherent time-frequency correlation in the log-magnitude spectrum. The goal is to improve the perceptual SNR of the decoded signals and, to reduce the distortions caused by signal sparsity. Objective measures show an average improvement of 1.5 dB for input perceptual SNR in range 4 to 18 dB. The improvement is especially prominent in components which had been quantized to zero.

### Temporal Noise Shaping with Companding

*Arijit Biswas[1], Per Hedelin[2], Lars Villemoes[2] and Vinay Melkote*
[1]Dolby Germany GmbH, Nürnberg, Germany
[2]DolbySweden AB, Stockholm, Sweden

**Thu-O-2-4-4, Time: 15:30-15:50**

Audio codecs are typically transform-domain based and efficiently code stationary musical signals but they struggle with speech and signals with dense transients such as applause. The temporal noise shaping (TNS) tool standardized in HE-AAC alleviates the issue of noise unmasking in these troublesome cases via signal-adaptive filtering of the transform domain quantization noise, albeit at the cost of significant additional side information in the bitstream. We present a novel alternative referred to as

**Notes**

companding that involves QMF domain pre- and post-processing around the core transform-domain coding system: prior to transform encoding, the dynamic range of the signal is reduced locally within a QMF time slot and restored again post decoding, which naturally shapes the coding noise temporally. A primary advantage is that the companding function is fixed and hence enables signal-adaptive noise shaping with just 1-2 bits of side-information per frame. Subjective tests illustrate that the proposed tool improves the quality of hard-to-code applause excerpts compared to TNS while achieving comparable performance on speech signals. The coding tool described in this paper is part of the Dolby AC-4 audio coding system standardized by ETSI and included in ATSC 3.0.

### Multi-frame Quantization of LSF Parameters Using a Deep Autoencoder and Pyramid Vector Quantizer

*Yaxing Li[1], Eshete Derb Emiru[1], Shengwu Xiong[1], Anna Zhu[1], Pengfei Duan[1] and Yichang Li[2]*

[1]School of Computer Science and Technology, Wuhan University of Technology, China

[2]School of Computer Science and Engineering, Nanjing University of Science & Technology, China

**Thu-O-2-4-5, Time: 15:50-16:10**

This paper presents a multi-frame quantization of line spectral frequency (LSF) parameters using a deep autoencoder (DAE) and pyramid vector quantizer (PVQ). The object is to provide sophisticated LSF quantization for the ultra-low bit rate speech coders with moderate delay. For the compression and de-correlation of multiple LSF frames, a DAE possessing linear coder-layer units with Gaussian noise is used. The DAE demonstrates a high degree of modelling flexibility for multiple LSF frames. To quantize the coder-layer vector effectively, a PVQ is considered. Comparing the discrete cosine model (DCM), the DAE-based compression shows better modelling accuracy of multi-frame LSF parameters and possesses an advantage in that the coder-layer dimensions could be any value. The compressed coder-layer dimensions of the DAE govern the trade-off between the modelling distortion and the coder-layer quantization distortion. The experimental results show that the proposed algorithm with determined optimal coder-layer dimension outperforms the DCM-based multi-frame LSF quantization approach in terms of spectral distortion (SD) performance and robustness across different speech segments.

### Multi-frame Coding of LSF Parameters Using Block-Constrained Trellis Coded Vector Quantization

*Yaxing Li[1], Shan Xu[1], Shengwu Xiong[1], Anna Zhu[1], Pengfei Duan[1] and Yueming Ding[2]*

[1]School of Computer Science and Technology, Wuhan University of Technology, China

[2]School of Computer Science and Engineering, Tianjin University of Technology, China

**Thu-O-2-4-6, Time: 16:10-16:30**

In this paper, the predictive block-constrained trellis coded vector quantization (BC-TCVQ) schemes are developed for quantization of multiple frames line spectral frequency (LSF) parameters. The consecutive LSF frames are interleaved to subvectors for trellis modeling. The predictive BC-TCVQ systems are then designed to encode multi-frame LSF parameters. The performance evaluation of proposed schemes is compared with the single-frame LSF encoding methods using multi-stage vector quantization (MSVQ) and predictive block-constrained trellis coded quantization (BC-TCQ), demonstrating significant reduction of bit rate for transparent coding. The developed multi-frame LSF quantization schemes show satisfactory performance even at very low encoding rate and thus can be efficiently applied to the speech coders with moderate delay.

## Thu-P-2-1: Speaker Verification Using Neural Network Methods II

Hall 4-6: Poster-1; 14:30-16:30; Thursday, 6 September, 2018

Chair: Koichi Shinoda

### Training Utterance-level Embedding Networks for Speaker Identification and Verification

*Heewoong Park[1], Sukhyun Cho[1], Kyubyong Park[2], Namju Kim[2] and Jonghun Park[1]*

[1]Dept. of Industrial Engineering & Center for Superintelligence, Seoul National University, Republic of Korea

[2]Kakao Brain Corporation, Republic of Korea

**Thu-P-2-1-1, Time: 14:30-16:30**

Encoding speaker-specific characteristics from speech signals into fixed length vectors is a key component of speaker identification and verification systems. This paper presents a deep neural network architecture for speaker embedding models where similarity in embedded utterance vectors explicitly approximates the similarity in vocal patterns of speakers. The proposed architecture contains an additional speaker embedding lookup table to compute loss based on embedding similarities. Furthermore, we propose a new feature sampling method for data augmentation. Experimentation based on two databases demonstrates that our model is more effective at speaker identification and verification when compared to a fully connected classifier and an end-to-end verification model.

### Analysis of Complementary Information Sources in the Speaker Embeddings Framework

*Mahesh Kumar Nandwana, Mitchell McLaren, Diego Castan, Julien van Hout and Aaron Lawson*

Speech Technology and Research Laboratory, SRI International, Menlo Park, California, USA

**Thu-P-2-1-2, Time: 14:30-16:30**

Deep neural network (DNN)-based speaker embeddings have resulted in new, state-of-the-art text-independent speaker recognition technology. However, very limited effort has been made to understand DNN speaker embeddings. In this study, our aim is analyzing the behavior of the speaker recognition systems based on speaker embeddings toward different front-end features, including the standard Mel frequency cepstral coefficients (MFCC), as well as power normalized cepstral coefficients (PNCC) and perceptual linear prediction (PLP). Using a speaker recognition system based on DNN speaker embeddings and probabilistic linear discriminant analysis (PLDA), we compared different approaches to leveraging complementary information using score-, embeddings- and feature-level combination. We report our results for Speakers in the Wild (SITW) and NIST SRE 2016 datasets. We found that first and second embeddings layers are complementary in nature. By applying score and embedding-level fusion we demonstrate relative improvements in equal error rate of 17% on NIST SRE 2016 and 10% on SITW over the baseline system.

### Self-Attentive Speaker Embeddings for Text-Independent Speaker Verification

*Yingke Zhu[1], Tom Ko[2], David Snyder[3], Brian Mak[1] and Daniel Povey[3]*

[1]Department of Computer Science & Engineering, The Hong Kong University of Science & Technology

[2]Huawei Noahs Ark Research Lab, Hong Kong, China

[3]Center for Language and Speech Processing & Human Language Technology Center of Excellence, The Johns Hopkins University, USA

**Thu-P-2-1-3, Time: 14:30-16:30**

**Notes**

This paper introduces a new method to extract speaker embeddings from a deep neural network (DNN) for text-independent speaker verification. Usually, speaker embeddings are extracted from a speaker-classification DNN that averages the hidden vectors over the frames of a speaker; the hidden vectors produced from all the frames are assumed to be equally important. We relax this assumption and compute the speaker embedding as a weighted average of a speaker's frame-level hidden vectors and their weights are automatically determined by a self-attention mechanism. The effect of multiple attention heads are also investigated to capture different aspects of a speaker's input speech. Finally, a PLDA classifier is used to compare pairs of embeddings. The proposed self-attentive speaker embedding system is compared with a strong DNN embedding baseline on NIST SRE 2016. We find that the self-attentive embeddings achieve superior performance. Moreover, the improvement produced by the self-attentive speaker embeddings is consistent with both short and long testing utterances.

## An Improved Deep Embedding Learning Method for Short Duration Speaker Verification

*Zhifu Gao[1], Yan Song[1], Ian McLoughlin[2], Wu Guo[1] and Lirong Dai[1]*

[1]National Engineering Laboratory of Speech and Language Information Processing, University of Science and Technology of China, Hefei, China

[2]School of Computing, University of Kent, Medway, UK

**Thu-P-2-1-4, Time: 14:30-16:30**

This paper presents an improved deep embedding learning method based on convolutional neural network (CNN) for short-duration speaker verification (SV). Existing deep learning-based SV methods generally extract frontend embeddings from a feed-forward deep neural network, in which the long-term speaker characteristics are captured via a pooling operation over the input speech. The extracted embeddings are then scored via a backend model, such as Probabilistic Linear Discriminative Analysis (PLDA). Two improvements are proposed for frontend embedding learning based on the CNN structure: (1) Motivated by the WaveNet for speech synthesis, dilated filters are designed to achieve a tradeoff between computational efficiency and receptive-filter size; and (2) A novel cross-convolutional-layer pooling method is exploited to capture 1st-order statistics for modelling long-term speaker characteristics. Specifically, the activations of one convolutional layer are aggregated with the guidance of the feature maps from the successive layer. To evaluate the effectiveness of our proposed methods, extensive experiments are conducted on the modified female portion of NIST SRE 2010 evaluations, with conditions ranging from 10s-10s to 5s-4s. Excellent performance has been achieved on each evaluation condition, significantly outperforming existing SV systems using i-vector and d-vector embeddings.

## Avoiding Speaker Overfitting in End-to-End DNNs Using Raw Waveform for Text-Independent Speaker Verification

*Jee-weon Jung, Hee-soo Heo, IL-ho Yang, Hye-jin Shim and Ha-jin Yu*

School of Computer Science, University of Seoul, South Korea

**Thu-P-2-1-5, Time: 14:30-16:30**

In this research, we propose a novel raw waveform end-to-end DNNs for text-independent speaker verification. For speaker verification, many studies utilize the speaker embedding scheme, which trains deep neural networks as speaker identifiers to extract speaker features. However, this scheme has an intrinsic limitation in which the speaker feature, trained to classify only known speakers, is required to represent the identity of unknown speakers. Owing to this mismatch, speaker embedding systems tend to well generalize towards unseen utterances from known speakers, but are overfitted to known speakers. This phenomenon is referred to as speaker overfitting. In this paper, we investigated regularization techniques, a multi-step training scheme and a residual connection with pooling layers in the perspective of mitigating speaker overfitting which lead to considerable performance improvements. Technique effectiveness is evaluated using the VoxCeleb dataset, which comprises over 1,200 speakers from various uncontrolled environments. To the best of our knowledge, we are the first to verify the success of end-to-end DNNs directly using raw waveforms in text-independent scenario. It shows an equal error rate of 7.4%, which is lower than i-vector/probabilistic linear discriminant analysis and end-to-end DNNs that use spectrograms.

## Deeply Fused Speaker Embeddings for Text-Independent Speaker Verification

*Gautam Bhattacharya[1,2], Md Jahangir Alam[2], Vishwa Gupta[2] and Patrick Kenny[2]*

[1]McGill University

[2]Computer Research Institute of Montreal (CRIM)

**Thu-P-2-1-6, Time: 14:30-16:30**

Recently there has been a surge of interest is learning speaker embeddings using deep neural networks. These models ingest time-frequency representations of speech and can be trained to discriminate between a known set speakers. While embeddings learned in this way perform well, they typically require a large number of training data points for learning. In this work we propose deeply fused speaker embeddings - speaker representations that combine neural speaker embeddings with i-vectors. We show that by combining the two speaker representations we are able to learn robust speaker embeddings in a computationally efficient manner.

We compare several different fusion strategies and find that the resulting speaker embeddings show significantly different verification performance. To this end we propose a novel fusion approach that uses an attention model to combine i-vectors with neural speaker embeddings. Our best performing embedding achieves an error rate of 3.17% using a simple cosine distance classifier. Combining our embeddings with a powerful Joint Bayesian classifier, we are able to further improve the performance of our speaker embeddings to 2.22%, which gave a 7.8% relative improvement over the baseline i-vector system.

## Employing Phonetic Information in DNN Speaker Embeddings to Improve Speaker Recognition Performance

*Md Hafizur Rahman[1], Ivan Himawan[1], Mitchell McLaren[2], Clinton Fookes[1] and Sridha Sridharan[1]*

[1]Speech and Audio Research Laboratory, Queensland University of Technology, Australia

[2]Speech Technology and Research Laboratory, SRI International, California, USA

**Thu-P-2-1-7, Time: 14:30-16:30**

The recent speaker embeddings framework has been shown to provide excellent performance on the task of text-independent speaker recognition. The framework is based on a deep neural network (DNN) trained to directly discriminate between speakers from traditional acoustic features such as Mel frequency cepstral coefficients. Prior studies on speaker recognition have found that phonetic information is valuable in the task of speaker identification, with systems being based on either bottleneck features (BFs) or tied-triphone state posteriors from a DNN trained for the task of speech recognition. In this paper, we analyze the role of phonetic BFs for DNN embeddings and explore methods to enhance the BFs further. Experimental results show that exploiting phonetic information encoded in BFs is very valuable for DNN speaker embeddings. Enriching the BFs using a cascaded DNN multi-task architecture is also shown to provide further improvements to the speaker embedding system.

## End-to-End Text-dependent Speaker Verification Using Novel Distance Measures

*Subhadeep Dey[1,2], Srikanth Madikeri[1] and Petr Motlicek[1]*

[1]Idiap Research Institute, Martigny, Switzerland

[2]Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

**Thu-P-2-1-8, Time: 14:30-16:30**

This paper explores novel ideas in building end-to-end deep neural network (DNN) based text-dependent speaker verification (SV) system. The baseline approach consists of mapping a variable length speech segment to a fixed dimensional speaker vector by estimating the mean of hidden

**Notes**

representations in DNN structure. The distance between two utterances is obtained by computing L2 norm between the vectors. This approach performs worse than the conventional Gaussian Mixture Model-Universal Background Model (GMM-UBM) based system in publicly available corpora. We believe that poor performance is due to the employed averaging operation, which may not capture the phonetic information of an utterance. Past studies indicate that techniques exploiting phonetic information in addition to speaker is beneficial for this task. This paper therefore proposes to incorporate content information of the speech signal by computing distance function with linguistic units co-occuring between enrollment and test data. The whole network is optimized by employing a triplet-loss objective in an end-to-end fashion to output SV scores. Experiments on the RSR2015 dataset show that the proposed approach outperforms GMM-UBM system by 48% and 36% relative equal error rate for fixed-phrase and Random-digit conditions respectively.

## Robust Speaker Clustering using Mixtures of von Mises-Fisher Distributions for Naturalistic Audio Streams

*Harishchandra Dubey, Abhijeet Sangwan and John H.L. Hansen*

Robust Speech Technologies Lab, Center for Robust Speech Systems, The University of Texas at Dallas, Richardson, TX-75080, USA

**Thu-P-2-1-9, Time: 14:30-16:30**

Speaker Diarization (i.e. determining who spoke and when?) for multi-speaker naturalistic interactions such as Peer-Led Team Learning (PLTL) sessions is a challenging task. In this study, we propose robust speaker clustering based on mixture of multivariate von Mises-Fisher distributions. Our diarization pipeline has two stages: (i) ground-truth segmentation; (ii) proposed speaker clustering. The ground-truth speech activity information is used for extracting i-Vectors from each speechsegment. We post-process the i-Vectors with principal component analysis for dimension reduction followed by lengthnormalization. Normalized i-Vectors are high-dimensional unit vectors possessing discriminative directional characteristics. We model the normalized i-Vectors with a mixture model consisting of multivariate von Mises-Fisher distributions. K-means clustering with cosine distance is chosen as baseline approach. The evaluation data is derived from: (i) CRSS-PLTL corpus; and (ii) three-meetings subset of AMI corpus. The CRSSPLTL data contain audio recordings of PLTL sessions which is student-led STEM education paradigm. Proposed approach is consistently better than baseline leading to upto 44.48% and 53.68% relative improvements for PLTL and AMI corpus, respectively.

## Triplet Network with Attention for Speaker Diarization

*Huan Song[1], Megan Willi[2], Jayaraman J. Thiagarajan[3], Visar Berisha[1,2] and Andreas Spanias[1]*

[1]SenSIP Center, School of ECEE, Arizona State University, Tempe, AZ

[2]Department of Speech and Hearing Science, Arizona State University, Tempe, AZ

[3]Lawrence Livermore National Labs, 7000 East Avenue, Livermore, CA

**Thu-P-2-1-10, Time: 14:30-16:30**

We present our research on continuous speech recognition based on Surface Electromyography (EMG), where speech information is captured by electrodes attached to the speaker's face. This method allows speech processing without requiring that an acoustic signal is present; however, reattachment of the EMG electrodes causes subtle changes in the recorded signal, which degrades the recognition accuracy and thus poses a major challenge for practical application of the system. Based on the growing body of recent work in domain-adversarial training of neural networks, we present a system which adapts the neural network frontend of our recognizer to data from a new reln automatic speech processing systems, speaker diarization is a crucial front-end component to separate segments from different speakers. Inspired by the recent success of deep neural networks (DNNs) in semantic inferencing, triplet loss-based architectures have been successfully used for this problem. However, existing work utilizes conventional i-vectors as the input representation and builds simple fully connected networks for metric learning, thus not fully leveraging the modeling power of DNN architectures. This paper investigates the importance of learning effective representations from the sequences directly in metric learning pipelines for speaker diarization. More specifically, we propose to employ attention models to learn embeddings and the metric jointly in an end-to-end fashion. Experiments are conducted on the CALLHOME conversational speech corpus. The diarization results demonstrate that, besides providing a unified model, the proposed approach achieves improved performance when compared against existing approaches.cording session, without requiring supervised enrollment.

## I-vector Transformation Using Conditional Generative Adversarial Networks for Short Utterance Speaker Verification

*Jiacen Zhang, Nakamasa Inoue and Koichi Shinoda*

Tokyo Institute of Technology

**Thu-P-2-1-11, Time: 14:30-16:30**

I-vector based text-independent speaker verification (SV) systems often have poor performance with short utterances, as the biased phonetic distribution in a short utterance makes the extracted i-vector unreliable. This paper proposes an i-vector compensation method using a generative adversarial network (GAN), where its generator network is trained to generate a compensated i-vector from a short-utterance i-vector and its discriminator network is trained to determine whether an i-vector is generated by the generator or the one extracted from a long utterance. Additionally, we assign two other learning tasks to the GAN to stabilize its training and to make the generated i-vector more speaker-specific. Speaker verification experiments on the NIST SRE 2008 "10sec-10sec" condition show that after applying our method, the equal error rate reduced by 11.3% from the conventional i-vector and PLDA system.

## Analysis of Length Normalization in End-to-End Speaker Verification System

*Weicheng Cai[2], Jinkun Chen[2] and Ming Li[1]*

[1]Data Science Research Center, Duke Kunshan University, Kunshan, China

[2]School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, China

**Thu-P-2-1-12, Time: 14:30-16:30**

The classical i-vectors and the latest end-to-end deep speaker embeddings are the two representative categories of utterance-level representations in automatic speaker verification systems. Traditionally, once i-vectors or deep speaker embeddings are extracted, we rely on an extra length normalization step to normalize the representations into unit-length hyperspace before back-end modeling. In this paper, we explore how the neural network learns length-normalized deep speaker embeddings in an end-to-end manner. To this end, we add a length normalization layer followed by a scale layer before the output layer of the common classification network. We conducted experiments on the verification task of the Voxceleb1 dataset. The results show that integrating this simple step in the end-to-end training pipeline significantly boosts the performance of speaker verification. In the testing stage of our L2-normalized end-to-end system, a simple inner-product can achieve the state-of-the-art.

## Angular Softmax for Short-Duration Text-independent Speaker Verification

*Zili Huang, Shuai Wang and Kai Yu*

Key Lab of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering

SpeechLab, Department of Computer Science and Engineering

Brain Science and Technology Research Center

Shanghai Jiao Tong University, Shanghai, China

**Thu-P-2-1-13, Time: 14:30-16:30**

Recently, researchers propose to build deep learning based end-to-end speaker verification (SV) systems and achieve competitive results

**Notes**

compared with the standard i-vector approach. In addition to deep learning architectures, optimization metric, such as softmax loss or triplet loss, is important for extracting speaker embeddings which are discriminative and generalizable to unseen speakers. In this paper, angular softmax (A-softmax) loss is introduced to improve speaker embedding quality. It is investigated in two SV frameworks: a CNN based end-to-end SV framework and an i-vector SV framework where deep discriminant analysis is used for channel compensation. Experimental results on a short-duration text-independent speaker verification dataset generated from SRE reveal that A-softmax achieves significant performance improvement compared with other metrics in both frameworks.

### An End-to-End Text-Independent Speaker Identification System on Short Utterances

*Ruifang Ji[1,2], Xinyuan Cai[1] and Xu Bo[1]*

[1]Institute of Automation, Chinese Academy of Sciences, China

[2]University of Chinese Academy of Sciences, China

**Thu-P-2-1-14, Time: 14:30-16:30**

In the field of speaker recognition, text-independent speaker identification on short utterances is still a challenging task, since it is rather tough to extract a robust and dicriminative speaker feature in short duration condition. This paper explores an end-to-end speaker identification system, which maps utterances to a speaker identity subspace where the similarity of speakers can be measured by Euclidean distance. To be specific, we apply GRU architectures to extract utterance-level feature. Then it is assumed that one's various utterances can be viewed as transformations of a single object in an ideal speaker identity subspace. Based on this assumption, the ResCNN architecture is utilized to model the transformation and the whole system is jointly optimized by speaker identity subspace loss. Experimental results demonstrate the effectiveness of our proposed system and superiority over pervious methods. For example, the GRU learned feature reduces the equal error rate by 27.53% relatively and the speaker identity subspace loss further brings 7.22% relative reduction compared to softmax loss.

### MTGAN: Speaker Verification through Multitasking Triplet Generative Adversarial Networks

*Wenhao Ding and Liang He*

Department of Electronic Engineering, Tsinghua University, China

**Thu-P-2-1-15, Time: 14:30-16:30**

In this paper, we propose an enhanced triplet method that improves the encoding process of embeddings by jointly utilizing generative adversarial mechanism and multitasking optimization. We extend our triplet encoder with Generative Adversarial Networks (GANs) and softmax loss function. GAN is introduced for increasing the generality and diversity of samples, while softmax is for reinforcing features about speakers. For simplification, we term our method Multitasking Triplet Generative Adversarial Networks (MTGAN). Experiment on short utterances demonstrates that MTGAN reduces the verification equal error rate (EER) by 67% (relatively) and 32% (relatively) over conventional i-vector method and state-of-the-art triplet loss method respectively. This effectively indicates that MTGAN outperforms triplet methods in the aspect of expressing the high-level feature of speaker information.

## Thu-P-2-2: Emotion Recognition and Analysis

Hall 4-6: Poster-2; 14:30-16:30; Thursday, 6 September, 2018

Chair: Chi-Chun Lee

### Categorical vs Dimensional Perception of Italian Emotional Speech

*Emilia Parada-Cabaleiro[1], Giovanni Costantini[2], Anton Batliner[1], Alice Baird[1] and Björn Schuller[1,3]*

[1]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany

[2]Department of Electronic Engineering, University of Rome Tor Vergata, Italy

[3]GLAM – Group on Language, Audio & Music, Imperial College London, UK

**Thu-P-2-2-1, Time: 14:30-16:30**

Culture and measurement strategies are influential factors when evaluating the perception of emotion in speech. However, multilingual databases suitable for such a study are missing and there is no agreement on the most suitable emotional model. To address this gap, we present EmoFilm, a new multilingual emotional speech corpus, consisting of 1115 English, Spanish and Italian emotional utterances extracted from 43 films and 207 speakers. We have performed a within-culture categorical vs dimensional perceptual evaluation, employing 225 native Italian listeners, who evaluated the Italian section of the database with the emotional states of anger, sadness, happiness, fear and contempt. The aim of this study is to assess whether the emotional model (categorical or dimensional), taken as reference for measurement, influences a listener's perception of emotional speech and—to what extent—both models are complementary or not. We show that the measurement strategy chosen does influence a listener's response, especially for some emotions, e.g. contempt. The confusion patterns typical of a categorical evaluation are not always mirrored by the dimensional assessment.

### A Three-Layer Emotion Perception Model for Valence and Arousal-Based Detection from Multilingual Speech

*Xingfeng Li and Masato Akagi*

Japan Advanced Institute of Science and Technology

**Thu-P-2-2-2, Time: 14:30-16:30**

Automated emotion detection from speech has recently shifted from monolingual to multilingual tasks for human-like interaction in real-life where a system can handle more than a single input language. However, most work on monolingual emotion detection is difficult to generalize in multiple languages, because the optimal feature sets of the work differ from one language to another. Our study proposes a framework to design, implement and validate an emotion detection system using multiple corpora. A continuous dimensional space of valence and arousal is first used to describe the emotions. A three-layer model incorporated with fuzzy inference systems is then used to estimate two dimensions. Speech features derived from prosodic, spectral and glottal waveform are examined and selected to capture emotional cues. The results of this new system outperformed the existing state-of-the-art system by yielding a smaller mean absolute error and higher correlation between estimates and human evaluators. Moreover, results for speaker independent validation are comparable to human evaluators.

### Cross-lingual Speech Emotion Recognition through Factor Analysis

*Brecht Desplanques and Kris Demuynck*

Ghent University - imec, IDLab, Department of Electronics and Information Systems, Belgium

**Thu-P-2-2-3, Time: 14:30-16:30**

Conventional speech emotion recognition based on the extraction of high level descriptors emerging from low level descriptors seldom delivers promising results in cross-corpus experiments. Therefore it might not perform well in real-life applications. Factor analysis, proven in the fields of language identification and speaker verification, could clear a path towards more robust emotion recognition. This paper proposes an iVector-based approach operating on acoustic MFCC features with a separate modeling of the speaker and emotion variabilities respectively. The speech analysis extracts two fixed-length low-dimensional feature vectors corresponding to the two mentioned sources of variation. To model the speaker-related nuisance variability speaker factors are extracted using an eigenvoice matrix. After compensating for this speaker variability in the supervector space, the emotion factors (one per targeted emotion) are extracted using an emotion variability matrix. The emotion factors are then fed to a basic

**Notes**

emotion classifier. Leave-one-speaker-out cross-validation on the Berlin Database of Emotional Speech EMO-DB (German) and IEMOCAP (English) datasets lead to results that are competitive with the current state-of-the-art. Cross-lingual experiments demonstrate the excellent robustness of the method: the classification accuracies degrade less than 15% relative when emotion models are trained on one corpus and tested on the other.

## Modeling Self-Reported and Observed Affect from Speech

*Jian Cheng[1], Jared Bernstein[1], Elizabeth Rosenfeld[1], Peter W. Foltz[2], Alex S. Cohen[3], Terje B. Holmlund[4] and Brita Elvevåg[4,5]*

[1]Analytic Measures Inc., Palo Alto, CA, U.S.A.

[2]Institute of Cognitive Science, University of Colorado & Pearson, CO, U.S.A.

[3]Department of Psychology, Louisiana State University, LA, U.S.A.

[4]Department of Clinical Medicine, University of Tromsø – the Arctic University of Norway, Norway

[5]Norwegian Centre for eHealth Research, University Hospital of North Norway, Tromsø, Norway

**Thu-P-2-2-4, Time: 14:30-16:30**

Listeners hear joy/sadness and engagement/indifference in speech, even when linguistic content is neutral. We measured audible emotion in spontaneous speech and related it to self-reports of affect in response to questions, such as "Are you hopeful?" Spontaneous speech and self-reports were both collected in sessions with an interactive mobile app and used to compare three affect measurements: self-report; listener judgement; and machine score. The app adapted a widely-used measure of affective state to collect self-reported positive/negative affect and it engaged users in spoken interactions. Each session elicited 11 affect self-reports and captured about 9 minutes of speech; with 118 sessions by psychiatric patients and 227 sessions by non-clinical users. Speech recordings were evaluated for arousal and valence by clinical experts and by computer analysis of acoustic (non-linguistic) variables. The affect self-reports were reasonably reliable ($\alpha$ 0.73 to 0.84). Combined affect ratings from clinical-expert listeners produced reliable ratings per session ($\alpha$ 0.75 to 0.99) and acoustic feature analysis matched the expert ratings fairly well ($0.36 < r < 0.72$, mean 0.57), but neither human nor computed scores had high correlation with standard affect self-reported values. These results are discussed in relation to common methods of developing and evaluating affect analysis.

## Stochastic Shake-Shake Regularization for Affective Learning from Speech

*Che-Wei Huang and Shrikanth Narayanan*

Signal Analysis and Interpretation Laboratory (SAIL), University of Southern California

**Thu-P-2-2-5, Time: 14:30-16:30**

We propose stochastic Shake-Shake regularization based on multi-branch residual architectures to mitigate over-fitting in affective learning from speech. Inspired by recent Shake-Shake [1] and ShakeDrop [2] regularization techniques, we introduce negative scaling into the Shake-Shake regularization algorithm while still maintain a consistent stochastic convex combination of branches to encourage diversity among branches whether they are scaled by positive or negative coefficients. In addition, we also employ the idea of stochastic depth to randomly relax the shaking mechanism during training as a method to control the strength of regularization.

Through experiments on speech emotion recognition with various levels of regularization strength, we discover that the shaking mechanism alone contributes much more to constraining the optimization of network parameters than to boosting the generalization power. However, stochastically relaxing the shaking regularization serves to conveniently strike a balance between them. With a flexible configuration of hybrid layers, promising experimental results demonstrate a higher unweighted accuracy and a smaller gap between training and validation, i.e. reduced over-fitting and shed light on the future direction for pattern recognition tasks with low resource.

**Notes**

## Investigating Speech Enhancement and Perceptual Quality for Speech Emotion Recognition

*Anderson R. Avila[1,2], Md Jahangir Alam[2], Douglas O'Shaughnessy[1] and Tiago Falk[1]*

[1]INRS-EMT, University of Quebec, Canada

[2]Computer Research Institute of Montreal (CRIM)

**Thu-P-2-2-6, Time: 14:30-16:30**

In this study, the performance of two enhancement algorithms is investigated in terms of perceptual quality as well as in respect to their impact on speech emotion recognition (SER). The SER system adopted is based on the same benchmark system provided for the AVEC Challenge 2016. The three objective measures adopted are the speech-to-reverberation modulation energy ratio (SRMR), the perceptual evaluation of speech quality (PESQ) and the perceptual objective listening quality assessment (POLQA). Evaluations are conducted on speech files from the RECOLA dataset, which provides spontaneous interactions in French of 27 subjects. Clean speech files are corrupted with different levels of background noise and reverberation. Results show that applying enhancement prior to the SER task can improve SER performance in more degraded scenarios. We also show that quality measures can be an important asset as indicator of enhancement algorithms performance towards SER, with SRMR and POLQA providing the most reliable results.

## Demonstrating and Modelling Systematic Time-varying Annotator Disagreement in Continuous Emotion Annotation

*Mia Atcheson, Vidhyasaharan Sethu and Julien Epps*

University of New South Wales

**Thu-P-2-2-7, Time: 14:30-16:30**

Continuous emotion recognition (CER) is the task of determining the emotional content of speech from audio or multimedia recordings. Training targets for machine learning must be generated by human annotation, generally as a time series of emotional parameter values. In typical contemporary CER systems and challenges, the mean over a pool of annotators is taken to represent this ground truth, but is this an appropriate model for the emotional content of speech? Using the RECOLA dataset, the primary contribution of this research is to show that a correlation exists between the time-varying disagreement from independent groups of annotators. Because the groups are completely isolated except via the speech signal, this agreement-about-disagreement demonstrates that there is a component of annotator disagreement which arises systematically from the signal itself, which qualitatively implies that the perceived emotional content of speech can exhibit some degree of inherent ambiguity. Additionally, we show that these human annotations exhibit a degree of temporal smoothness. Neither of these characteristics is represented by the standard series-of-means ground-truth model, so we propose two alternative ground-truth models: a mean-variance model that incorporates ambiguity and a more general Gaussian process model that incorporates ambiguity and temporal smoothness in a well-defined probability distribution.

## Speech Emotion Recognition from Variable-Length Inputs with Triplet Loss Function

*Jian Huang[1,2], Ya Li[1], Jianhua Tao[1,2,3] and Zhen Lian[1,2]*

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China

[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

[3]CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China

**Thu-P-2-2-8, Time: 14:30-16:30**

Automatic emotion recognition is a crucial element on understanding human behavior and interaction. Prior works on speech emotion recognition focus on exploring various feature sets and models. Compared with these methods, we propose a triplet framework based on Long Short-Term Memory Neural Network (LSTM) for speech emotion recognition. The system learns

a mapping from acoustic features to discriminative embedding features, which are regarded as basis of testing with SVM. The proposed model is trained with triplet loss and supervised loss simultaneously. The triplet loss makes intra-class distance shorter and inter-class distance longer and supervised loss incorporates class label information. In view of variable-length inputs, we explore three different strategies to handle this problem and meanwhile make better use of temporal dynamic process information. Our experimental results on the Interactive Emotional Motion Capture (IEMOCAP) database reveal that the proposed methods are beneficial to performance improvement. We demonstrate promise of triplet framework for speech emotion recognition and present our analysis.

## Imbalance Learning-based Framework for Fear Recognition in the MediaEval Emotional Impact of Movies Task

*Xiaotong Zhang, Xingliang Cheng, Mingxing Xu and Thomas Fang Zheng*

Department of Computer Science and Technology, Center for Speech and Language Technologies, Research Institute of Information Technology

Tsinghua University, Beijing, China

**Thu-P-2-2-9, Time: 14:30-16:30**

Fear recognition, which aims at predicting whether a movie segment can induce fear or not, is a promising area in movie emotion recognition. Research in this area, however, has reached a bottleneck. Difficulties may partly result from the imbalanced database. In this paper, we propose an imbalance learning-based framework for movie fear recognition. A data rebalance module is adopted before classification. Several sampling methods, including the proposed softsampling and hardsampling which combine the merits of both undersampling and oversampling, are explored in this module. Experiments are conducted on the MediaEval 2017 Emotional Impact of Movies Task. Compared with the current state-of-the-art, we achieve an improvement of 8.94% on F1, proving the effectiveness of proposed framework.

## Emotion Recognition from Variable-Length Speech Segments Using Deep Learning on Spectrograms

*Xi Ma[1,3], Zhiyong Wu[1,2,3], Jia Jia[1,3], Mingxing Xu[1,3], Helen Meng[1,2] and Lianhong Cai[1,3]*

[1]Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China

[2]Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China

[3]Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

**Thu-P-2-2-10, Time: 14:30-16:30**

Deep neural network (DNN) dereverberation preprocessing has been shown to be a viable strategy for speech enhancement and increasing the accuracy of automatic speech recognition and automatic speaker verification. In this paper, an improved DNN technique based on convolutional neural networks is presented and compared to existing methods for speech enhancement and speaker verification in the presence of reverberation. This new technique is first shown to enhance speech quality as compared to other existing methods. Then, a more thorough set of experiments is presented that assesses cross-corpora speaker verification performance on data that contains real reverberation and noise. A discussion of the applicability In this work, an approach of emotion recognition is proposed for variable-length speech segments by applying deep neutral network to spectrograms directly. The spectrogram carries comprehensive para-lingual information that are useful for emotion recognition. We tried to extract such information from spectrograms and accomplish the emotion recognition task by combining Convolutional Neural Networks (CNNs) with Recurrent Neural Networks (RNNs). To handle the variable-length speech segments, we proposed a specially designed neural network structure that accepts variable-length speech sentences directly as input. Compared to the traditional methods that split the sentence into smaller fixed-length segments, our method can solve the problem of accuracy degradation introduced in the speech segmentation process. We evaluated the emotion recognition model on the IEMOCAP dataset over four emotions. Experimental results demonstrate that the proposed method outperforms the fixed-length neural network on both weighted accuracy (WA) and unweighted accuracy (UA).and generalizability of such techniques is given.

## Speech Emotion Recognition Using Spectrogram & Phoneme Embedding

*Promod Yenigalla, Abhay Kumar, Suraj Tripathi, Chirag Singh, Sibsambhu Kar and Jithendra Vepa*

Samsung R&D Institute India - Bangalore

**Thu-P-2-2-11, Time: 14:30-16:30**

This paper proposes a speech emotion recognition method based on phoneme sequence and spectrogram. Both phoneme sequence and spectrogram retain emotion contents of speech which is missed if the speech is converted into text. We performed various experiments with different kinds of deep neural networks with phoneme and spectrogram as inputs. Three of those network architectures are presented here that helped to achieve better accuracy when compared to the state-of-the-art methods on benchmark dataset. A phoneme and spectrogram combined CNN model proved to be most accurate in recognizing emotions on IEMOCAP data. We achieved more than 4% increase in overall accuracy and average class accuracy as compared to the existing state-of-the-art methods.

## On Enhancing Speech Emotion Recognition Using Generative Adversarial Networks

*Saurabh Sahu[1], Rahul Gupta[2] and Carol Espy-Wilson[1]*

[1]Speech Communication Laboratory, University of Maryland, College Park, MD, USA

[2]Amazon.com, USA

**Thu-P-2-2-12, Time: 14:30-16:30**

Generative Adversarial Networks (GANs) have gained a lot of attention from machine learning community due to their ability to learn and mimic an input data distribution. GANs consist of a discriminator and a generator working in tandem playing a min-max game to learn the complex underlying data distribution when fed with data-points sampled from a simpler distribution like Uniform or Gaussian. Once trained, they allow synthetic generation of examples sampled from the learned distribution. We investigate the application of GANs to generate synthetic feature vectors used for speech emotion recognition. Specifically, we investigate two set ups: (i) a vanilla GAN that learns the distribution of a lower dimensional representation of the actual higher dimensional feature vector and (ii) a conditional GAN that learns the distribution of the higher dimensional feature vectors conditioned on the labels or the emotional class to which it belongs. As a potential practical application of these synthetically generated samples, we measure any improvement in a classifier's performance when the synthetic data was used along with real for training it. We perform cross validation analyses followed by a cross-corpus study.

## Ladder Networks for Emotion Recognition: Using Unsupervised Auxiliary Tasks to Improve Predictions of Emotional Attributes

*Srinivas Parthasarathy and Carlos Busso*

Multimodal Signal Processing(MSP) lab, Department of Electrical and Computer Engineering The University of Texas at Dallas, Richardson TX 75080, USA

**Thu-P-2-2-13, Time: 14:30-16:30**

Recognizing emotions using few attribute dimensions such as arousal, valence and dominance provides the flexibility to effectively represent complex range of emotional behaviors. Conventional methods to learn these emotional descriptors primarily focus on separate models to recognize each of these attributes. Recent work has shown that learning these attributes together regularizes the models, leading to better feature representations. This study explores new forms of regularization by adding

**Notes**

unsupervised auxiliary tasks to reconstruct hidden layer representations. This auxiliary task requires the denoising of hidden representations at every layer of an auto-encoder. The framework relies on ladder networks that utilize skip connections between encoder and decoder layers to learn powerful representations of emotional dimensions. The results show that ladder networks improve the performance of the system compared to baselines that individually learn each attribute and conventional denoising autoencoders. Furthermore, the unsupervised auxiliary tasks have promising potential to be used in a semi-supervised setting, where few labeled sentences are available.

# Thu-P-2-3: Acoustic Modelling

Hall 4-6: Poster-3; 14:30-16:30; Thursday, 6 September, 2018
Chair: Rohit Prabhavalkar

## Knowledge Distillation for Sequence Model

*Mingkun Huang[1], Yongbin You[2], Zhehuai Chen[1], Yanmin Qian[1] and Kai Yu[1]*

[1]Key Lab. of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, SpeechLab, Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China

[2]AISpeech Ltd.

**Thu-P-2-3-1, Time: 14:30-16:30**

Knowledge distillation, or teacher-student training, has been effectively used to improve the performance of a relatively simpler deep learning model (the student) using a more complex model (the teacher). It is usually done by minimizing the Kullback-Leibler divergence (KLD) between the output distributions of the student and the teacher at each frame. However, the gain from frame-level knowledge distillation is limited for sequence models such as Connectionist Temporal Classification (CTC), due to the mismatch between the sequence-level criterion used in teacher model training and the frame-level criterion used in distillation. In this paper, sequence-level knowledge distillation is proposed to achieve better distillation performance. Instead of calculating a teacher posterior distribution given the feature vector of the current frame, sequence training criterion is employed to calculate the posterior distribution given the whole utterance and the teacher model. Experiments are conducted on both English Switchboard corpus and a large Chinese corpus. The proposed approach achieves significant and consistent improvements over the traditional frame-level knowledge distillation using both labeled and unlabeled data.

## Improving CTC-based Acoustic Model with Very Deep Residual Time-delay Neural Networks

*Sheng Li[1], Xugang Lu[1], Ryoichi Takashima[1], Peng Shen[1], Tatsuya Kawahara[1,2] and Hisashi Kawai[1]*

[1]National Institute of Information and Communications Technology, Kyoto, Japan

[2]Kyoto University, Kyoto, Japan

**Thu-P-2-3-2, Time: 14:30-16:30**

Connectionist temporal classification (CTC) has shown great potential in end-to-end (E2E) acoustic modeling. The current state-of-the-art architecture for a CTC-based E2E model is based on a deep bidirectional long short-term memory (BLSTM) network that provides frame-wise outputs estimated from both forward and backward directions (BLSTM-CTC). Since this architecture can lead to a serious time latency problem in decoding, it cannot be applied to real-time speech recognition tasks. Considering that the CTC label of one current frame can only be affected by a few neighboring frames, we argue that using BLSTM traversing on a whole utterance from both directions is not necessary. In this paper, we use a very deep residual time-delay (VResTD) network for CTC-based E2E acoustic modeling (VResTD-CTC). The VResTD network provides frame-wise outputs with local bidirectional information without needing to wait for the whole utterance. Speech recognition experiments on Corpus of Spontaneous Japanese were carried out to test our proposed VResTD-CTC and the state-of-the-art BLSTM-CTC model. Comparable performance was obtained while the proposed VResTD-CTC does not suffer from the decoding time latency problem.

## Filter Sampling and Combination CNN (FSC-CNN): A Compact CNN Model for Small-footprint ASR Acoustic Modeling Using Raw Waveforms

*Jinxi Guo[1], Ning Xu[2], Xin Chen[2], Yang Shi[1], Kaiyuan Xu[1] and Abeer Alwan[1]*

[1]Department of Electrical and Computer Engineering, University of California, Los Angeles, CA, USA

[2]Snap Inc., Venice, CA, USA

**Thu-P-2-3-3, Time: 14:30-16:30**

Learning an ASR acoustic model directly from raw waveforms using CNNs has proved to be effective, where convolutional layers with learnable filters are able to automatically extract useful features. However, these filters, with independent parameters, can be highly redundant resulting in inefficient systems. In this paper, we propose a novel method to generate CNN filter parameters by first sampling from a low-dimensional parameter space and then using a trainable scalar vector to perform a linear combination. This filter sampling and combination method (denoted as FSC) not only naturally enforces parameter sharing in the low-dimensional sampling space, but also adds to the learning capacity of filters. The FSC-CNN model has a significantly smaller number of parameters and is more efficient compared to conventional CNN models, which makes it feasible for small-footprint ASR. Experimental results on the WSJ LVCSR task show that FSC-CNNs are able to achieve a WER of 3.67 with a standard decoder set-up with only 1.19M nonlinear-layer parameters (better than a strong baseline CNN model with 3.2x more parameters). It also outperforms a CNN model with a similar number of parameters by a relative improvement of 10.26%.

## Twin Regularization for Online Speech Recognition

*Mirco Ravanelli, Dmitriy Serdyuk and Yoshua Bengio*

Montreal Institute for Learning Algorithms (MILA) Université de Montréal, Canada

**Thu-P-2-3-4, Time: 14:30-16:30**

Online speech recognition is crucial for developing natural human-machine interfaces. This modality, however, is significantly more challenging than off-line ASR, since real-time/low-latency constraints inevitably hinder the use of future information, that is known to be very helpful to perform robust predictions.

A popular solution to mitigate this issue consists of feeding neural acoustic models with context windows that gather some future frames. This introduces a latency which depends on the number of employed look-ahead features.

This paper explores a different approach, based on estimating the future rather than waiting for it. Our technique encourages the hidden representations of a unidirectional recurrent network to embed some useful information about the future. Inspired by a recently proposed technique called Twin Networks, we add a regularization term that forces forward hidden states to be as close as possible to cotemporal backward ones, computed by a "twin" neural network running backwards in time.

The experiments, conducted on a number of datasets, recurrent architectures, input features and acoustic conditions, have shown the effectiveness of this approach. One important advantage is that our method does not introduce any additional computation at test time if compared to standard unidirectional recurrent networks.

## Self-Attentional Acoustic Models

*Matthias Sperber[1], Jan Niehues[1], Graham Neubig[2], Sebastian Stüker[1] and Alex Waibel[1,2]*

[1]Karlsruhe Institute of Technology

**Notes**

[2]Carnegie Mellon University

**Thu-P-2-3-5, Time: 14:30-16:30**

Self-attention is a method of encoding sequences of vectors by relating these vectors to each-other based on pairwise similarities.
These models have recently shown promising results for modeling discrete sequences, but they are non-trivial to apply to acoustic modeling due to computational and modeling issues.

In this paper, we apply self-attention to acoustic modeling, proposing several improvements to mitigate these issues:

First, self-attention memory grows quadratically in the sequence length, which we address through a downsampling technique.

Second, we find that previous approaches to incorporate position information into the model are unsuitable and explore other representations and hybrid models to this end.

Third, to stress the importance of local context in the acoustic signal, we propose a Gaussian biasing approach that allows explicit control over the context range.

Experiments find that our model approaches a strong baseline based on LSTMs with network-in-network connections while being much faster to compute.

Besides speed, we find that interpretability is a strength of self-attentional acoustic models and demonstrate that self-attention heads learn a linguistically plausible division of labor.

## Hierarchical Recurrent Neural Networks for Acoustic Modeling

*Jinhwan Park, Iksoo Choi, Yoonho Boo and Wonyong Sung*
Department of Electrical and Computer Engineering, Seoul National University

**Thu-P-2-3-6, Time: 14:30-16:30**

Recurrent neural network (RNN)-based acoustic models are widely used in speech recognition and end-to-end training with CTC (connectionist temporal classification) shows good performance. In order to improve the ability to keep temporarily distant information, we employ hierarchical recurrent neural networks (HRNNs) to the acoustic modeling in speech recognition. HRNN consists of multiple RNN layers that operate on different time-scales and the frequency of operation at each layer is controlled by learned gates from training data. We employ gate activation regularization techniques to control the operation of the hierarchical layers. When tested with the WSJ eval92, our best model obtained the word error rate of 5.19% with beam search decoding using RNN based character-level language models. Compared to an LSTM based acoustic model with a similar parameter size, we achieved a relative word error rate improvement of 10.5%. Even though this model employs uni-directional RNN models, it showed the performance improvements over the previous bi-directional RNN based acoustic models.

## Dictionary Augmented Sequence-to-Sequence Neural Network for Grapheme to Phoneme Prediction

*Antoine Bruguier[1], Anton Bakhtin[2] and Dravyansh Sharma[1]*
[1]Google LLC, Mountain View, CA
[2]Facebook AI Research, New York, NY

**Thu-P-2-3-7, Time: 14:30-16:30**

Both automatic speech recognition and text to speech systems need accurate pronunciations, typically obtained by using both a lexicon dictionary and a grapheme to phoneme (G2P) model. G2P typically struggle with predicting pronunciations for tail words and we hypothesized that one reason is because they try to discover general pronunciation rules without using prior knowledge of the pronunciation of related words. Our new approach expands a sequence-to-sequence G2P model by injecting prior knowledge. In addition, our model can be updated without having to retrain a system. We show that our new model has significantly better performance for German, both on a tightly controlled task and on our real-world system. Finally, the simplification of the system allows for faster and easier scaling to other languages.

## Leveraging Second-Order Log-Linear Model for Improved Deep Learning Based ASR Performance

*Ankit Raj, Shakti P Rath and Jithendra Vepa*
Samsung Research Institute India - Bangalore

**Thu-P-2-3-8, Time: 14:30-16:30**

Gaussian generative models have been shown to be equivalent to discriminative log-linear models under weak assumptions for acoustic modeling in speech recognition systems. In this paper, we note that the output layer of deep learning model consists of a first-order log-linear model, also known as logistic regression, which induces a set of homoscedastic distributions in the generative model space, resulting in linear decision boundaries. We leverage the above equivalence to make the deep learning models more expressive by replacing the first order log-linear model with a second-order model, which leads to heteroscedastic distributions, as a result, the linear decision boundaries are replaced with quadratic ones. We observe that the proposed architecture yields a significant improvement in speech recognition accuracy compared to the conventional model having a comparable number of parameters. Relative improvement of 8.37% and 3.92% in word error rate (WER) is obtained for shallow and deep feed-forward networks respectively. Moreover, with Long Short-Term Memory (LSTM) networks with projection matrix, we obtain significant relative improvement in WER over the standard architecture.

## Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks

*Daniel Povey[1,2], Gaofeng Cheng[3,4], Yiming Wang[1], Ke Li[1], Hainan Xu[1], Mahsa Yarmohammadi[1] and Sanjeev Khudanpur[1,2]*
[1]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA
[2]Human Language Technology Center of Excellence, Johns Hopkins University, Baltimore, MD, USA
[3]University of Chinese Academy of Sciences, Beijing, China
[4]Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, Beijing, China

**Thu-P-2-3-9, Time: 14:30-16:30**

Time Delay Neural Networks (TDNNs), also known as one-dimensional Convolutional Neural Networks (1-d CNNs), are an efficient and well-performing neural network architecture for speech recognition. We introduce a factored form of TDNNs (TDNN-F) which is structurally the same as a TDNN whose layers have been compressed via SVD, but is trained from a random start with one of the two factors of each matrix constrained to be semi-orthogonal. This gives substantial improvements over TDNNs and performs about as well as TDNN-LSTM hybrids.

## Completely Unsupervised Phoneme Recognition by Adversarially Learning Mapping Relationships from Audio Embeddings

*Da-Rong Liu, Kuan-Yu Chen, Hung-yi Lee and Lin-shan Lee*
National Taiwan University

**Thu-P-2-3-10, Time: 14:30-16:30**

Unsupervised discovery of acoustic tokens from audio corpora without annotation and learning vector representations for these tokens have been widely studied. Although these techniques have been shown successful in some applications such as query-by-example Spoken Term Detection (STD), the lack of mapping relationships between these discovered tokens and real phonemes has limited the down-stream applications. This paper represents probably the first attempt towards the goal of completely unsupervised phoneme recognition, or mapping audio signals to phoneme sequences without phoneme-labeled audio data. The basic idea is to cluster the embedded acoustic tokens and learn the mapping between the cluster sequences and the unknown phoneme sequences with a Generative

**Notes**

Adversarial Network (GAN). An unsupervised phoneme recognition accuracy of 36% was achieved in the preliminary experiments.

### Phone Recognition Using a Non-Linear Manifold with Broad Phone Class Dependent DNNs

*Mengjie Qian, Linxue Bai, Peter Jančovič and Martin Russell*
Department of Electronic, Electrical & Systems Engineering, The University of Birmingham, UK

**Thu-P-2-3-11, Time: 14:30-16:30**

Although it is generally accepted that different broad phone classes (BPCs) have different production mechanisms and are better described by different types of features, most automatic speech recognition (ASR) systems use the same features and decision criteria for all phones. Motivated by this observation, this paper proposes a two-level DNN structure, referred to as a BPC-DNN, inspired by the notion of a topological manifold. In the first level, several small separate BPC-dependent DNNs are applied to different broad phonetic classes and in the second level the outputs of these DNNs are fused to obtain senone-dependent posterior probabilities, which can be used for frame level classification or integrated into Viterbi decoding for phone recognition. In a previous paper using this approach we reported improved frame classification accuracy on the TIMIT corpus compared with a conventional DNN. The contribution of the present paper is to demonstrate that this advantage extends to full phone recognition. Our most recent results show that the BPC-DNN achieves reductions in error rate relative to a conventional DNN of 16% and 8% for frame classification and phone recognition, respectively.

### A Multi-Discriminator CycleGAN for Unsupervised Non-Parallel Speech Domain Adaptation

*Ehsan Hosseini-Asl, Yingbo Zhou, Caiming Xiong and Richard Socher*
Salesforce Research

**Thu-P-2-3-12, Time: 14:30-16:30**

Domain adaptation plays an important role for speech recognition models, in particular, for domains that have low resources. We propose a novel generative model based on cyclic-consistent generative adversarial network (CycleGAN) for unsupervised non-parallel speech domain adaptation. The proposed model employs multiple independent discriminators on the power spectrogram, each in charge of different frequency bands. As a result we have 1) better discriminators that focus on fine-grained details of the frequency features and 2) a generator that is capable of generating more realistic domain-adapted spectrogram. We demonstrate the effectiveness of our method on speech recognition with gender adaptation, where the model only has access to supervised data from one gender during training, but is evaluated on the other at test time. Our model is able to achieve an average of 7.41% on phoneme error rate and 11.10% word error rate relative performance improvement as compared to the baseline, on TIMIT and WSJ dataset, respectively. Qualitatively, our model also generates more natural sounding speech, when conditioned on data from the other domain.

## Thu-P-2-4: Speech and Speaker Perception

Hall 4-6: Poster-4; 14:30-16:30; Thursday, 6 September, 2018
Chair: Chris Davis

### Interactions between Vowels and Nasal Codas in Mandarin Speakers' Perception of Nasal Finals

*Chong Cao, Wei Wei, Wei Wang, Yanlu Xie and Jinsong Zhang*
Beijing Advanced Innovation Center for Language Resources
Beijing Language and Culture
University, Beijing, China

**Thu-P-2-4-1, Time: 14:30-16:30**

Previous studies have shown vowels' effect on the perception of nasal codas from the perspectives of acoustics and phonetics with evidence from Mandarin. While few studies investigated the processing interactions between vowels and nasal codas during the perceptual processing of nasal Finals. Using a speeded classification paradigm which requires participants to attend to one dimension selectively and to ignore other dimensions, we aimed to explore the processing interactions between vowels and nasal codas in the process of nasal Finals' perception by Mandarin speakers. Results of the speeded classification experiment showed that there was a mutual and symmetrical pattern of interference effect between vowel and nasal coda dimensions. More specifically, participants were affected by irrelevant variation of the nasal coda dimension when they were attending to the vowel dimension and affected by irrelevant variation of the vowel dimension when they were attending to the nasal coda dimension. The extent of such mutual interference effect was equal. Results were discussed in terms of phonemic differences in acoustic properties and relative discriminability between vowels and nasal codas. The present study may be helpful to second language learners of Mandarin during their course of acquiring nasal Finals.

### Weighting Pitch Contour and Loudness Contour in Mandarin Tone Perception in Cochlear Implant Listeners

*Qinglin Meng[1,2], Nengheng Zheng[3], Ambika Prasad Mishra[2], Jacinta Dan Luo[2] and Jan W. H. Schnupp[2]*
[1]Acoustics Lab, School of Physics and Optoelectronics, South China University of Technology, China
[2]Hearing Research Group, Biomedical Science Department, City University of Hong Kong, Hong Kong SAR of China
[3]College of Information Engineering, Shenzhen University, China

**Thu-P-2-4-2, Time: 14:30-16:30**

Previous investigations found that loudness-contours within individual Mandarin monosyllables can drive categorical perception of Mandarin tone for cochlear implant (CI) users, while in normal hearing (NH) subjects the pitch contour is phonologically acknowledged to be the dominant cue. Here we further examine the weighting strategy of pitch induced and loudness induced contour identification on Mandarin tone perception by CI users. Twenty-seven versions of the disyllabic utterance /Lao3 Shi/ with orthogonally manipulated loudness-contour and pitch-contour of the voiced portion of the second monosyllable /Shi/ served as the stimuli to both CI and NH subjects. In Mandarin, if /Shi/ is pronounced with high-flat-pitched Tone 1 the word means "teacher", with rising Tone 2 it means "well-behaved", or with falling Tone 4 it means "always". CI users generally had poorer word-recognition scores and their inter-subject variance was large. While NH subjects recognized tone reliably based on pitch-contour, half of the CI users relied on pitch-contour, the other half on loudness-contour, implying systematic differences in pitch coding in their CI processing. This paradigm of orthogonal manipulation of pitch and loudness contours could be developed into improved audiometric tests of Mandarin tone perception and pitch coding with CIs.

### Implementing DIANA to Model Isolated Auditory Word Recognition in English

*Filip Nenadić[1], Louis ten Bosch[2] and Benjamin V. Tucker[1]*
[1]University of Alberta
[2]Radboud University Nijmegen

**Thu-P-2-4-3, Time: 14:30-16:30**

DIANA, an end-to-end computational model of spoken word recognition, was previously used to simulate auditory lexical decision experiments in Dutch. A single test conducted for North American English showed promising results as well. However, this simulation used a relatively small amount of data collected in the pilot phase of the Massive Auditory Lexical Decision (MALD) project. Additionally, already existing acoustic models were implemented. In this paper, we expand the analysis of MALD data by including a larger sample of both stimuli and participants. Acknowledging that most speech humans hear is conversational speech, we also test new acoustic models created using spontaneous speech corpora. Simulations successfully replicate expected trends in word competition and show

**Notes**

plausible competitors as the signal unfolds, but acoustic model accuracy should be improved. Despite the number of responses per word being relatively small (never more than five), correlations between model estimates and participants' responses are moderate. Future directions in acoustic model training and simulating MALD data are discussed.

## Effects of Homophone Density on Spoken Word Recognition in Mandarin Chinese

*Bhamini Sharma*

The Hong Kong Polytechnic University, Hong Kong SAR

**Thu-P-2-4-4, Time: 14:30-16:30**

Homophones, words that sound same, influence spoken word recognition. Whether the effects of homophone density (i.e., number of same-sounding words) on spoken word recognition are facilitatory or inhibitory or complex is a matter of ongoing debate. In addition, there are limited studies investigating the effects of homophone density, probably due to paucity of homophones in the examined languages (e.g., English). In comparison, languages such as Mandarin Chinese have abundant homophony that makes it a suitable tool to investigate the effects of homophone density. In the current study, an auditory naming task was conducted using Mandarin Chinese to investigate the effects of homophone density on spoken word recognition. Using mixed modeling, a significant inhibitory effect of homophone density ($\beta$ = 0.0098, t = 2.10) on reaction time was found. Participants were slower in naming words with high homophone density, possibly due to competition posed by more number of homophones, as compared to the words with low homophone density. Further, an interaction between homophone density and syllable frequency was found i.e., for high syllable frequency, homophone density effects were inhibitory but for low syllable frequency, the inhibitory effect was reduced. Taken together, the effects of homophone density are not straightforward but complex.

## Visual Timing Information in Audiovisual Speech Perception: Evidence from Lexical Tone Contour

*Hui Xie[1,2], Biao Zeng[3] and Rui Wang[4]*

[1]Faculty of Architecture and Urban Planning, Chongqing University, China.

[2]Key Laboratory of New Technology for Construction of Cities in Mountain Areas, Ministry of Education, Chongqing University, China.

[3]School of Psychology and Therapeutic Studies, University of South Wales, UK

[4]Department of Psychology, Bournemouth University, UK.

**Thu-P-2-4-5, Time: 14:30-16:30**

The present study investigated whether duration of lip movement could improve intelligibility of lexical pitch contours under noisy condition. Eighteen Chinese speakers were asked to identify a Mandarin lexical tone in one pair of tones under auditory only (AO) and audiovisual (AV) condition. Two types of tone pairs were used in the study: maximum contrastive pair (falling vs. dipping tones, the durational difference of lip movement was 100ms) and minimum contrastive pair (rising vs. falling tones, the difference was 33ms). The results showed that duration of lip movement enhanced discrimination in the maximum pair whereas the similar lengths of rising and dipping tones attenuated such visual benefit. The finding suggested that visual timing information could be a specific cue for audiovisual lexical tone perception.

## COSMO SylPhon: A Bayesian Perceptuo-motor Model to Assess Phonological Learning

*Marie-Lou Barnaud[1,2], Juien Diard[2], Pierre Bessière[3] and Jean-Luc Schwartz[1]*

[1]Univ. Grenoble Alpes, CNRS, Grenoble INP*, GIPSA-lab, 38000 Grenoble, France

[2]Univ. Grenoble Alpes, CNRS, LPNC, 38000 Grenoble, France

[3]CNRS - Sorbonne Université - ISIR, Paris, France

**Thu-P-2-4-6, Time: 14:30-16:30**

During speech development, babies learn to perceive and produce speech units, especially syllables and phonemes. However, the mechanisms underlying the acquisition of speech units still remain unclear. We propose a Bayesian model of speech communication, named "COSMO SylPhon", for studying the acquisition of both syllables and phonemes. In this model, speech development involves a sensory learning phase mainly related to perception development and a motor learning phase mainly related to production development. We analyze how an agent can learn speech units during these two phases through an unsupervised learning process based on syllable stimuli. We show that the learning process enables to efficiently learn the distribution of syllabic stimuli provided in the environment. Importantly, we show that if agents are equipped with a bootstrap process inspired by the Frame-Content Theory of speech development, they learn to associate consonants to specific articulatory gestures, providing the basis for consonantal articulatory invariance.

## Experience-dependent Influence of Music and Language on Lexical Pitch Learning Is Not Additive

*Akshay Raj Maggu [1], Patrick C. M. Wong [1], Hanjun Liu [2] and Francis C. K. Wong [3]*

[1]The Chinese University of Hong Kong, Hong Kong SAR

[2]Sun Yat-sen University, China

[3]Nanyang Technological University, Singapore

**Thu-P-2-4-7, Time: 14:30-16:30**

Research studies provide evidence for the facilitative effects of musical and linguistic experience on lexical pitch learning. However, the effect of interaction of linguistic and musical pitch experience on lexical pitch processing is a matter of ongoing research. In the current study, we sought to examine the effect of combination of musical and linguistic pitch experience on learning of novel lexical pitch. Using a 10-session pseudoword-picture association training paradigm, we compared the learning performance of musicians and non-musicians who either spoke a non-tone language, spoke one tone language, or spoke two tone languages. Among the non-tone language speakers, we found that musicians showed enhanced learning of novel lexical pitch as compared to non-musicians. In comparison, among the tone-language speakers, we found no significant difference in the learning performance of musicians and non-musicians no matter they spoke one or more tone languages. We conclude that though musical experience facilitates linguistic pitch learning, the effects of combination of musical and linguistic pitch experience are not additive i.e. possessing both types of pitch experience is no better than possessing either one of them and knowing two tone languages does not facilitate the learning of a new tone language beyond the knowledge of one.

## Influences of Fundamental Oscillation on Speaker Identification in Vocalic Utterances by Humans and Computers

*Volker Dellwo[1], Thayabaran Kathiresan[1], Elisa Pellegrino[1], Lei He[1], Sandra Schwab[1] and Dieter Maurer[2]*

[1]Department of Computational Linguistics, University of Zurich

[2]Department of the Performing Arts, Zurich University of the Arts

**Thu-P-2-4-8, Time: 14:30-16:30**

We tested the influence of fundamental oscillation (fo) on human and machine speaker recognition performance in vocalic test utterances. In experiment I, we trained a Gaussian-Mixture model on 15 speakers (80 multi-word utterances each) and tested it with sustained vowel utterances (/a:/, /i:/ and /u:/) under six fo conditions, three changing (fall, rise, fall-rise) and three steady-state (high, mid, low). Results revealed better performance for the steady-state compared to the changing conditions and within the steady-state condition, performance was poorest for high fo. In experiment II, we tested 9 human listeners on a subset of 4 speakers from experiment I. They went through two training tasks (training 1: multi-word utterances; training 2: words). In the test, they recognized speakers based on the same vocalic utterances as in experiment I (for these 4 speakers). Results showed that performance was about equally high for the changing and steady-state vowels, however, in the steady-state condition performance was best for high fo vowels. The experiments suggest that (a) fo has an influence on the strength of speaker specific characteristics in vowels and (b) humans - compared to machines - pay attention to different acoustic information in vocalic utterances for speaker recognition.

**Notes**

# Session Chair Index


Abraham, Ajish..... 122
Alku, Paavo.....177
186
Anguera, Xavier..... 160
Bäckström, Tom.....158
Bali, Kalika.....183
235
246
Barker, Jon..... 216
Batliner, Anton..... 109
125
Baumeister, Harald.....109
125
Bhat, Chitralekha.....203
Bilgi, Raghavendra..... 246
Biyani, Pravesh.....247
Blanco, Noelia do Carmo.....203
Bonastre, Jean-François.....106
Boril, Hynek.....178
Boves, Lou..... 164
Breen, Andrew..... 141
Busso, Carlos..... 142
Cahill, Peter..... 230
Černocký, Jan.....197
Church, Kenneth..... 219
Cieri, Christopher..... 219
Compernolle, Dirk Van.....212
Cristia, Alejandrina.....209
219
Cui, Xiaodong.....107
Cummins, Nicholas..... 231
D, Dileep A..... 158
Davis, Chris.....259
Deepak, K T.....195
Delgado, Héctor.....197
Dellwo, Volker..... 161
193
Deshpande, Gauri.....158
Diab, Mona..... 183
Diener, Lorenz.....233
Dong, Fengquan.....109
125
Du, Jun.....219
Earnshaw, Katherine.....217
Epps, Julien..... 231
Esch, Daan van.....146
Evans, Nicholas.....106
Ganapathy, Sriram.....148
160
219
Garimella, Sri.....229
Gerlach, Johanna.....201
Gobl, Christer.....161
Gowda, Dhananjaya N.....158
Green, Phil..... 225
Hagedorn, Christina.....173
182
Hagmuller, Martin..... 229
Hain, Thomas.....140
Hasegawa-Johnson, Mark.....159
Hermansky, Hynek..... 214
Hirschberg, Julia..... 183
Hoffmeister, Bjorn.....249
Ishikawa, Haruko.....146
Jayagopi, Dinesh..... 249
Jayaraman, Rajagopal.....165
Jyothi, Preethi.....146
K, Samudravijaya..... 198
246
Karafiat, Martin..... 223
Karpov, Prof Alexey.....218
Kim, Jeesun.....198
Knill, Kate.....136
Kotti, Margarita.....178
Kubin, Gernot..... 250
Kurimo, Mikko..... 204
241
Lee, Chi-Chun..... 254
Liberman, Mark..... 163
219
Lin, Qiguang.....199
M, Mahesh.....236
Madikeri, Srikanth.....246
Magimai Doss, Mathew.....117
138
Mandel, Michael.....247
Mariani, Joseph..... 218
Mehta, Rupesh Kumar.....235
Meng, Helen.....145
Metze, Florian..... 214
Meyer, Bernd T..... 143
Mitra, Vikramjit..... 119
Mixdorff, Hansjörg.....180
Mohan, C. Krishna.....127
195
Mohan, Krishna Doss.....235
246
Moniz, Helena.....128
Moogi, Pratibha.....195
Moreno, Pedro.....146
Murthy, Hema.....140
Namboodiri, Vinay..... 239
Narayanan, Shrikanth.....173
177
182
Nayak, Niranjan.....148
235
Nguyen, Patrick..... 195
Ni Chasaide, Ailbhe..... 121
Nixon, Jessie..... 163
Noeth, Elmar..... 142
243
P, Laxminarayana.....148
P, Narenda N.....112
Padmanabhan, R.....111
Pandey, P.C.....232
Patil, Hemant.....158
Payne, Elinor.....217
Prabhavalkar, Rohit.....168
257
Prasanna, S R Mahadeva.....195
Ramasubramanian, V.....191
Rao, K. Sreenivasa.....185
Rao, Preeti.....228
Rastrow, Ariya.....120
Rayner, Manny.....201
Riccardi, Giuseppe.....246
Riley, Michael.....195
Russell, Martin.....201
Ryant, Neville..... 219
Sainath, Tara.....177
Sankar, Ananth.....158
Sarmah, Priyankoo.....246
Schlüter, Ralf.....177
Schuller, Björn.....109
125
Schultz, Tanja.....200
233
Schwartz, Rich..... 140
Seelamantula, Chandra Sekhar.....195
232
Sharkh, Sarah Abu.....146
Shinoda, Koichi.....251
Silva, Thiago Fraga da.....107
188
Simko, Juraj.....180
Singh, Mittul..... 120
Singh, Rita..... 146
Sinha, Rohit..... 143
158
Sinha, Uttam.....173
182
Sitaram, Sunayana.....183
235
Solorio, Thamar..... 183
Sreenivas, T V.....166
226
Srikanth, Radhakrishnan.....235
Srinivasamurthy, Ajay.....153
158
165
Steidl, Stefan.....109
125
Strik, Helmer.....171
201
Stylianou, Yannis.....124
181
T, Balamurali B.....199
T, Veena.....127
Toda, Tomoki.....124
Trancoso, Isabel.....250
van Son, Rob.....122
Velmurugan, Rajbabu.....151
158
Vepa, Jitendra.....236
Vijayasenan, Deepu.....246
Vuppala, Anil..... 195
Wang, Weiran.....159
Watanabe, Shinji.....237
Watts, Oliver.....105
230
Weinstein, Eugene.....146
Wu, Chung-Hsien.....175
Ye, Wang.....156
Yegnanarayana, Bayya.....104
Yu, Dong.....104
Zhang, Zixing.....121

# Author Index


**A**

A G, Ramakrishnan Wed-P-1-3-2 191
Wed-S&T-1-1-5 185
A. Salim, Fahim Mon-O-2-2-4 122
Aare, Kätlin Tue-O-2-4-3 162
Abad, Alberto Thu-P-1-4-9 245
Abe, Masanobu Wed-P-2-2-1 206
Abraham, Basil Thu-SS-1-2-7 236
Abrol, Vinayak Thu-P-1-2-5 240
Wed-P-1-3-7 192
Adachi, Hiroyoshi Tue-SS-2-2-5 174
Adda-Decker, Martine Wed-O-3-3-6 218
Adeel, Ahsan Wed-O-3-2-6 216
Afouras, Triantafyllos Thu-O-2-3-3 249
Thu-P-1-1-6 238
Afshan, Amber Mon-SS-1-1-11 111
Tue-SS-2-2-2 173
Agarwal, Dharmesh Wed-S&T-2-1-1 203
Aggarwal, Ritu Thu-O-1-4-4 233
Aggarwal, Sidharth Tue-S&T-1-1-5 148
Agrawal, Purvi Wed-O-3-2-4 216
Wed-P-3-2-6 224
Agurto, Carla Mon-P-1-4-5 120
Aharon, Dan Thu-S&T-1-1-3 237
Ahmadi, Farzaneh Thu-SS-1-1-2 234
Ahmed, Beena Tue-SS-2-2-1 173
Ainger, Eloise Wed-O-2-5-3 201
Airaksinen, Manu Mon-P-2-3-2 133
Wed-P-1-1-10 187
Ajili, Moez Tue-P-1-1-1 148
Akagi, Masato Thu-P-2-2-2 254
Akhanov, Egor Tue-P-1-1-8 149
Akuzawa, Kei Thu-O-1-2-4 230
Al Hanai, Tuka Tue-SS-2-2-10 175
Alam, Md Jahangir Thu-P-2-1-6 252
Thu-P-2-2-6 255
Alcorn, Alyssa Wed-O-2-5-3 201
Aleksic, Petar Wed-O-2-1-4 196
Wed-O-2-1-5 196
Alharbi, Sadeen Thu-P-1-4-10 245
Ali, Hussnain Tue-SS-2-2-4 174
Alku, Paavo Mon-P-2-3-2 133
Thu-P-1-4-4 244
Wed-P-1-1-10 187
Alleva, Fil Thu-O-1-1-4 229
Alwan, Abeer Mon-SS-1-1-11 111
Thu-P-2-3-3 257
Tue-P-2-3-9 172
Tue-SS-2-2-2 173
Ambikairajah, Eliathamby Mon-P-2-2-10 132
Mon-P-2-2-14 133
Mon-P-2-2-3 131
Mon-P-2-2-8 132
Wed-O-1-2-5 179
Amiriparian, Shahin Mon-SS-1-1-1 109
Mon-SS-1-1-2 109
Mon-SS-1-1-4 109
Mon-SS-1-1-8 110
Mon-SS-2-1-1 125
Mon-SS-2-1-10 127
Wed-O-2-5-3 201
Amoh, Justice Tue-O-2-2-6 160
An, Guozhen Mon-O-2-2-2 121
Tue-P-2-5-9 176
An, Kwanghoon Wed-P-3-3-3 225
Wed-SS-1-1-6 183
Ananthakrishnan, Shankar Wed-P-1-2-6 189
Ananthapadmanabha, TV Wed-P-1-3-2 191
Anastasopoulos, Antonios Tue-P-1-4-6 157
Ando, Atsushi Tue-P-1-4-2 156
Tue-P-2-5-3 175
André, Elisabeth Mon-SS-1-1-9 111
Andrés-Ferrer, Jesús Thu-P-1-3-8 243
Aneeja, Gunnam Wed-O-2-4-2 199
Angrick, Miguel Wed-O-2-5-1 200
Anumula, Jithendar Mon-O-1-1-4 104
Mon-P-2-5-4 138
Aono, Yushi Mon-O-2-1-4 121
Tue-P-1-4-2 156
Aono, Yushi Tue-P-1-4-4 156
Tue-P-2-5-3 175
Wed-P-2-1-2 204
Wed-P-2-1-7 205
Arai, Takayuki Mon-S&T-1-1-3 112
Araki, Shoko Wed-O-1-4-2 181
Aralikatti, Rohith Wed-O-1-4-5 181
Arias, Tomas Mon-O-2-3-3 123
Arkhangorodsky, Arkady Tue-P-1-4-7 157
Armstrong, Zeb Tue-P-2-2-2 169
Asaei, Afsaneh Wed-P-1-2-8 190
Asakawa, Reine Tue-P-2-5-3 175
Ashkenazi, Amit Wed-S&T-1-1-4 185
Atal, Bishnu S. Mon-Medal-1 104
Atcheson, Mia Thu-P-2-2-7 255
Ateeq, Mohammad Wed-SS-2-1-4 202
Athar, Awais Tue-SS-1-1-3 146
Atkins, David Thu-P-1-4-6 245
Wed-O-2-5-5 201
Wed-SS-1-1-5 183
Aubanel, Vincent Wed-O-2-3-3 198
Auris, Bastian Mon-P-1-1-5 113
Avila, Anderson R. Thu-P-2-2-6 255
Axtmann, Nico Thu-S&T-1-1-1 236
Ayyavu, Madhavaraj Wed-S&T-1-1-5 185
Azarbayejani, Ali Mon-SS-2-1-2 125

**B**

B T, Balamurali Wed-O-2-4-5 200
B, Ganga Gowri Wed-P-1-3-12 193
Baayen, R. Harald Tue-O-1-4-3 144
Baby, Arun Wed-SS-1-2-7 184
Baby, Deepak Thu-P-1-1-10 239
Bacchiani, Michiel Thu-O-1-1-2 229
Tue-O-1-1-6 141
Bachhav, Pramod Tue-P-1-2-11 153
Bäckström, Tom Thu-O-2-4-1 250
Thu-O-2-4-2 250
Thu-O-2-4-3 250
Badino, Leonardo Mon-P-2-1-2 128
Tue-P-1-4-5 156
Bae, Jaesung Mon-P-2-4-4 136
Bagi, Anita Thu-P-1-4-5 244
Bahmaninezhad, Fahimeh Tue-P-1-1-3 149
Bai, Haichuan Thu-P-1-2-6 240
Bai, Linxue Thu-P-2-3-11 259
Tue-SS-2-1-4 165
Bai, Zhongxin Tue-P-1-1-14 151
Bailly, Gérard Wed-P-2-2-13 208
Baird, Alice Mon-SS-1-1-1 109
Mon-SS-1-1-2 109
Mon-SS-1-1-4 109
Baird, Alice Mon-SS-1-1-8 110
Mon-SS-2-1-1 125
Mon-SS-2-1-10 127
Thu-P-2-2-1 254
Wed-O-2-5-3 201
Wed-P-3-1-9 222
Bakhtin, Anton Thu-P-2-3-7 258
Baldwijns, Daan Wed-S&T-1-1-3 185
Baljekar, Pallavi Wed-P-2-2-3 207
Ballard, Kirrie Tue-SS-2-2-1 173
Bandini, Andrea Tue-SS-2-2-9 175
Bangalore, Srinivas Mon-P-2-1-11 130
Banno, Hideki Wed-P-1-1-13 188
Bansal, Sameer Tue-P-1-4-10 157
Bao, Feilong Mon-O-1-2-6 106
Bao, Xiuguo Thu-P-1-2-6 240
Barker, Jon Mon-P-2-3-1 133
Tue-P-2-2-1 168
Wed-O-3-2-6 216
Barnaud, Marie-Lou Thu-P-2-4-6 260
Barras, Claude Tue-O-2-3-6 161
Barrett, Tyson S. Mon-P-2-1-6 129
Barrios, Maria A. Tue-P-2-2-2 169
Thu-O-2-2-1 247
Bartels, Chris Tue-P-2-2-2 169
Wed-P-3-3-5 226
Bartl-Pokorny, Katrin D. Mon-SS-1-1-1 109
Bartl-Pokorny, Katrin D. Mon-SS-1-1-2 109
Mon-SS-1-1-4 109
Mon-SS-1-1-8 110
Mon-SS-2-1-1 125
Mon-SS-2-1-10 127
Baskar, Murali Karthick Thu-SS-1-2-3 235
Wed-P-2-4-9 213
Basu, Joyanta Tue-S&T-2-1-3 165
Batista, Fernando Wed-P-1-4-5 194
Batliner, Anton Mon-SS-1-1-1 109
Mon-SS-1-1-2 109
Mon-SS-1-1-4 109
Mon-SS-1-1-8 110
Mon-SS-2-1-1 125
Mon-SS-2-1-10 127
Thu-P-2-2-1 254
Baucom, Brian Thu-P-1-4-8 245
Wed-O-2-5-4 201
Baumann, Timo Mon-S&T-1-1-1 111
Wed-P-1-4-1 193
Wed-P-1-4-2 193
Baumeister, Harald Mon-P-1-2-6 116
Mon-SS-1-1-1 109
Mon-SS-1-1-2 109
Mon-SS-1-1-4 109
Mon-SS-1-1-8 110
Mon-SS-2-1-1 125
Mon-SS-2-1-10 127
Thu-P-1-4-1 243
Baur, Claudia Wed-SS-2-1-1 201
Baziotis, Christos Tue-O-1-3-1 142
Beare, Richard Wed-P-1-4-6 194
Beautemps, Denis Wed-P-2-4-10 213
Béchet, Frédéric Thu-O-2-1-2 247
Beck, Eugen Mon-P-2-4-2 136
Beke, András Wed-P-2-4-7 213
Bell, Peter Tue-O-1-1-1 140
Bellamkonda, M. R. Tue-S&T-1-1-3 148
Belyaev, Andrey Tue-O-2-3-5 161
Ben-David, Shay Mon-S&T-2-1-3 127
Beneš, Karel Thu-P-1-3-11 243
Bengio, Yoshua Mon-O-1-1-5 105
Bengio, Yoshua Thu-P-2-3-4 257
Ben-Reuven, Ehud Wed-O-1-2-3 179
Bepari, Milton S. Tue-S&T-2-1-3 165
Berard, Alexandre Wed-O-3-1-3 215
Berisha, Visar Mon-P-2-1-6 129
Thu-P-2-1-10 253
Tue-P-2-3-4 171
Berkling, Kay Wed-SS-2-1-2 202
Bernstein, Jared Thu-P-2-2-4 255
Besacier, Laurent Wed-O-3-1-3 215
Besirli, Aslı Thu-P-1-4-3 244
Bessière, Pierre Thu-P-2-4-6 260
Bhat, Chitralekha Mon-O-2-3-2 123
Bhat, Chitralekha Mon-O-2-3-6 123
Bhattacharya, Gautam Thu-P-2-1-6 252
Biasutto-Lervat, Théo Thu-O-1-4-1 232
Biemann, Chris Wed-O-3-1-6 215
Billa, Jayadev Thu-SS-1-2-8 236
Billington, Rosey Wed-P-1-4-8 194
Bin Zia, Haris Tue-SS-1-1-3 146
Biswas, Arijit Thu-O-2-4-4 250
Biswas, Astik Wed-P-2-4-2 212
Wed-SS-1-2-2 184
Biyani, Pravesh Thu-O-2-2-5 248
Black, Alan W Mon-P-2-1-12 130
Mon-SS-2-1-4 125
Wed-P-2-2-3 207
Blackburn, Daniel Wed-SS-1-1-2 182
Blood, Ian Mon-S&T-2-1-2 127
Bo, Xu Thu-P-2-1-14 254
Bodenstab, Nathan Thu-P-1-3-8 243
Boeddeker, Christoph Thu-O-1-1-5 229
Boito, Marcely Zanon Wed-O-3-1-3 215
Boll-Avetisyan, Natalie Tue-O-2-5-2 163
Bollepalli, Bajibabu Wed-P-1-1-10 187
Bonafonte, Antonio Thu-O-1-2-3 230
Tue-S&T-2-1-4 166

Bonafonte, Antonio Wed-P-1-1-7 187
Bonastre, Jean-François Tue-P-1-1-1 148
Wed-O-3-4-4 219
Bondale, Nandini Wed-SS-1-2-1 183
Boo, Yoonho Thu-P-2-3-6 258
Borrie, Stephanie A. Mon-P-2-1-6 129
Bothe, Chandrakant Tue-O-1-5-3 145
Boulard, Yannick Tue-P-1-1-8 149
Boulianne, Gilles Wed-P-2-4-12 214
Bourlard, Hervé Mon-O-1-4-1 107
Tue-O-2-1-3 159
Wed-P-1-2-8 190
Wed-P-2-1-14 206
Wed-Plenary-2 177
Bouserhal, Rachel E. Tue-P-2-1-4 166
Boves, Lou Tue-O-1-4-4 144
Tue-O-2-5-6 164
Tue-SS-2-1-1 164
Braga, Daniela Wed-S&T-1-1-3 185
Brasser, Ferdinand Tue-P-1-4-11 157
Braun, Sebastian Tue-O-2-1-2 158
Braun, Stefan Mon-O-1-1-4 104
Braunschweiler, Norbert Mon-P-2-1-5 129
Breakstone, Micha Wed-S&T-1-1-4 185
Bredin, Hervé Tue-O-2-3-6 161
Brenndoerfer, Michael Thu-S&T-1-1-1 236
Bridle, John Wed-P-1-2-13 190
Broux, Pierre-Alexandre Tue-O-2-3-1 160
Broz, Frank Tue-P-1-3-8 155
Bruguier, Antoine Thu-P-2-3-7 258
Tue-P-1-4-7 157
Brumfitt, Shelagh Thu-P-1-4-10 245
Brümmer, Niko Mon-O-1-3-3 107
Bruno, James Mon-S&T-2-1-2 127
Brutti, Richard Mon-SS-2-1-2 125
Bu, Suliang Thu-O-1-1-6 230
Buanzur, Tuarik Mon-P-2-1-14 130
Bui, Trung Mon-S&T-1-1-2 111
Bullock, Barbara E. Wed-P-2-3-1 209
Bulut, Ahmet E. Wed-O-3-4-1 218
Bunnell, H Timothy Mon-O-2-3-4 123
Burget, Lukáš Mon-O-1-3-3 107
Thu-P-1-3-11 243
Thu-SS-1-2-3 235
Wed-P-2-4-9 213
Wed-SS-3-1-3 220
Burkhardt, Felix Wed-P-3-1-9 222
Buschermöhle, Eric Mon-P-1-2-5 115
Busso, Carlos Mon-P-1-2-3 115
Thu-P-2-2-13 256
Tue-O-1-3-4 143
Tue-O-1-3-6 143
Tue-P-1-3-10 155
Bušta, Michal Thu-O-2-3-6 250
Butcher, Andrew Wed-P-1-4-6 194


## C


C M, Vikram Mon-O-1-4-6 108
Mon-P-1-4-6 120
Wed-O-2-4-4 200
C, Mahima Thu-SS-1-2-6 236
Tue-SS-1-1-5 147
Wed-S&T-2-1-3 203
CA, Valliappan Thu-O-1-4-5 233
Cabarrão, Vera Wed-P-1-4-5 194
Cabral, João Wed-P-3-4-5 227
Cahill, Aoife Wed-SS-2-1-6 202
Cai, Lianhong Mon-O-1-4-3 108
Thu-P-2-2-10 256
Cai, Weicheng Thu-P-2-1-12 253
Cai, Xinyuan Thu-P-2-1-14 254
Caines, Andrew Tue-P-2-3-5 172
Wed-SS-2-1-1 201
Camelin, Nathalie Tue-P-1-4-8 157
Campbell, Nick Tue-P-2-5-4 176
Campbell, Thomas Tue-SS-2-2-3 173
Wed-SS-1-1-6 183
Can, Dogan Thu-P-1-4-6 245
Cao, Beiming Thu-SS-1-1-4 234
Wed-P-3-3-3 225
Cao, Chong Thu-P-2-4-1 259
Cao, Yuewen Thu-O-1-2-5 231
Caraty, Marie-José Mon-SS-2-1-9 126
Cardinal, Patrick Tue-P-2-1-4 166
Carlozzi, Noelle Wed-SS-1-1-3 182
Carmiel, Yishay Mon-P-1-2-2 115
Wed-P-2-4-8 213
Carrive, Jean Tue-O-2-3-1 160
Casillas, Marisa Tue-P-1-3-1 154
Wed-P-2-3-11 211
Castan, Diego Thu-P-2-1-2 251
Caudrelier, Tiphaine Tue-O-1-4-1 143
Cecchi, Guillermo Mon-P-1-4-5 120
Celik, Selime Thu-P-1-4-3 244
Ceolini, Enea Mon-O-1-1-4 104
Mon-P-2-5-4 138
Cerda, Mauricio Wed-O-3-4-5 219
Černocký, Jan Thu-SS-1-2-3 235
Tue-O-2-1-6 159
Wed-P-2-4-9 213
Wed-P-3-2-1 223
Cerva, Petr Wed-O-1-2-2 179
Cervone, Alessandra Tue-O-1-5-6 146
Cestero, Nishmar Wed-SS-1-2-5 184
Chabot, Philippe Tue-P-2-1-4 166
Chai, Li Thu-P-1-1-11 239
Chakrabarty, Soumitro Tue-O-2-1-2 158
Chakraborty, Kishalay Wed-S&T-2-1-5 203
Champagne, Benoit Thu-P-1-1-2 237
Chanchaochai, Nattanun Mon-P-1-1-4 113
Chandra Shekar, Ram C. Tue-SS-2-2-4 174
Chandrashekaran, Akshay Mon-P-2-4-8 137
Chang, Shuangyu Thu-P-1-3-10 243
Chang, Walter Mon-S&T-1-1-2 111
Chang, Wei-Wen Mon-O-2-2-5 122
Chang, Xuankai Tue-P-2-2-6 169
Chaudhuri, Sourish Tue-P-1-3-9 155
Chauhan, Surbhi Wed-S&T-2-1-1 203
Chebolu, Pooja Wed-O-3-4-2 218
Chellapriyadharshini, M. Tue-SS-1-1-7 147
Chellu, Chandra Sekhar Tue-P-1-3-4 154
Chen, Bo Wed-P-1-1-6 187
Chen, Deming Wed-P-2-4-5 212
Chen, Hangting Thu-P-1-2-6 240
Chen, Hongjie Mon-O-1-4-2 108
Chen, Hsin-Hsi Tue-P-2-5-1 175
Chen, Huan-Yu Mon-SS-2-1-8 126
Chen, Jer-Ming Wed-O-2-4-5 200
Chen, Jie Mon-P-1-3-2 117
Wed-P-2-4-4 212
Chen, Jing Tue-O-1-4-2 144
Chen, Jingdong Tue-P-1-1-14 151
Chen, Jinkun Thu-P-2-1-12 253
Chen, Kai Wed-P-1-2-10 190
Chen, KaWai Thu-P-1-3-5 242
Chen, Kuan Wed-P-1-1-6 187
Chen, Kuan-Yu Thu-P-2-3-10 258
Tue-P-2-5-1 175
Chen, Lei Wed-SS-2-1-3 202
Chen, Lianwu Mon-P-1-3-1 117
Mon-P-1-3-2 117
Wed-P-2-4-4 212
Chen, Mengzhe Wed-P-2-4-14 214
Chen, Michael Thu-P-1-4-2 244
Chen, Minbo Tue-O-2-5-1 163
Chen, Nancy F. Wed-P-1-2-4 189
Chen, Nanxin Mon-O-1-3-6 107
Mon-P-1-2-8 116
Wed-O-2-1-1 196
Chen, Oscar Thu-P-1-3-2 242
Chen, Peixin Wed-P-3-1-7 222
Chen, Si Tue-O-2-4-4 162
Chen, Sin-Horng Mon-O-1-2-3 105
Chen, Szu-Jui Thu-P-1-1-7 238
Tue-P-2-2-3 169
Chen, Wei Mon-P-2-4-12 138
Chen, Wenda Wed-P-1-2-4 189
Chen, Xiao Tue-O-2-1-5 159
Chen, Xiaomin Tue-O-1-3-2 142
Chen, Xie Thu-P-1-3-2 242
Chen, Xin Thu-P-2-3-3 257
Chen, Ying Wed-P-2-3-6 210
Chen, Zhehuai Thu-P-2-3-1 257
Wed-O-2-1-2 196
Chen, Zhi Wed-P-3-1-7 222
Chen, Zhuo Thu-O-1-1-4 229
Chen, Zhuohao Thu-P-1-4-6 245
Cheng, Gaofeng Thu-P-2-3-9 258
Wed-O-1-1-6 178
Wed-P-3-2-2 223
Cheng, Jian Thu-P-2-2-4 255
Tue-P-2-3-1 171
Cheng, Xingliang Thu-P-2-2-9 256
Cheon, Sung Jun Tue-O-1-2-5 142
Chhabra, Diksha Thu-SS-1-2-4 235
Chiang, Chen-Yu Mon-O-1-2-3 105
Chiang, David Tue-P-1-4-6 157
Chiba, Yuya Tue-P-2-5-6 176
Chibuye, Nick K Wed-P-2-4-6 213
Ching, Joey Tue-P-1-2-12 153
Chiu, Chung-Cheng Mon-O-1-1-6 105
Wed-P-3-3-8 226
Chng, Eng Siong Mon-S&T-2-1-5 128
Thu-O-2-2-2 248
Thu-P-1-3-3 242
Wed-SS-1-2-3 184
Cho, Jaejin Mon-P-1-2-2 115
Cho, Sukhyun Thu-P-2-1-1 251
Chodroff, Eleanor Wed-O-1-3-5 180
Choi, Byoung Jin Tue-O-1-2-5 142
Choi, Heejin Tue-P-2-1-12 168
Choi, Iksoo Mon-O-2-1-6 121
Thu-P-2-3-6 258
Choi, Yeunju Tue-P-1-3-3 154
Chou, Ju-chieh Mon-O-2-4-6 125
Chou, Katherine Wed-P-3-3-8 226
Christensen, Heidi Wed-SS-1-1-2 182
Chua, Cathy Wed-SS-2-1-1 201
Chua, Zhi Ming Mon-SS-1-1-11 111
Chukharev-Hudilainen, E. Wed-O-3-4-6 219
Chung, Joon Son Thu-O-2-3-3 249
Thu-P-1-1-6 238
Tue-P-1-1-6 149
Chung, Pei-Hung Wed-P-1-2-1 188
Chung, Yu-An Mon-P-2-4-11 138
Cieri, Christopher Mon-P-1-1-4 113
Cleland, Joanne Wed-SS-1-1-1 182
Co, Chris Wed-P-3-3-8 226
Coates, Adam Mon-O-2-1-1 120
Cohen, Alex S. Thu-P-2-2-4 255
Cohen, Raphael Wed-S&T-1-1-4 185
Cole, Jennifer Wed-O-1-3-5 180
Collobert, Ronan Mon-P-2-4-5 136
Çöltekin, Çağri Tue-O-2-5-2 163
Conlan, Owen Mon-O-2-2-4 122
Cooke, Martin Wed-O-2-3-1 198
Cooper, Erica Wed-P-3-1-11 222
Correia, Joana Wed-P-3-3-6 226
Cortes, Elisabet Eir Tue-P-2-1-6 167
Costantini, Giovanni Thu-P-2-2-1 254
Cox, Felicity Tue-O-2-4-6 162
Creed, Torrey Wed-SS-1-1-5 183
Cristia, Alejandrina Tue-O-2-3-4 161
Wed-P-2-3-11 211
Wed-P-2-3-12 211
Crowgey, Erin Mon-O-2-3-4 123
Csapó, Tamás Gábor Thu-SS-1-1-8 235
Cui, Jia Mon-P-2-4-1 136
Mon-P-2-4-6 137
Cui, Jie Wed-P-2-3-6 210
Cui, Yang Wed-P-1-1-11 188
Cummins, Nicholas Mon-P-1-2-6 116
Thu-P-1-4-1 243
Wed-O-2-5-3 201
Wed-P-3-1-9 222
Curry, COL Ian Tue-P-2-5-8 176


## D


D, Govind Wed-P-1-3-12 193
D. S., Pavan Kumar Mon-P-1-2-11 116
Mon-SS-2-1-6 126
D.V.L.N, Dheeraj Sai Mon-P-2-5-6 139
d'Heureuse, Christian Tue-O-2-4-5 162
Dai, Lirong Thu-O-1-3-2 231
Thu-P-2-1-4 252
Wed-O-1-1-4 178
Dai, Li-Rong Wed-P-1-1-4 186
Wed-P-2-2-10 208
Dancette, Corentin Wed-P-2-4-13 214
Dandapat, S Mon-P-1-4-4 119
Mon-P-1-4-3 119
Dang, Jianwu Mon-P-2-2-7 132
Tue-O-2-5-1 163
Tue-P-2-2-11 170
Darcy, Dan Wed-P-1-1-2 186

Das Gupta, Mithun Tue-S&T-1-1-3 148
Tue-S&T-1-1-4 148
Das, Amit Wed-P-2-1-9 205
Das, Biswajit Mon-O-2-3-2 123
Das, Partha Pratim Mon-P-2-5-3 138
Das, Samarjit Thu-P-1-2-1 239
Das, Sneha Thu-O-2-4-1 250
Thu-O-2-4-2 250
Thu-O-2-4-3 250
Dasgupta, Hirak Wed-P-1-3-8 192
Dash, Debadatta Tue-SS-1-1-8 147
D'Ausilio, Alessandro Mon-P-2-1-2 128
Davis, Chris Wed-O-2-3-3 198
Wed-P-2-3-8 210
Dawalatabad, Nauman Tue-P-1-3-4 154
Dayalu, Praveen Wed-SS-1-1-3 182
de Mori, Renato Mon-O-1-1-5 105
de Wet, Febe Wed-P-2-4-2 212
Wed-SS-1-2-2 184
Debrah, Japhet Mon-P-1-1-4 113
Degottex, Gilles Tue-O-1-2-1 141
Dehak, Najim Mon-O-1-3-6 107
Mon-P-1-2-2 115
Mon-P-1-2-8 116
Thu-O-1-3-4 232
Tue-P-1-1-11 150
Tue-SS-2-1-6 165
Wed-O-1-2-6 179
Wed-P-1-2-5 189
Wed-P-2-4-8 213
Wed-SS-3-1-5 220
Deka, Abhash Thu-S&T-1-1-2 237
Deka, Barsha Thu-S&T-1-1-2 237
Delcroix, Marc Mon-O-1-1-1 104
Thu-O-1-1-5 229
Wed-P-2-1-11 206
Wed-P-2-1-2 204
Delgado, Héctor Mon-O-1-3-4 107
Wed-O-3-4-4 219
Wed-SS-3-1-6 220
Dellwo, Volker Thu-P-2-4-8 260
Tue-O-2-4-5 162
Delvaux, Veronique Mon-P-1-4-2 119
Demiroglu, Cenk Thu-P-1-4-3 244
Demuynck, Kris Mon-P-1-4-1 119
Thu-P-2-2-3 254
Deng, Liuhui Mon-O-2-1-2 120
Deng, Shiwen Thu-P-1-2-2 240
Thu-P-1-2-4 240
Deng, Yue Thu-P-1-3-5 242
Dentel, Laure Wed-O-2-3-5 199
Depra Cuozzo, Luis G. Wed-SS-3-1-7 220
DeRenzi, Brian Wed-P-2-4-6 213
Dernoncourt, Franck Mon-S&T-1-1-2 111
Desnous, Florent Tue-O-2-3-1 160
Desplanques, Brecht Thu-P-2-2-3 254
Detey, Sylvain Wed-P-2-3-3 209
Devnath, Sanjeevan Wed-S&T-2-1-5 203
Dey, Abhishek Thu-S&T-1-1-2 237
Tue-SS-1-1-6 147
Dey, Prasenjit Mon-P-2-2-12 133
Dey, Subhadeep Thu-P-2-1-8 252
Tue-P-1-1-9 150
Dhiman, Jitendra Kumar Mon-P-2-3-9 135
Diakoloukas, Vassilios Tue-O-1-5-1 145
Diard, Juien Thu-P-2-4-6 260
Di-Caterina, Gaetano Mon-O-2-3-1 122
Diener, Lorenz Thu-SS-1-1-6 234
Diez, Mireia Wed-SS-3-1-3 220
Dillon, Caitlin Mon-P-1-2-12 117
D'Imperio, Mariapaola Wed-P-3-4-1 227
Ding, Hongwei Mon-P-1-1-4 113
Ding, Shaojin Mon-O-2-4-1 124
Mon-O-2-4-2 124
Ding, Shuoyang Wed-P-2-1-1 204
Ding, Wenhao Thu-P-2-1-15 254
Ding, Yueming Thu-O-2-4-6 251
Djamali, Julia Thu-P-1-4-1 243
Djegdjiga, Amazouz Wed-O-3-3-6 218
Dmitriev, Evgeny Tue-P-1-1-8 149
Do Carmo Blanco, Noelia Wed-O-2-3-5 199
Wed-O-2-3-6 199
Do, Cong-Thanh Tue-P-2-2-7 170
Dollaghan, Christine Tue-SS-2-2-3 173
Dong, Fengquan Mon-SS-1-1-1 109
Dong, Fengquan Mon-SS-1-1-2 109
Mon-SS-1-1-4 109
Mon-SS-1-1-8 110
Mon-SS-2-1-1 125
Mon-SS-2-1-10 127
Dong, Linhao Mon-P-2-4-12 138
Mon-P-2-4-7 137
Dong, Qiusi Wed-P-2-3-13 211
Dorca, Georgina Wed-P-1-1-7 187
Dörfler, Oskar Wed-P-1-4-1 193
Dötsch, Patrick Mon-P-2-4-2 136
Drory, Tal Mon-S&T-2-1-3 127
Drossos, Konstantinos Mon-P-1-3-7 118
Drude, Lukas Thu-O-1-1-5 229
Du, Jun Thu-P-1-1-11 239
Wed-SS-3-1-2 219
Du, Mingxing Wed-O-3-1-4 215
Du, Yufan Tue-O-1-4-2 144
Du, Yulun Mon-P-2-1-12 130
Duan, Pengfei Thu-O-2-4-5 251
Thu-O-2-4-6 251
Dubey, Akhilesh Kumar Mon-P-1-4-4 119
Dubey, Harishchandra Thu-P-2-1-9 253
Dumpala, Sri Harsha Tue-P-2-5-7 176
Dupoux, Emmanuel Mon-P-2-4-5 136
Wed-O-3-1-4 215
Wed-P-2-4-13 214
Dyrenko, Natalia Wed-P-2-3-7 210

**E**

Earnshaw, Kate Wed-O-3-3-4 217
Wed-O-3-3-5 217
Ebbers, Janek Wed-O-3-1-5 215
Edler, Bernd Tue-O-2-1-2 158
Edwards, Erik Thu-S&T-1-1-1 236
Egorova, Ekaterina Thu-SS-1-2-3 235
Ehr, Heiko Thu-O-2-3-4 249
Einspieler, Christa Mon-SS-1-1-1 109
Mon-SS-1-1-2 109
Mon-SS-1-1-4 109
Mon-SS-1-1-8 110
Mon-SS-2-1-1 125
Mon-SS-2-1-10 127
Ekon, Nicole Mon-P-1-2-12 117
Elg, Aija Tue-S&T-1-1-1 148
Ellis, Daniel P. W. Tue-P-1-3-9 155
Elvevåg, Brita Thu-P-2-2-4 255
Emiru, Eshete Derb Thu-O-2-4-5 251
Endo, Hajime Wed-P-2-2-2 207
Enrique Yalta Soplin, N. Wed-O-2-1-1 196
Epps, Julien Mon-P-1-2-4 115
Mon-P-2-2-3 131
Mon-P-2-2-8 132
Thu-O-1-3-6 232
Thu-P-1-4-2 244
Thu-P-2-2-7 255
Erdogan, Hakan Thu-O-1-1-4 229
Thu-O-2-2-6 248
Erickson, Donna Wed-P-1-4-10 195
Ernestus, Mirjam Tue-O-1-4-4 144
Erzin, Engin Mon-SS-1-1-5 110
Tue-P-2-5-5 176
Escudero, Paola Tue-O-2-5-4 163
Eshky, Aciel Wed-SS-1-1-1 182
Eskenazi, Maxine Mon-P-2-1-12 130
Espic, Felipe Wed-P-1-1-12 188
Espy-Wilson, Carol Thu-O-1-4-6 233
Thu-P-2-2-12 256
Estève, Yannick Tue-P-1-4-8 157
Wed-P-2-1-5 204
Wed-P-2-2-6 207
Evanini, Keelan Mon-S&T-2-1-2 127
Wed-S&T-1-1-2 185
Wed-SS-2-1-6 202
Evans, Nicholas Mon-O-1-3-4 107
Tue-P-1-2-11 153
Wed-O-3-4-4 219
Wed-SS-3-1-6 220

**F**

Fadiga, Luciano Mon-P-2-1-2 128
Fainberg, Joachim Tue-O-1-1-1 140
Falk, Tiago Thu-P-2-2-6 255
Fang, Qiang Thu-O-1-4-2 232
Farah, Chadi Wed-P-3-4-1 227
Farinas, Jérôme Wed-P-3-3-4 225
Farrús, Mireia Tue-S&T-2-1-4 166
Fathima, Noor Thu-SS-1-2-6 236
Tue-SS-1-1-5 147
Wed-S&T-2-1-3 203
Faytak, Matthew Mon-P-1-1-6 113
Fedotov, Dmitrii Mon-SS-2-1-5 126
Fels, Sidney Tue-P-1-3-11 155
Feng, Gang Wed-P-2-4-10 213
Feng, Siyuan Thu-P-1-4-7 245
Wed-O-3-1-2 215
Wed-P-2-1-10 205
Feng, Zhe Mon-SS-1-1-3 109
Fernandez, Raul Wed-P-2-1-4 204
Fernando, Sarith Wed-O-1-2-5 179
Ferrer, Luciana Mon-O-1-3-5 107
Fezza, Nassima Wed-P-3-4-2 227
Fingscheidt, Tim Mon-P-1-2-5 115
Finkelstein, Katherine Mon-P-1-2-12 117
Finley, Gregory Thu-S&T-1-1-1 236
Fischer, Johannes Thu-O-2-4-1 250
Flemotomos, Nikolaos Thu-P-1-4-6 245
Tue-O-2-3-3 161
Wed-SS-1-1-5 183
Flint, Jonathan Tue-SS-2-2-2 173
Fohr, Dominique Wed-O-3-2-2 216
Foltz, Peter W. Thu-P-2-2-4 255
Fone, James Thu-S&T-1-1-1 236
Fontan, Lionel Wed-P-2-3-3 209
Fookes, Clinton Thu-P-2-1-7 252
Foulkes, Paul Mon-P-1-1-11 114
Franco, Horacio Tue-P-2-2-2 169
Wed-P-3-3-5 226
Frassetto, Tommaso Tue-P-1-4-11 157
Frederiksen, Peter S. Wed-O-1-2-6 179
Fredouille, Corinne Wed-P-3-3-2 225
Freimoser, Cem Philipp Wed-SS-2-1-2 202
Freitas, João Wed-S&T-1-1-3 185
French, Peter Mon-P-1-1-11 114
Friedrichs, Daniel Tue-O-2-4-5 162
Fringi, Eva Wed-P-2-3-9 211
Fu, Ruibo Tue-O-1-2-3 141
Tue-O-1-2-6 142
Wed-P-2-2-11 208
Fu, Szu-wei Wed-O-1-4-4 181
Fuchs, Robert Wed-P-2-3-7 210
Fujisaki, Hiroya Tue-P-2-5-2 175
Fujita, Yusuke Wed-P-3-2-9 224
Fukuda, Takashi Wed-P-2-1-4 204

**G**

G, Gayathri Mon-S&T-2-1-4 128
G, Kasthuri Tue-S&T-2-1-1 165
Gabbay, Aviv Tue-P-1-2-8 152
Gafos, Adamantios Tue-O-2-4-2 162
Tue-O-2-5-4 163
Gaillard, Pascal Wed-P-3-3-4 225
Gales, Mark J.F. Tue-O-1-2-1 141
Gales, Mark Thu-P-1-3-2 242
Tue-O-1-1-2 140
Tue-P-2-3-2 171
Tue-P-2-3-5 172
Wed-O-2-1-3 196
Gallagher, Andrew Tue-P-1-3-9 155
Galvez, Rodrigo Wed-O-3-4-5 219
Gamble, Paul Thu-O-2-2-1 247
Tue-P-2-2-2 169
Ganapathy, Sriram Thu-O-1-1-1 229
Thu-Perspective-4 246
Tue-P-1-1-13 151
Tue-P-1-1-7 149
Wed-P-2-3-11 211
Wed-P-3-2-6 224
Gandhe, Ankur Thu-P-1-3-1 241
Gandhi, Rushabh Tue-S&T-1-1-3 148
Ganesh, Shobhana Wed-P-2-3-11 211
Gang, Arpita Thu-O-2-2-5 248
Gangashetty, S. V. Mon-P-1-1-1 112
Ganji, Sreeram Wed-SS-1-2-8 185
Gao, Guanglai Mon-O-1-2-6 106
Gao, Jie Wed-P-2-4-14 214
Gao, Yi Wed-P-2-4-4 212
Gao, Zhifu Thu-P-2-1-4 252
Garcia, Nicanor Wed-O-2-5-6 201
Garcia-Romero, Daniel Mon-O-1-3-3 107

Wed-SS-3-1-5 220
Garg, Saurabh Wed-P-2-4-1 212
Garner, Philip N. Thu-SS-1-1-3 234
Wed-P-2-1-14 206
Garudadri, Harinath Tue-P-1-2-10 153
Gau, Susan Shur-Fen Wed-O-2-5-2 200
Gaudrain, Etienne Wed-O-2-3-2 198
Gautam, Digvijaysingh Thu-P-1-2-7 240
Ge, Fengpei Wed-P-3-2-2 223
Ge, Hancheng Mon-P-2-1-9 129
Geng, Puyang Tue-P-2-5-2 175
Geng, Tianzhu Wed-P-1-4-9 194
Georgiou, Panayiotis Thu-P-1-4-8 245
Tue-O-2-3-2 161
Tue-P-1-1-15 151
Wed-O-2-5-4 201
Gerazov, Branislav Wed-P-2-2-13 208
Gerlach, Johanna Wed-SS-2-1-1 201
Ghaffarzadegan, S. Mon-SS-1-1-3 109
Ghahremani, Pegah Mon-P-1-2-8 116
Thu-O-1-3-4 232
Tue-P-2-2-8 170
Ghannay, Sahar Tue-P-1-4-8 157
Ghassemi, Mohammad Tue-SS-2-2-10 175
Ghio, Alain Wed-P-3-3-2 225
Ghorbani, Shahram Wed-P-2-1-12 206
Ghosal, Deepanway Wed-P-1-2-12 190
Ghosh, Debayan Tue-P-1-3-5 154
Ghosh, Prasanta Kumar Mon-O-2-4-4 124
Mon-P-2-5-9 139
Mon-S&T-2-1-1 127
Thu-O-1-4-3 233
Thu-O-1-4-4 233
Thu-O-1-4-5 233
Thu-P-1-1-8 238
Tue-P-2-1-10 168
Wed-P-3-4-4 227
Wed-P-3-4-7 228
Wed-S&T-2-1-4 203
Gibson, James Thu-P-1-4-6 245
Tue-O-2-3-3 161
Wed-SS-1-1-5 183
Gideon, John Wed-SS-1-1-4 183
Gimeno, Pablo Wed-SS-3-1-4 220
Girau, Bernard Wed-O-3-4-5 219
Girija Ramesan, K. Mon-P-2-5-9 139
Glarner, Thomas Wed-O-3-1-5 215
Glass, James Mon-P-2-4-11 138
Tue-P-2-2-4 169
Tue-SS-2-1-2 164
Tue-SS-2-2-10 175
Wed-P-3-2-10 224
Glembek, Ondřej Wed-SS-3-1-3 220
Gobl, Christer Mon-P-1-1-10 114
Mon-P-1-1-9 114
Godard, Pierre Wed-O-3-1-3 215
Goehner, Kyle Tue-P-1-3-6 154
Goel, Nagendra Kumar Thu-O-1-3-4 232
Goel, Nagendra Wed-S&T-2-1-1 203
Gogate, Mandar Wed-O-3-2-6 216
Gogoi, Pamir Mon-P-1-1-13 114
Gogoi, Parismita Mon-P-1-1-13 114
Tue-SS-1-1-6 147
Gold, Erica Wed-O-1-3-1 180
Wed-O-3-3-4 217
Wed-O-3-3-5 217
Goldberger, Jacob Wed-O-1-2-3 179
Goldwater, Sharon Tue-P-1-4-10 157
Wed-O-3-1-1 214
Gómez Alanís, A. Mon-P-2-2-11 132
Gomez, Angel Mon-P-2-2-11 132
Gómez, Juan C. Wed-O-3-4-5 219
Gong, Rong Mon-P-2-3-5 134
Gong, Yifan Thu-P-1-1-9 238
Tue-P-1-2-7 152
Wed-O-1-1-1 177
Wed-O-2-1-1 195
Gong, Yuan Wed-O-3-2-1 216
Gonzalez, Jose A. Mon-P-2-2-11 132
Gopikishore, Pebbili Wed-P-3-4-7 228
Gorrostieta, Cristina Mon-SS-2-1-2 125
Goswami, Nabarun Wed-O-3-2-4 216
Gosztolya, Gábor Mon-SS-2-1-7 126
Gosztolya, Gábor Thu-P-1-4-5 244
Thu-SS-1-1-8 235
Wed-P-2-4-7 213
Govender, Avashna Wed-P-3-1-4 221
Wed-P-3-1-5 221
Grabek, Anna Tue-P-2-3-4 171
Graciarena, Martin Tue-P-1-1-10 150
Tue-P-2-2-2 169
Green, Jordan Tue-SS-2-2-9 175
Wed-SS-1-1-6 183
Green, Phil Thu-P-1-4-10 245
Greenberg, Craig Wed-O-1-2-1 178
Greenwood, David Wed-P-2-2-5 207
Greer, Timothy Wed-O-2-5-5 201
Grézl, František Wed-P-2-4-9 213
Grondin, François Wed-P-3-2-10 224
Grósz, Tamás Mon-SS-2-1-7 126
Thu-SS-1-1-8 235
Gu, Wentao Tue-P-2-5-2 175
Gu, Yu Wed-P-1-1-9 187
Guan, Haotian Mon-P-2-2-7 132
Tue-P-2-2-11 170
Guarino Reid, Loretta Tue-P-1-3-9 155
Guerry, Marine Wed-P-1-4-10 195
Guha, Tanaya Tue-O-1-3-5 143
Gui, Xiangquan Tue-SS-2-2-6 174
Gunendradasan, T. Mon-P-2-2-3 131
Guo, Jinxi Thu-P-2-3-3 257
Tue-SS-2-2-2 173
Guo, Lili Tue-P-2-2-11 170
Guo, Michelle Wed-O-2-4-1 199
Guo, Pengcheng Wed-SS-1-2-3 184
Guo, Wu Mon-O-1-3-1 106
Thu-O-1-3-2 231
Thu-P-2-1-4 252
Wed-P-3-1-7 222
Gupta, Arpit Mon-P-2-1-9 129
Gupta, Chitralekha Tue-P-2-1-3 166
Gupta, Raghav Thu-O-2-1-6 247
Gupta, Rahul Thu-P-2-2-12 256
Gupta, Sanjay Kumar Wed-P-3-4-6 228
Gupta, Suyog Mon-O-1-1-6 105
Gupta, Vishwa Thu-P-2-1-6 252
Wed-P-2-4-12 214
Gurugopinath, Sanjeev Tue-P-1-3-5 154
Gurugubelli, Krishna Thu-SS-1-2-5 236
Gutierrez-Osuna, Ricardo Mon-O-2-4-1 124
Mon-O-2-4-2 124
Wed-O-3-4-6 219
Guzewich, Peter Tue-O-2-1-5 159
Guzmán, Gualberto Wed-P-2-3-1 209
Gweon, Gahgene Tue-P-2-3-6 172

## H

H R, Shiva Kumar Wed-S&T-1-1-5 185
Ha, Thanh-Le Tue-P-1-4-9 157
Habets, Emanuël Tue-O-2-1-2 158
Hadian, Hossein Mon-O-1-1-3 104
Tue-P-2-2-8 170
Haeb-Umbach, Reinhold Thu-O-1-1-5 229
Wed-O-3-1-5 215
Haghani, Parisa Tue-O-1-1-6 141
Hahn, Minsoo Tue-P-2-1-12 168
Haider, Adnan Wed-P-3-2-8 224
Haider, Fasih Mon-O-2-2-4 122
Tue-P-2-5-4 176
Hain, Thomas Mon-P-2-3-1 133
Tue-P-2-3-10 173
Hakkani-Tür, Dilek Thu-O-2-1-6 247
Tue-Perspective-1 158
Han, Jing Mon-SS-1-1-7 110
Thu-O-1-3-1 231
Han, Jiqing Thu-P-1-2-2 240
Thu-P-1-2-4 240
Han, Kyu Mon-P-2-4-8 137
Han, Wenjing Tue-O-1-3-2 142
Hanani, Abualsoud Wed-SS-2-1-4 202
Hanebrink, Patrick Wed-O-3-1-5 215
Hannemann, Mirko Mon-P-2-4-2 136
Hansen, John H.L. Mon-P-2-3-11 135
Thu-P-2-1-9 253
Tue-P-1-1-3 149
Tue-SS-2-2-3 173
Tue-SS-2-2-4 174
Wed-O-3-4-1 218
Hansen, John H.L. Wed-P-1-3-5 191
Wed-P-2-1-12 206
Hantke, Simone Mon-SS-1-1-1 109
Mon-SS-1-1-2 109
Mon-SS-1-1-4 109
Mon-SS-1-1-8 110
Mon-SS-2-1-1 125
Mon-SS-2-1-10 127
Thu-O-2-3-1 249
Wed-P-3-1-9 222
Hao, Lixia Wed-P-3-4-9 228
Haque, Albert Wed-O-2-4-1 199
Hara, Kohei Tue-O-1-5-2 145
Hara, Sunao Wed-P-2-2-1 206
Harish, Madhumita Thu-O-1-1-1 229
Harman, Craig Wed-P-1-2-5 189
Harmegnies, Bernard Mon-P-1-4-2 119
Harrison, Philip Mon-P-1-1-11 114
Harte, Naomi Mon-P-2-1-7 129
Harvey, Richard Wed-P-2-4-11 214
Hasan, Madina Thu-P-1-4-10 245
Hasan, Taufiq Mon-SS-1-1-3 109
Hasegawa-Johnson, M. Mon-P-1-2-1 115
Thu-O-2-3-5 249
Tue-SS-2-1-6 165
Wed-P-1-2-4 189
Wed-P-2-1-9 205
Wed-P-2-4-5 212
Hayamizu, Satoru Wed-P-2-2-2 207
Hayashi, Tomoki Mon-P-2-4-9 137
Wed-O-2-1-1 196
Wed-P-1-1-5 186
Haynes, Rob Wed-P-1-2-13 190
He, Di Wed-P-2-4-5 212
He, Lei Thu-P-2-4-8 260
Wed-P-1-1-11 188
He, Liang Thu-P-2-1-15 254
Wed-O-2-2-3 197
He, Qianhua Mon-P-2-2-6 131
He, Shan Wed-SS-3-1-2 219
He, Weipeng Mon-P-1-3-3 117
Heaton, Kristin Mon-P-1-2-12 117
Tue-P-2-5-8 176
Hebbar, Rajat Mon-P-1-2-9 116
Hedayatnia, Behnam Thu-P-1-3-1 241
Hedelin, Per Thu-O-2-4-4 250
Heitzman, Daragh Wed-SS-1-1-6 183
Heldner, Mattias Tue-O-2-4-3 162
Helmke, Hartmut Thu-O-2-3-4 249
Heo, Hee-soo Thu-P-2-1-5 252
Hermann, Enno Wed-O-3-1-1 214
Hermansky, Hynek Thu-O-1-1-3 229
Hermes, Anne Mon-P-1-1-5 113
Tue-P-2-1-7 167
Hernandez, Nidia Wed-O-3-3-1 217
Hernandez-Cordero, J. Wed-O-1-2-1 178
Hershey, John Wed-O-3-2-3 216
Hetherly, Jeffrey Thu-O-2-2-1 247
Tue-P-2-2-2 169
Heymann, Jahn Thu-O-1-1-5 229
Wed-O-2-1-1 196
Hilt, Pauline Mon-P-2-1-2 128
Himawan, Ivan Thu-P-2-1-7 252
Wed-P-1-3-3 191
Hirose, Keikichi Wed-P-3-4-8 228
Hirschberg, Julia Mon-O-2-2-1 121
Mon-O-2-2-2 121
Thu-O-1-3-3 232
Wed-P-3-1-11 222
Wed-SS-1-2-5 184
Hirschfeld, Diane Mon-P-2-1-8 129
Hitschfeld-Kahler, Nancy Wed-O-3-4-5 219
Hodari, Zack Mon-O-1-2-1 105
Hoen, Michel Wed-O-2-3-6 199
Hoffmann, Ildikó Thu-P-1-4-5 244
Hoffmeister, Björn Tue-P-1-3-6 154
Holmlund, Terje B. Thu-P-2-2-4 255
Holzenberger, Nils Wed-O-3-1-4 215
Honda, Kiyoshi Thu-O-1-4-2 232
Hönemann, Angelika Mon-O-2-2-3 121
Hong, Qian-Bei Tue-O-1-5-5 145
Hong, Yoon Seok Tue-P-2-3-6 172
Hoory, Ron Mon-S&T-2-1-3 127
Wed-P-3-1-10 222
Horgan, Michael Wed-P-1-1-2 186
Hori, Takaaki Wed-O-2-1-1 196
Horio, Kento Wed-P-2-2-2 207

Hosseini-Asl, Ehsan Thu-P-2-3-12 259
Houde, John Tue-P-2-1-1 166
Howson, Phil Mon-P-1-1-2 112
Hrúz, Marek Thu-O-2-3-6 250
Wed-SS-3-1-1 219
Hsu, Wei-Ning Tue-P-2-2-4 169
Tue-SS-2-1-2 164
Wed-P-3-2-10 224
Hu, Chih Chi Thu-P-1-3-7 242
Hu, Fang Mon-P-1-1-7 113
Tue-O-2-4-1 162
Hu, Shoukang Wed-O-1-1-3 177
Wed-P-3-3-1 225
Hu, Wenping Tue-P-2-3-3 171
Hua, Kanru Mon-P-1-3-8 118
Huang, Che-Wei Thu-P-2-2-5 255
Huang, Dongyan Mon-SS-1-1-10 111
Huang, Hen-Hsen Tue-P-2-5-1 175
Huang, Huai-Hung Tue-O-1-5-5 145
Huang, Jian Thu-P-2-2-8 255
Huang, Jinfeng Tue-O-2-5-1 163
Huang, Kun-Yi Tue-O-1-5-5 145
Huang, Lu Wed-O-1-1-6 178
Huang, Mingkun Thu-P-2-3-1 257
Huang, Zhaocheng Thu-P-1-4-2 244
Huang, Zhongqiang Wed-P-1-2-5 189
Huang, Zili Thu-P-2-1-13 253
Huber, Adrian Mon-P-2-5-4 138
Huber, Rainer Tue-O-1-4-5 144
Tue-O-1-4-6 144
Huckvale, Mark Mon-SS-1-1-6 110
Hueber, Thomas Wed-P-2-4-10 213
Huet, Kathy Mon-P-1-4-2 119
Huet, Moïra-Phoebé Wed-O-2-3-2 198
Hughes, Vincent Mon-P-1-1-11 114
Hullebus, Marc Antony Tue-O-2-4-2 162
Humayun, Ahmed Imtiaz Mon-SS-1-1-3 109
Hunger-Schoppe, C. Thu-P-1-4-1 243
Hussain, Amir Wed-O-3-2-6 217
Hussein, Hussein Wed-P-1-4-2 193
Hwang, Hsin-Te Mon-O-2-4-3 124
Wed-O-1-4-4 181
Hwang, Meiyuh Thu-O-1-1-6 230
Hwang, Mei-Yuh Wed-P-2-1-3 204
Hwang, Min-Jae Tue-O-1-2-4 142

## I

Ikeda, Manabu Tue-SS-2-2-5 174
Illa, Aravind Thu-O-1-4-3 233
Imani, Siddika Thu-S&T-1-1-2 237
Inez Watson, Catherine Wed-O-3-4-3 218
Inoue, Koji Mon-P-2-1-13 130
Tue-O-1-5-2 145
Inoue, Nakamasa Thu-P-2-1-11 253
Tue-SS-2-2-8 174
Inoue, Yusuke Tue-P-2-3-7 172
Ioannis Mimilakis, S. Mon-P-1-3-7 118
Iqbal, Zikra Wed-S&T-2-1-1 203
Irie, Kazuki Mon-O-1-1-2 104
Mon-O-2-1-2 120
Irino, Toshio Wed-O-1-4-2 181
Wed-P-1-1-13 188
Irtza, Saad Mon-P-2-2-8 132
Isei-Jaakkola, Toshiko Wed-P-3-4-8 228
Ito, Akinori Tue-P-2-5-6 176
Itoh, Yoshiaki Wed-P-1-2-7 189
Iwasawa, Yusuke Thu-O-1-2-4 230
Iwata, Tomoharu Mon-O-1-1-1 104
Iyengar, Anuroop Thu-SS-1-2-6 236
Tue-SS-1-1-5 147
Wed-S&T-2-1-3 203

## J

J H Simons, Anthony Thu-P-1-4-10 245
Jacob, Namita Tue-S&T-2-1-1 165
Jain, Abhinav Wed-P-2-1-13 206
Jain, Mohit Mon-P-2-2-12 133
Jaiswal, Mimansa Wed-SS-1-1-4 183
Jaitly, Navdeep Wed-P-3-3-8 226
James, Jesin Wed-O-3-4-3 218
Jančovič, Peter Thu-P-2-3-11 259
Tue-SS-2-1-4 165
Wed-SS-2-1-5 202
Jati, Arindam Tue-P-1-1-15 151
Jauk, Igor Thu-O-1-2-3 230
Jaunzeikare, Diana Wed-P-1-2-10 190
Wed-P-3-3-8 226
Jayagopi, Dinesh Babu Tue-SS-1-1-2 146
Jayaraman, Rajagopal Tue-S&T-1-1-2 148
Tue-S&T-1-1-4 148
Jelil, Sarfaraz Mon-P-2-2-2 131
Ji, Jim X. Tue-SS-2-2-1 173
Ji, Ruifang Thu-P-2-1-14 254
Ji, Xuan Wed-P-2-4-4 212
Jia, Jia Thu-P-2-2-10 256
Jia, Yuan Wed-P-1-1-4 186
Wed-P-1-4-4 193
Jian, Hu Tue-P-2-1-13 168
Jiang, Chao Wed-SS-3-1-2 219
Jiang, Yuan Wed-P-1-1-4 186
Jiang, Yue Mon-P-1-1-4 113
Jin, Hongxia Thu-O-2-1-1 246
Thu-O-2-1-3 247
Thu-O-2-1-5 247
Thu-P-1-3-5 242
Jin, Wenyu Wed-SS-1-1-3 182
Joachim, Dale Thu-P-1-4-2 244
Joglekar, Aditya Wed-O-3-4-1 218
Johnson, Michael T. Wed-O-2-2-3 197
Joshi, Pankaj Thu-P-1-2-7 240
JPrakash, Jeena Thu-SS-1-2-2 235
Juang, Biing-Hwang (Fred) Thu-P-1-1-9 238
Tue-P-1-2-7 152
Jülg, Dominik Wed-SS-2-1-2 202
Julka, Sahib Wed-O-2-5-3 201
Jun, Heewoo Mon-O-2-1-1 120
Jung, Jee-weon Thu-P-2-1-5 252
Jung, Youngmoon Tue-P-1-3-3 154
Juvela, Lauri Mon-P-2-3-2 133
Wed-P-1-1-10 187
Jyothi, Preethi Thu-P-1-2-7 240
Wed-P-2-1-13 206
Wed-P-2-4-1 212

## K

K Abraham, Ajish Mon-P-1-4-6 120
K E, Manjunath Tue-SS-1-1-2 146
K S, Girish Mon-P-1-4-6 120
K T, Raseena Thu-O-1-4-4 233
K. M., Srinivasa R. Tue-SS-1-1-7 147
K., Gopakumar Mon-S&T-1-1-4 112
K.P, Soman Wed-P-1-3-12 193
K.S, Kishor Mon-P-2-5-6 139
Kabashima, Suguru Tue-P-2-3-7 172
Kabil, Selen Hande Mon-P-1-2-10 116
Kaburagi, Tokihiko Wed-P-2-2-8 208
Kadiri, Sudarsana Reddy Mon-O-2-2-6 122
Mon-P-1-1-1 112
Mon-P-1-1-12 114
Wed-O-2-4-2 199
Wed-O-2-4-6 200
Kalita, Sishir Mon-P-1-1-13 114
Mon-P-1-4-3 119
Mon-P-2-2-2 131
Tue-P-2-1-9 167
Tue-SS-2-2-7 174
Kallio, Heini Wed-O-1-3-3 180
Kamble, Madhu Mon-P-2-2-4 131
Mon-P-2-2-5 131
Mon-P-2-2-9 132
Thu-SS-1-2-4 235
Kamiyama, Hosana Tue-P-2-5-3 175
Kamper, Herman Tue-P-1-4-10 157
Wed-P-2-4-3 212
Kanamura, Kumi Tue-P-2-3-7 172
Kanda, Naoyuki Wed-P-3-2-9 224
Kane, John Mon-SS-2-1-2 125
Kang, Hong-Goo Tue-O-1-2-4 142
Kang, Shiyin Thu-O-1-2-5 231
Kang, Yongguo Thu-O-1-2-6 231
Wed-P-1-1-9 187
Kang, Yoonjung Tue-P-2-1-8 167
Kannan, Anjuli Wed-O-2-1-5 196
Wed-P-1-2-10 190
Wed-P-3-3-8 226
Kao, Chieh-Chi Tue-O-2-2-2 159
Tue-O-2-2-5 160
Kar, Sibsambhu Thu-P-2-2-11 256
Karadayi, Julien Tue-O-2-3-4 161
Wed-O-3-1-4 215
Wed-P-2-4-13 214
Karafiát, Martin Thu-SS-1-2-3 235
Wed-P-2-4-9 213
Karhila, Reima Tue-S&T-1-1-1 148
Karisma, Sonya Wed-O-2-3-3 198
Karita, Shigeki Mon-O-1-1-1 104
Wed-O-2-1-1 196
Wed-P-2-1-11 206
Karjol, Pavan Thu-P-1-1-8 238
Karki, Bhavya Mon-SS-2-1-4 125
Karpov, Alexey Mon-SS-2-1-5 126
Kathiresan, Thayabaran Thu-P-2-4-8 260
Tue-O-2-4-5 162
Kato, Akihiro Mon-P-1-3-6 118
Kato, Tsuneo Wed-P-1-4-7 194
Katz, William Tue-P-2-1-2 166
Kaushik, Lakshmish Wed-O-3-4-1 218
Kavanagh, Colleen Mon-P-1-1-11 114
Kaver, Liat Tue-P-1-3-9 155
Kawahara, Hideki Wed-P-1-1-13 188
Kawahara, Tatsuya Mon-P-2-1-13 130
Thu-P-2-3-2 257
Tue-O-1-5-2 145
Wed-O-2-1-6 196
Wed-P-2-1-7 205
Kawai, Hisashi Thu-P-2-3-2 257
Tue-O-2-2-4 160
Wed-O-1-2-4 179
Wed-P-3-1-1 221
Kaya, Heysem Mon-SS-2-1-5 126
Kazimirova, Evdokia Tue-O-2-3-5 161
Kazui, Hiroaki Tue-SS-2-2-5 174
Keller, Orgad Wed-S&T-1-1-4 185
Kelley, Matthew C. Tue-P-1-3-2 154
Kenefick, Robert Mon-P-1-2-12 117
Kenny, Patrick Thu-P-2-1-6 252
Kenyon, Robert V. Wed-O-1-3-6 181
Kesiraju, Santosh Thu-P-1-3-11 243
Thu-SS-1-2-3 235
Khan, Md. Tauhiduzzaman Mon-SS-1-1-3 109
Khan, Soma Tue-S&T-2-1-3 165
Khassanov, Yerbolat Thu-P-1-3-3 242
Khatri, Chandra Thu-P-1-3-1 241
Kheyrkhah, Timothee Wed-O-1-2-1 178
Khokhlov, Yuri Wed-P-2-1-5 204
Wed-P-3-2-5 224
Khorram, Soheil Wed-SS-1-1-4 183
Khoury, Elie Thu-S&T-1-1-4 237
Khudanpur, Sanjeev Mon-O-1-1-3 104
Mon-P-1-2-8 116
Thu-P-1-3-9 243
Thu-P-2-3-9 258
Tue-P-2-2-8 170
Wed-O-1-1-6 178
Wed-O-2-1-2 196
Wed-P-1-2-5 189
Wed-SS-3-1-5 220
Ki, Kyung Seo Tue-P-2-3-6 172
Kim, Chanwoo Thu-O-1-1-2 229
Kim, Dae-Shik Mon-P-2-4-4 136
Kim, Hoirin Tue-P-1-3-3 154
Kim, Jeesun Wed-O-2-3-3 198
Wed-P-2-3-8 210
Kim, Jin-Seob Tue-O-1-2-4 142
Kim, John Tue-O-1-3-3 143
Kim, Jong-Jin Tue-P-2-1-12 168
Kim, Joo-Kyung Mon-P-2-1-1 128
Kim, Jungsuk Mon-P-2-4-8 137
Kim, Juntae Tue-P-2-1-12 168
Kim, Myungjong Thu-SS-1-1-4 234
Tue-SS-1-1-8 147
Wed-P-3-3-3 225
Wed-SS-1-1-6 183
Kim, Nam Soo Tue-O-1-2-5 142
Kim, Namju Thu-P-2-1-1 251
Kim, Sangjin Tue-P-2-1-12 168
Kim, So Hyun Wed-O-3-4-2 218
Kim, Suyoun Wed-P-3-2-7 224
Kim, Taehwan Wed-P-3-1-3 221
Kim, Yeon-Jun Wed-P-3-2-11 225
Kim, Young-Bum Mon-P-2-1-1 128
Kim, Younggwan Tue-P-1-3-3 154
Kim, Yulia Tue-P-1-1-8 149
King, Simon Mon-O-1-2-1 105
Wed-O-2-3-1 198

Wed-P-1-1-12 188
Wed-P-3-1-4 221
Wed-P-3-1-5 221
Kinnunen, Tomi Mon-O-1-3-4 107
Mon-P-1-3-6 118
Kinoshita, Keisuke Thu-O-1-1-5 229
Wed-O-1-4-2 181
Kirkedal, Andreas Søeborg Wed-P-3-2-11 225
Kiselev, Ilya Mon-P-2-5-4 138
Kitza, Markus Tue-O-1-1-3 140
Kleinert, Matthias Thu-O-2-3-4 249
Klejch, Ondřej Tue-O-1-1-1 140
Knill, Kate Tue-P-2-3-2 171
Knill, Kate Tue-P-2-3-5 172
Ko, Ming-Ya Mon-SS-2-1-8 126
Ko, Tom Thu-P-2-1-3 251
Wed-P-3-3-7 226
Kobashikawa, Satoshi Tue-P-2-5-3 175
Kobayashi, Kazuhiro Wed-P-1-1-5 186
Kobayashi, Yosuke Wed-O-1-4-1 181
Kocharov, Daniil Wed-P-1-4-3 193
Kochetov, Alexei Mon-P-1-1-2 112
Mon-P-1-1-6 113
Tue-P-2-1-8 167
Kodrasi, Ina Tue-O-2-1-3 159
Kodukula, Sri Rama Murty Mon-P-2-5-6 139
Köhn, Arne Mon-S&T-1-1-1 111
Wed-P-1-4-1 193
Kolekar, Maheshkumar H. Wed-P-1-2-12 190
Koller, Alexander Mon-S&T-1-1-1 111
Kollmeier, Birger Tue-O-1-4-6 144
Kolly, Marie-José Wed-O-3-3-3 217
Kondo, Kazuhiro Wed-O-1-4-1 181
Kong, Anthony Pak Hin Thu-P-1-4-7 245
Konopnicki, David Wed-P-3-1-10 222
Kons, Zvi Mon-S&T-2-1-3 127
Koolagudi, Shashidhar G Thu-P-1-2-9 241
Kopparapu, Sunil Kumar Mon-O-2-3-2 123
Mon-O-2-3-6 123
Tue-P-2-5-7 176
Koreman, Jacques Wed-P-2-3-10 211
Koshinaka, Takafumi Wed-O-2-2-4 197
Kothalkar, Prasanna Tue-SS-2-2-3 173
Kothinti, Sandeep Reddy Tue-O-1-1-4 140
Kotti, Margarita Tue-O-1-5-1 145
Kozlov, Alexandr Wed-O-2-2-2 197
Kraljevski, Ivan Mon-P-2-1-8 129
Kranzusch, Paul Tue-O-1-4-6 144
Kremnev, Ivan Wed-O-2-2-2 197
Krikke, Teun Tue-P-1-3-8 155
Krishna, Hari Thu-SS-1-2-5 236
Krishnamurthy, Rahul Wed-P-3-4-7 228
Krüger, Melanie Tue-O-1-4-6 144
Kucza, Maren Wed-P-1-2-9 190
Kudo, Takashi Tue-SS-2-2-5 174
Kulkarni, Purva Mon-P-1-2-2 115
Kumar, Abhay Thu-P-2-2-11 256
Kumar, Anurendra Thu-O-2-2-4 248
Kumar, Avinash Mon-P-2-3-13 136
Tue-P-2-2-9 170
Kumar, Deepak Tue-SS-1-1-5 147
Wed-S&T-2-1-3 203
Kumar, Kanhaiya Tue-S&T-2-1-2 165
Wed-SS-1-2-1 183
Kumar, Manoj Mon-P-1-2-11 116
Wed-O-3-4-2 218
Kumar, Mari Ganesh Tue-S&T-1-1-5 148
Kumar, Rajath Mon-P-1-3-10 119
Tue-P-1-1-13 151
Kumar, Vivek Wed-P-1-1-2 186
Kunešová, Marie Wed-SS-3-1-1 219
Kuniyoshi, Yasuo Wed-P-2-2-9 208
Kunstek, Mario Wed-SS-2-1-2 202
Kurata, Gakuto Wed-P-2-1-4 204
Wed-P-3-2-3 223
Kuriakose, Jom Tue-P-1-3-4 154
Kurimo, Mikko Tue-S&T-1-1-1 148
Kushwah, Tejendra Wed-S&T-2-1-1 203
Kyaw, Win Thuzar Wed-P-2-3-4 210
Kyaw, Zin Tun Mon-S&T-2-1-5 128
Kyriakopoulos, K. Tue-P-2-3-2 171
Tue-P-2-3-5 172

**L**

Laaridh, Imed Wed-P-3-3-2 225
Lagoudakis, Michail Tue-O-1-5-1 145
Lai, Cheng-I Tue-P-1-1-11 150
Lai, Jiahao Wed-P-1-1-6 187
Lala, Divesh Mon-P-2-1-13 130
Lalain, Muriel Wed-P-3-3-2 225
Lalhminghlui, Wendy Tue-SS-1-1-6 147
Lam, Max W. Y. Wed-O-1-1-3 177
Wed-P-3-3-1 225
Lambert, Mathias Mon-P-2-1-9 129
Lamel, Lori Wed-O-3-3-1 217
Wed-O-3-3-6 218
Lammert, Adam Mon-P-1-2-12 117
Tue-P-2-5-8 176
Lan, Haipeng Mon-P-2-5-2 138
Lancia, Leonardo Mon-P-2-1-2 128
Landini, Federico Wed-SS-3-1-3 220
Lane, David Tue-P-1-3-8 155
Lane, Ian Mon-P-2-4-8 137
Thu-P-1-3-7 242
Lang, Yue Thu-P-1-2-8 241
Lange, Patrick Mon-S&T-2-1-2 127
Wed-S&T-1-1-2 185
Lapidot, Itshak Wed-O-3-4-4 219
Larcher, Anthony Tue-O-2-3-1 160
Larson, Martha Wed-O-2-3-4 198
Latif, Siddique Mon-P-1-2-4 115
Thu-O-1-3-6 232
Law, Bradley Wed-P-1-3-3 191
Law, Vina Tue-O-2-5-3 163
Lawson, Aaron Thu-P-2-1-2 251
Tue-P-1-1-10 150
Tue-P-2-2-2 169
Laycock, Stephen Wed-P-2-2-5 207
Le Coz, Maxime Wed-P-2-3-3 209
Le Franc, Adrien Tue-O-2-3-4 161
Le Roux, Jonathan Wed-O-3-2-3 216
Le, Duc Wed-SS-1-1-3 182
Le, Nam Wed-O-2-2-5 197
Le, Ngoc Phu Mon-P-2-2-3 131
Lee, Chia-Hsuan Thu-O-2-1-4 247
Lee, Chi-Chun Mon-O-2-2-5 122
Mon-SS-2-1-8 126
Thu-O-1-3-5 232
Thu-P-1-4-11 246
Wed-O-2-5-2 200
Lee, Ching-Hua Tue-P-1-2-10 153
Lee, Chin-Hui Thu-P-1-1-11 239
Wed-SS-3-1-2 219
Lee, Hung-yi Mon-O-2-4-6 125
Thu-O-2-1-4 247
Thu-P-2-3-10 258
Wed-P-1-2-1 188
Lee, Jeremy Mon-S&T-2-1-2 127
Lee, Joun Yeop Tue-O-1-2-5 142
Lee, Kai-Zhan Wed-P-3-1-11 222
Lee, Kong Aik Mon-O-1-3-4 107
Tue-P-1-1-2 148
Lee, Lin-shan Mon-O-2-4-6 125
Thu-P-2-3-10 258
Lee, Sang-goo Wed-P-1-2-11 190
Lee, Shi-wook Wed-P-1-2-7 189
Lee, Tan Mon-O-2-2-3 121
Thu-P-1-4-7 245
Wed-O-3-1-2 215
Wed-P-2-1-10 205
Leemann, Adrian Wed-O-3-3-3 217
Legou, Thierry Mon-P-2-1-2 128
Lei, Ming Mon-P-2-4-3 136
Wed-P-2-4-14 214
Lei, Zhihong Mon-O-2-1-2 120
Lentz, Tomas O. Tue-O-2-5-2 163
Leung, Cheung-Chi Mon-O-1-4-2 108
Lev, Guy Wed-P-3-1-10 222
Levis, John Wed-O-3-4-6 219
Levit, Michael Thu-P-1-3-10 243
Levitan, Rivka Mon-O-2-2-2 121
Tue-P-2-5-9 176
Levitan, Sarah Ita Mon-O-2-2-1 121
Mon-O-2-2-2 121
Levy, Jason Wed-S&T-1-1-4 185
Levy, Russell Wed-S&T-1-1-4 185
Li, Bin Thu-P-1-1-1 237
Tue-O-2-4-4 162
Li, Bo Tue-O-1-1-6 141
Li, Chao Mon-P-1-2-7 116
Li, Chenxing Tue-O-2-1-1 158
Li, Dongbo Mon-P-2-2-7 132
Li, Haifeng Tue-O-1-3-2 143
Li, Haizhou Mon-O-1-2-5 106
Mon-O-1-4-2 108
Mon-S&T-2-1-5 128
Thu-O-2-2-2 248
Tue-P-1-1-2 148
Tue-P-2-1-3 166
Wed-P-1-1-3 186
Li, Hao Mon-S&T-2-1-5 128
Li, Hao Thu-O-1-2-6 231
Li, Jeng-Lin Mon-SS-2-1-8 126
Thu-O-1-3-5 232
Thu-P-1-4-11 246
Li, Jie Wed-O-1-1-5 178
Li, Jinyu Thu-P-1-1-9 238
Tue-P-1-2-7 152
Wed-O-1-1-1 177
Wed-P-3-2-7 224
Li, Juncheng Thu-P-1-2-1 239
Tue-O-2-2-1 159
Li, Junfeng Mon-P-2-5-7 139
Li, Ke Thu-P-1-3-9 243
Thu-P-2-3-9 258
Li, Kun Wed-P-2-3-5 210
Li, Ming Mon-SS-1-1-12 111
Thu-P-2-1-12 253
Li, Na Wed-O-2-2-6 198
Li, Pengcheng Thu-O-1-3-2 231
Li, Ruizhi Thu-O-1-1-3 229
Li, Runnan Mon-O-1-4-3 108
Li, Sheng Thu-P-2-3-2 257
Tue-O-2-2-4 160
Wed-O-1-2-4 179
Li, Wei Mon-O-1-4-4 108
Li, Xiangang Mon-P-2-2-7 132
Tue-P-2-2-11 170
Li, Xingfeng Thu-P-2-2-2 254
Li, Xu Wed-P-2-3-5 210
Li, Ya Mon-O-1-2-4 106
Thu-P-2-2-8 255
Li, Yan Wed-O-1-1-5 178
Li, Yaxing Thu-O-2-4-5 251
Thu-O-2-4-6 251
Li, Yichang Thu-O-2-4-5 251
Li, Yijie Wed-P-3-2-4 223
Li, Yujiang Mon-P-2-1-3 128
Li, Zhifeng Wed-O-2-2-6 198
Lian, Zhen Thu-P-2-2-8 255
Liang, Jiaen Wed-P-3-2-4 223
Liang, Shan Thu-P-1-1-1 237
Liao, Sishi Mon-P-1-1-4 113
Liberatore, Christopher Mon-O-2-4-1 124
Mon-O-2-4-2 124
Liberman, Mark Mon-P-1-1-4 113
Lilley, Jason Mon-O-2-3-4 123
Lim, Boon Pang Wed-P-2-4-5 212
Lim, Zhi Hao Mon-S&T-2-1-5 128
Lin, Hui Wed-P-2-3-13 211
Lin, Huibin Mon-O-1-3-2 106
Tue-P-1-1-4 149
Tue-P-1-1-5 149
Lin, Qiguang Wed-P-1-3-1 191
Lin, Shoufeng Wed-P-1-3-11 192
Lin, Yun-Shao Wed-O-2-5-2 200
Linarès, Georges Mon-O-1-1-5 105
Ling, Zhen-Hua Wed-P-1-1-4 186
Wed-P-2-2-10 208
Liou, Guan-Ting Mon-O-1-2-3 105
Lippus, Pärtel Tue-O-2-4-3 162
Liss, Julie Tue-P-2-3-4 171
Liu, Bin Thu-P-1-1-1 237
Liu, Bing Thu-P-1-3-7 242
Liu, Changliang Wed-O-1-1-1 177
Liu, Chi-Liang Thu-O-2-1-4 247
Liu, Chunxi Wed-P-1-2-5 189
Liu, Da-Rong Thu-P-2-3-10 258
Liu, Hanjun Thu-P-2-4-7 260
Liu, Hong Mon-P-2-5-2 138
Liu, Jia Wed-O-2-2-3 197
Liu, Li Wed-P-2-4-10 213
Liu, Li-Juan Wed-P-1-1-4 186
Liu, Linda Thu-P-1-3-1 241

Liu, Liquan Tue-O-2-5-2 163
Liu, Liu Mon-O-1-3-2 106
Tue-P-1-1-4 149
Tue-P-1-1-5 149
Liu, Meng Mon-P-2-2-7 132
Liu, Rui Mon-O-1-2-6 106
Liu, Rujie Mon-O-1-3-2 106
Tue-P-1-1-4 149
Tue-P-1-1-5 149
Liu, Shansong Wed-O-1-1-3 177
Wed-P-3-3-1 225
Liu, Shih-Chii Mon-O-1-1-4 104
Mon-P-2-5-4 138
Liu, Songxiang Mon-O-2-4-5 124
Thu-O-1-2-5 231
Liu, Wenju Thu-P-1-1-1 237
Liu, Xunying Mon-O-2-4-5 124
Thu-O-1-2-5 231
Thu-O-2-3-2 249
Wed-O-1-1-3 177
Wed-P-2-3-5 210
Wed-P-3-3-1 225
Liu, Yang Wed-P-2-3-13 211
Wed-SS-2-1-3 202
Liu, Yi Wed-O-2-2-3 197
Liu, Yong Wed-P-2-4-14 214
Liu, Yun Tue-P-1-2-4 152
Livescu, Karen Tue-P-1-4-10 157
Lleida, Eduardo Wed-SS-3-1-4 220
Long, Yanhua Wed-P-3-2-4 223
Lopez, Adam Tue-P-1-4-10 157
Lord, Catherine Wed-O-3-4-2 218
Lorenzo-Trueba, Jaime Thu-O-1-2-3 230
Lotfian, Reza Tue-O-1-3-6 143
Loweimi, Erfan Mon-P-2-3-1 133
Lowit, Anja Mon-O-2-3-1 122
Lu, Xugang Thu-P-2-3-2 257
Tue-O-2-2-4 160
Wed-O-1-2-4 179
Luan, Huan Wed-P-2-3-13 211
Lucic, Ivana Wed-O-3-4-6 219
Lugosch, Loren Wed-O-2-4-3 199
Luitjens, Justin Wed-O-2-1-2 196
Luo, Danqing Mon-SS-1-1-10 111
Luo, Dean Tue-P-2-3-8 172
Luo, Jacinta Dan Thu-P-2-4-2 259
Luo, Yi Mon-P-1-3-10 119
Mon-P-1-3-9 118
Luong, Hieu-Thi Mon-O-1-2-2 105
Wed-P-2-2-7 207
Luque, Jordi Wed-P-3-2-1 223
Luz, Saturnino Mon-O-2-2-4 122
Tue-P-2-5-4 176
Lv, Hang Tue-P-2-2-8 170

**M**

M A, Rohit Tue-P-1-3-7 155
M S, Saranya Mon-P-2-2-13 133
M, Gurunath Reddy Mon-P-2-5-3 138
M, Mahesh Mon-P-1-1-8 113
M, Nikhil Tue-O-2-1-4 159
M, Pushpavathi Mon-P-1-4-6 120
M, Ragesh Rajan Wed-P-1-2-3 189
M, Vikram C Tue-SS-2-2-7 174
Ma, Bin Mon-O-1-4-2 108
Ma, Matthew K. H. Mon-O-2-2-3 121
Ma, Xi Thu-P-2-2-10 256
Ma, Xiaoxiao Wed-P-1-4-4 193
Maas, Roland Tue-P-1-3-6 154
Maciejewski, Matthew Wed-SS-3-1-5 220
Mack, Wolfgang Tue-O-2-1-2 158
Madhavi, Maulik C. Wed-P-1-1-1 186
Madikeri, Srikanth Thu-P-2-1-8 252
Tue-P-1-1-9 150
Magg, Sven Tue-O-1-5-3 145
Maggu, Akshay Raj Thu-P-2-4-7 260
Tue-O-2-5-3 163
Magimai.-Doss, Mathew Mon-P-1-2-10 116
Mon-P-1-2-11 116
Mon-SS-2-1-6 126
Tue-P-1-1-12 150
Magnen, Cynthia Wed-P-3-3-4 225
Magron, Paul Mon-P-1-3-7 118
Mon-P-2-5-8 139
Mahanta, Shakuntala Wed-P-3-1-2 221
Maiti, Soumi Tue-P-1-2-12 153
Mak, Brian Mon-O-1-4-4 108
Thu-P-2-1-3 251
Malenovský, Vladimír Wed-P-2-4-9 213
Malinen, Jarmo Wed-P-3-4-3 227
Malinin, Andrey Tue-P-2-3-5 172
Mallidi, Sri Harish Tue-P-1-3-6 154
Mandal, Tanumay Wed-P-3-4-6 228
Mandel, Michael I Tue-P-2-2-12 171
Mandel, Michael Tue-P-1-2-12 153
Tue-P-1-2-13 153
Mangalam, Karttikeya Tue-O-1-3-5 143
Mannem, Renuka Thu-O-1-4-5 233
Manohar, Vimal Wed-P-1-2-5 189
Wed-SS-3-1-5 220
Mao, Shaoguang Wed-P-2-3-5 210
Marcel, Sebastien Tue-P-1-1-12 150
Marchi, Erik Wed-P-1-2-13 190
Maredia, Angel Mon-O-2-2-1 121
Margam, Dilip Kumar Wed-O-1-4-5 181
Markó, Alexandra Thu-SS-1-1-8 235
Marschik, Peter B. Mon-SS-1-1-1 109
Mon-SS-1-1-2 109
Mon-SS-1-1-4 109
Mon-SS-1-1-8 110
Mon-SS-2-1-1 125
Mon-SS-2-1-10 127
Marshall, David Mon-SS-2-1-3 125
Martinez, Kassandra Wed-O-3-4-2 218
Martinez, Victor Wed-SS-1-1-5 183
Maruthachalam, Srihari Tue-S&T-1-1-5 148
Marvin, Radhika Tue-P-1-3-9 155
Marxer, Ricard Wed-O-3-2-6 216
Masataki, Hirokazu Mon-O-2-1-4 121
Tue-P-1-4-2 156
Mason, Lisa Wed-O-1-2-1 178
Mass, Yosi Wed-P-3-1-10 222
Masumura, Ryo Mon-O-2-1-4 121
Tue-P-1-4-2 156
Tue-P-2-5-3 175
Mata, Ana Isabel Wed-P-1-4-5 194
Matějka, Pavel Tue-O-2-1-6 159
Wed-SS-3-1-3 220
Mateju, Lukas Wed-O-1-2-2 179
Mathew, Lani Mon-S&T-1-1-4 112
Mathivanan, Mithul Tue-S&T-2-1-5 166
Matosova, Anna Tue-P-2-1-5 167
Matoušek, Jindřich Wed-P-1-3-4 191
Matsoukas, Spyros Tue-P-1-3-6 154
Matsuo, Yutaka Thu-O-1-2-4 230
Matthews, Iain Wed-P-2-2-5 207
Matveeva, Anastasia Tue-P-1-1-8 149
Mau, Ted Thu-SS-1-1-4 234
Maurer, Dieter Thu-P-2-4-8 260
Tue-O-2-4-5 162
Maxwell, Olga Wed-P-1-4-8 194
McCree, Alan Wed-SS-3-1-5 220
McElwain, Nancy Mon-P-1-2-1 115
McFee, Brian Tue-P-2-2-12 171
McGlothlin, Jennifer Tue-SS-2-2-3 173
McInnis, Melvin Wed-SS-1-1-4 183
McLaren, Mitchell Mon-O-1-3-5 107
Thu-P-2-1-2 251
Thu-P-2-1-7 252
Tue-P-1-1-10 150
McLoughlin, Ian Thu-O-1-3-2 231
Thu-P-1-2-8 241
Thu-P-2-1-4 252
Wed-O-1-1-4 178
Medennikov, Ivan Wed-P-3-2-5 224
Meenakshi, G. Nisha Mon-O-2-4-4 124
Tue-P-2-1-10 168
Wed-P-3-4-4 227
Mehta, Daryush D. Tue-P-2-5-8 176
Meignier, Sylvain Tue-O-2-3-1 160
Melkote, Vinay Thu-O-2-4-4 250
Meltzner, Geoffrey S. Wed-P-3-1-12 223
Meng, Helen Mon-O-1-4-3 108
Mon-O-2-4-5 124
Mon-P-1-3-4 118
Thu-O-1-2-5 231
Thu-P-2-2-10 256
Meng, Helen Thu-Plenary-3 228
Wed-O-1-1-3 177
Wed-P-2-3-5 210
Wed-P-3-3-1 225
Meng, Qinglin Thu-P-2-4-2 259
Meng, Zhong Thu-P-1-1-9 238
Tue-P-1-2-7 152
Menon, Raghav Wed-P-2-4-3 212
Menshikova, Alla Wed-P-1-4-3 193
Merkx, Danny Wed-P-1-3-10 192
Mertens, Jane Tue-P-2-1-7 167
Mesgarani, Nima Mon-P-1-3-10 119
Mon-P-1-3-9 118
Wed-Perspective-3 195
Metallinou, Angeliki Thu-P-1-3-1 241
Metze, Florian Mon-O-2-1-3 120
Thu-P-1-2-1 239
Tue-O-2-2-1 159
Tue-O-2-3-4 161
Meunier, Fanny Wed-O-2-3-5 199
Wed-O-2-3-6 199
Meyer, Bernd T. Tue-O-1-4-5 144
Tue-O-1-4-6 144
Meyer, Julien Wed-O-1-3-4 180
Wed-O-2-3-5 199
Wed-O-2-3-6 199
Meyer, Patrick Mon-P-1-2-5 115
Meyer-Sickendiek, B. Wed-P-1-4-2 193
Miasato Filho, Valter Akira Wed-SS-3-1-7 220
Micheyl, Christophe Wed-O-2-3-2 198
Miculicich, Lesly Mon-O-1-4-1 107
Middag, Catherine Mon-P-1-4-1 119
Miguel, Antonio Wed-SS-3-1-4 220
Milde, Benjamin Wed-O-3-1-6 215
Miller, Mark Thu-S&T-1-1-1 236
Millwala, Zainab Tue-S&T-1-1-2 148
Tue-S&T-1-1-4 148
Milner, Ben Wed-P-2-2-4 207
Mimura, Masato Wed-O-2-1-6 196
Wed-P-2-1-7 205
Minagi, Shogo Wed-P-2-2-1 206
Minematsu, Nobuaki Tue-O-1-5-4 145
Tue-P-2-3-7 172
Mirheidari, Bahman Wed-SS-1-1-2 182
Mishra, Ambika Prasad Thu-P-2-4-2 259
Misra, Ananya Tue-O-1-1-6 141
Mitchell, Katherine Mon-P-1-2-12 117
Mitra, Vikramjit Thu-O-1-4-6 233
Wed-P-3-3-5 226
Mitsufuji, Yuki Wed-O-3-2-4 216
Mittag, Gabriel Wed-O-1-4-6 182
Mittal, Sanjeev Kumar Thu-O-1-4-4 233
Mixdorff, Hansjörg Mon-O-2-2-3 121
Mizgajski, Jan Wed-P-2-4-8 213
Mohammadi, S. H. Wed-P-3-1-3 221
Möller, Sebastian Wed-O-1-4-6 182
Monath, Nicholas Wed-P-1-2-6 189
Moniz, Helena Wed-P-1-4-5 194
Montacié, Claude Mon-SS-2-1-9 126
Tue-P-1-1-8 149
Moore, Meredith Mon-O-2-3-5 123
Moran, Joseph Mon-SS-2-1-2 125
Morchid, Mohamed Mon-O-1-1-5 105
Mordechay, Moran Wed-P-3-1-10 222
Morency, Louis-Philippe Mon-P-2-1-12 130
Moreno, Pedro Wed-O-2-1-4 196
Mori, Hiroki Thu-O-1-2-1 230
Mori, Koichi Tue-P-2-1-11 168
Mori, Takuma Mon-P-2-4-10 137
Morise, Masanori Wed-P-1-1-13 188
Moriya, Takafumi Tue-P-1-4-4 156
Wed-P-2-1-2 204
Wed-P-2-1-7 205
Mošner, Ladislav Tue-O-2-1-6 159
Wed-SS-3-1-3 220
Motlicek, Petr Mon-P-1-3-3 117
Thu-O-2-3-4 249
Thu-P-2-1-8 252
Tue-P-1-1-9 150
Mower Provost, Emily Wed-SS-1-1-3 182
Wed-SS-1-1-4 183
Mücke, Doris Mon-P-1-1-5 113
Tue-P-2-1-7 167
Muckenhirn, Hannah Mon-P-1-2-10 116
Tue-P-1-1-12 150
Mukherjee, Sankar Mon-P-2-1-2 128
Mukherjee, Subhadip Tue-P-1-2-2 151
Mulder, Kimberley Tue-O-2-5-6 164

Mulholland, Matthew Wed-SS-2-1-6 202
Mulimani, Manjunath Thu-P-1-2-9 241
Müller, Markus Wed-P-2-1-6 205
Murakami, Hiroki Wed-P-2-2-1 206
Murphy, Andy Mon-P-1-1-10 114
Mon-P-1-1-9 114
Murthy, Hema Mon-P-1-1-8 113
Mon-P-1-2-11 116
Mon-P-2-2-13 133
Mon-S&T-2-1-4 128
Thu-SS-1-2-2 235
Tue-P-1-3-4 154
Tue-S&T-1-1-5 148
Tue-S&T-2-1-1 165
Wed-SS-1-2-7 184
Murthy, Savitha Tue-SS-1-1-4 146
Murtola, Tiina Wed-P-3-4-3 227
Myer, Samuel Tue-P-1-4-3 156

**N**

N K, Kausthubha Mon-S&T-2-1-1 127
Thu-O-1-4-4 233
Wed-S&T-2-1-4 203
N P, Narendra Thu-P-1-4-4 244
N, Abinay Reddy Tue-P-2-1-10 168
N, Krishna D Tue-SS-1-1-5 147
Wed-S&T-2-1-3 203
N, Mohana Mon-S&T-2-1-4 128
Nagamatsu, Kenji Wed-P-3-2-9 224
Nagano, Tohru Wed-P-3-2-3 223
Nagarajan, Srikantan Tue-P-2-1-1 166
Nagarsheth, Parav Thu-S&T-1-1-4 237
Nagata, Tomohiro Thu-O-1-2-1 230
Nagrani, Arsha Tue-P-1-1-6 149
Naik, Chetan Mon-P-2-1-9 129
Nair, Krishnakumar N Thu-O-2-3-5 249
Nakagawa, Seiichi Mon-P-2-2-7 132
Nakamura, Satoshi Mon-P-2-4-10 137
Tue-O-1-1-5 141
Tue-O-1-2-2 141
Tue-SS-2-2-5 174
Nakashika, Toru Wed-P-2-2-12 208
Wed-P-2-2-14 209
Nakatani, Tomohiro Thu-O-1-1-5 229
Wed-O-1-4-2 181
Wed-P-2-1-11 206
Nallan Chakravarthula, S. Wed-O-2-5-4 201
Namboodiri, Vinay Thu-O-2-2-4 248
Namyalo, Saudah Mon-P-2-1-14 130
Nandwana, Mahesh Kumar Thu-P-2-1-2 251
Tue-P-1-1-10 150
Tue-P-2-2-2 169
Nara, Kiranpreet Mon-P-1-1-6 113
Narayanamurthy, BHVS Wed-S&T-2-1-2 203
Narayanan, Arun Thu-O-1-1-2 229
Tue-O-1-1-6 141
Narayanan, Shrikanth Mon-P-1-2-11 116
Mon-P-1-2-9 116
Thu-P-1-4-6 245
Thu-P-1-4-8 245
Thu-P-2-2-5 255
Tue-O-2-3-3 161
Tue-P-1-2-3 152
Wed-O-2-5-5 201
Wed-O-3-4-2 218
Wed-SS-1-1-5 183
Naresh, Annam Tue-S&T-1-1-3 148
Nasir, Md Thu-P-1-4-8 245
Nataraj, K S Wed-P-1-3-8 192
Nathwani, Karan Mon-O-1-4-5 108
Neil, Daniel Mon-O-1-1-4 104
Nenadić, Filip Thu-P-2-4-3 259
Neubig, Graham Thu-P-2-3-5 257
Ney, Hermann Mon-O-1-1-2 104
Mon-O-2-1-2 120
Mon-P-2-4-2 136
Thu-P-1-3-6 242
Tue-O-1-1-3 140
Ng, Chip-Jin Thu-P-1-4-11 246
Ng, Manwa Tue-SS-2-2-6 174
Ngente, Lalnunsiami Tue-SS-1-1-6 147
Nguyen, Huy Wed-SS-2-1-3 202
Nguyen, Noël Mon-P-2-1-2 128
Nguyen, Patrick Wed-P-3-3-8 226
Ní Chasaide, Ailbhe Mon-P-1-1-10 114
Mon-P-1-1-9 114
Ni, Jinfu Wed-P-3-1-1 221
Ni, Karl Thu-O-2-2-1 247
Tue-P-2-2-2 169
Nicolao, Mauro Tue-P-2-3-10 173
Nicolson, Aaron Tue-P-2-2-10 170
Nidadavolu, Phani Sankar Mon-P-1-2-8 116
Tue-P-1-1-11 150
Nie, Shuai Thu-P-1-1-1 237
Niehues, Jan Thu-P-2-3-5 257
Tue-P-1-4-9 157
Wed-P-1-2-9 190
Niesler, Thomas Wed-P-2-4-2 212
Wed-P-2-4-3 212
Wed-SS-1-2-2 184
Nigam, Aditya Wed-P-1-3-6 192
Nishitoba, Jiro Wed-O-2-1-1 196
Nishizawa, Nobuyuki Mon-O-1-2-2 105
Nixon, Jessie S. Tue-O-2-5-2 163
Tue-O-2-5-5 164
Nomo Sudro, Protima Tue-P-2-1-9 167
Norel, Raquel Mon-P-1-4-5 120
Nose, Takashi Tue-P-2-5-6 176
Nöth, Elmar Mon-O-2-3-3 123
Wed-O-2-5-6 201
Novak III, John S. Wed-O-1-3-6 181
Novoselov, Sergey Wed-O-2-2-2 197
Novotný, Ondřej Tue-O-2-1-6 159
Wed-SS-3-1-3 220
Nyberg, Eric Mon-SS-2-1-4 125

**O**

Ochi, Keiko Tue-P-2-1-11 168
Wed-P-3-4-8 228
Ochiai, Tsubasa Wed-O-2-1-1 196
Odobez, Jean-Marc Mon-P-1-3-3 117
Wed-O-2-2-5 197
Ogawa, Atsunori Mon-O-1-1-1 104
Wed-P-2-1-11 206
Ohashi, Narumi Wed-O-1-4-2 181
Ohsugi, Yasuhito Tue-O-1-5-4 145
Okabe, Koji Wed-O-2-2-4 197
Öktem, Alp Tue-S&T-2-1-4 166
Oliveira, Sara Wed-S&T-1-1-3 185
Ondel, Lucas Wed-O-3-1-3 215
Wed-P-1-2-5 189
Oo, Zeyan Mon-P-2-2-7 132
Ooster, Jasper Tue-O-1-4-5 144
Oraby, Shereen Thu-O-1-2-2 230
Orife, Iroro Wed-P-3-1-6 222
Orozco-Arroyave, J. R. Mon-O-2-3-3 123
Wed-O-2-5-6 201
Ortega, Alfonso Wed-SS-3-1-4 220
O'Shaughnessy, Douglas Thu-P-2-2-6 255
Ostendorf, Mari Wed-P-2-1-3 204
Ottl, Sandra Wed-O-2-5-3 201
Oualil, Youssef Thu-O-2-3-4 249
Ouni, Slim Thu-O-1-4-1 232
Wed-O-3-4-5 219
Ouyang, Zhiheng Thu-P-1-1-2 237
Özkanca, Yasin Thu-P-1-4-3 244

**P**

P A, Anand Mon-S&T-2-1-1 127
Wed-S&T-2-1-4 203
Padmasundari Mon-P-2-1-11 130
Pal, Madhab Tue-S&T-2-1-3 165
Pal, Radhika Mon-S&T-2-1-4 128
Palaniappan, R. Thu-P-1-2-8 241
Paliwal, Kuldip K. Tue-P-2-2-10 170
Pallavi, Priya Tue-P-1-2-6 152
Palmer, Jeffrey Tue-P-2-5-8 176
Panchanathan, S. Mon-O-2-3-5 123
Panda, Ashish Tue-P-2-5-7 176
Pandey, Abhishek Tue-S&T-2-1-5 166
Pandey, Ashutosh Tue-P-1-2-1 151
Pandey, Laxmi Mon-O-1-4-5 108
Thu-O-2-2-4 248
Pandey, Prem C. Tue-P-1-2-9 153
Wed-P-1-3-8 192
Pandya, Mugdha Wed-SS-1-2-1 183
Pang, Cheng Mon-P-2-5-2 138
Pang, Ruoming Mon-O-1-1-6 105
Pankajakshan, Arjun Wed-P-1-3-6 192
Pantofaru, Caroline Tue-P-1-3-9 155
Papadopoulos, Pavlos Tue-O-2-3-3 161
Tue-P-1-2-3 152
Papangelis, Alexandros Mon-P-2-1-5 129
Tue-O-1-5-1 145
Papayiannis, Constantinos Tue-O-2-2-6 160
Pappagari, Raghavendra Mon-P-1-2-2 115
Parada-Cabaleiro, Emilia Thu-P-2-2-1 254
Wed-P-3-1-9 222
Paramasivam, P. Wed-S&T-1-1-1 185
Paraskevopoulos, G. Tue-O-1-3-1 142
Parcollet, Titouan Mon-O-1-1-5 105
Parekh, Tanmay Wed-P-2-4-1 212
Parizet, Etienne Wed-O-2-3-2 198
Park, Heewoong Thu-P-2-1-1 251
Park, Jinhwan Mon-O-2-1-6 121
Thu-P-2-3-6 258
Park, Jinuk Tue-P-2-1-12 168
Park, Jonghun Thu-P-2-1-1 251
Park, Kyubyong Thu-P-2-1-1 251
Park, Soo Jin Mon-SS-1-1-11 111
Tue-SS-2-2-2 173
Park, Tae Jin Tue-O-2-3-2 161
Parrell, Benjamin Tue-P-2-1-1 166
Parthasarathy, S. Mon-O-2-1-5 121
Thu-P-1-3-10 243
Parthasarathy, Srinivas Mon-P-1-2-3 115
Thu-P-2-2-13 256
Tue-O-1-3-4 143
Pascual, Santiago Wed-P-1-1-7 187
Patel, Abhishek Thu-O-2-1-1 246
Patel, Rupal Wed-P-3-1-12 223
Patel, Tanvina Thu-SS-1-2-6 236
Tue-SS-1-1-5 147
Wed-S&T-2-1-3 203
Pati, Debadatta Mon-P-2-3-12 135
Patil, Ankur T. Thu-SS-1-2-4 235
Patil, Hemant Mon-P-2-2-4 131
Mon-P-2-2-5 131
Mon-P-2-2-9 132
Mon-P-2-3-3 134
Mon-P-2-3-4 134
Mon-P-2-3-6 134
Mon-P-2-3-7 134
Thu-SS-1-1-5 234
Thu-SS-1-2-4 235
Wed-P-1-1-1 186
Patino, Jose Wed-SS-3-1-6 220
Pautler, David Wed-S&T-1-1-2 185
Pavitt, Joe Mon-S&T-2-1-3 127
Payne, Elinor Wed-P-1-4-8 194
Peinado, Antonio M. Mon-P-2-2-11 132
Peleg, Shmuel Tue-P-1-2-8 152
Pellegrino, Elisa Thu-P-2-4-8 260
Peng, Yu-Huai Mon-O-2-4-3 124
Penney, Joshua Tue-O-2-4-6 162
Percival, Maida Tue-P-2-1-8 167
Perez, Matthew Wed-SS-1-1-3 182
Perrier, Pascal Tue-O-1-4-1 143
Peshterliev, Stanislav Mon-P-2-1-4 129
Petitrenaud, Simon Tue-O-2-3-1 160
Petrović, Sunčica Wed-O-2-5-3 201
Pham, Lam Dang Thu-P-1-2-8 241
Pham, Ngoc-Quan Tue-P-1-4-9 157
Pham, Van Tung Mon-S&T-2-1-5 128
Phan, Huy Thu-P-1-2-8 241
Piccaluga, Myriam Mon-P-1-4-2 119
Pietrowicz, Mary Mon-P-1-4-5 120
Pinquier, Julien Wed-P-3-3-4 225
Plchot, Oldřich Tue-O-2-1-6 159
Wed-SS-3-1-3 220
Podsiadło, Monika Wed-P-1-1-8 187
Poellabauer, Christian Wed-O-3-2-1 216
Pokorny, Florian B. Mon-SS-1-1-1 109
Mon-SS-1-1-2 109
Mon-SS-1-1-4 109
Mon-SS-1-1-8 110
Mon-SS-2-1-1 125
Mon-SS-2-1-10 127
Popov, Dmitry Wed-P-3-2-5 224
Potamianos, Alexandros Tue-O-1-3-1 142
Povey, Daniel Mon-O-1-1-3 104
Mon-P-1-2-8 116
Thu-O-1-3-4 232
Thu-P-1-3-9 243
Thu-P-2-1-3 251

Thu-P-2-3-9 258
Tue-P-2-2-8 170
Wed-O-1-1-6 178
Wed-O-2-1-2 196
Wed-SS-3-1-5 220
Prabhakar, Anil Tue-S&T-2-1-1 165
Prabhakaran, Divya Tue-P-2-1-2 166
Prabhavalkar, Rohit Mon-O-1-1-6 105
Pradhan, Gayadhar Mon-P-2-3-13 136
Tue-P-1-2-5 152
Tue-P-2-2-9 170
Prakash, Anusha Thu-SS-1-2-7 236
Wed-SS-1-2-7 184
Prakash, Jeena J Mon-P-1-1-8 113
Prasad, RaviShankar Mon-P-1-1-1 112
Mon-P-1-1-3 112
Prasanna, S R Mahadeva Mon-O-1-4-6 108
Mon-P-1-1-13 114
Mon-P-1-4-3 119
Mon-P-1-4-4 119
Mon-P-1-4-6 120
Mon-P-2-2-2 131
Mon-SS-1-1-6 147
Thu-S&T-1-1-2 237
Tue-P-2-1-9 167
Tue-SS-2-2-7 174
Wed-O-2-4-4 200
Wed-S&T-2-1-5 203
Pražák, Aleš Thu-O-2-3-6 250
Prieto, Ramon Wed-SS-2-1-3 202
Priyadharshini, Veeramani Wed-P-3-4-7 228
Pucher, Michael Tue-P-2-1-5 167
Pugh, Robert Wed-SS-2-1-6 202
Pulugundla, Bhargav Thu-SS-1-2-3 235
Pundak, Golan Tue-O-1-1-6 141
PV, Muhammed Shifas Wed-O-1-4-3 181

## Q

Qadir, Junaid Mon-P-1-2-4 115
Thu-O-1-3-6 232
Qaroush, Aziz Wed-SS-2-1-4 202
Qi, Xiaoke Mon-P-2-5-5 139
Qian, Kun Mon-SS-1-1-1 109
Mon-SS-1-1-2 109
Mon-SS-1-1-4 109
Mon-SS-1-1-7 110
Mon-SS-1-1-8 110
Mon-SS-2-1-1 125
Mon-SS-2-1-10 127
Qian, Mengjie Thu-P-2-3-11 259
Wed-SS-2-1-1 201
Wed-SS-2-1-2 202
Wed-SS-2-1-5 202
Mon-P-1-3-1 117
Qian, Yanmin Mon-P-1-3-2 117
Thu-P-2-3-1 257
Tue-P-2-2-6 169
Qian, Yao Wed-SS-2-1-6 202
Qin, Ying Thu-P-1-4-7 245
Qin, Zhen Wed-P-3-3-7 226
Quatieri, Thomas Mon-P-1-2-12 117
Tue-P-2-5-8 176
Quinn, John Wed-P-2-4-3 212

## R

R, Muralishankar Tue-P-1-3-5 154
Rabinovitz, Carmel Mon-S&T-2-1-3 127
Ragni, Anton Thu-P-1-3-2 242
Tue-P-2-3-5 172
Wed-O-2-1-3 196
Rahman, Md Hafizur Thu-P-2-1-7 252
Rai, Atul Wed-O-2-2-1 197
Raj, Ankit Thu-P-2-3-8 258
Raj, Bhiksha Wed-P-3-3-6 226
Rajan, Golda Brunet Thu-SS-1-2-2 235
Rajan, Padmanabhan Wed-P-1-3-6 192
Wed-P-1-3-7 192
Rajkomar, Alvin Wed-P-1-2-10 190
Raju, Anirudh Thu-P-1-3-1 241
Rallabandi, SaiKrishna Mon-SS-2-1-4 125
Wed-P-2-2-3 207
Ram, Dhananjay Mon-O-1-4-1 107
Ram, Dhananjay Wed-P-1-2-8 190
Ramabhadran, B. Tue-Perspective-2 177
Wed-P-2-1-4 204
Ramakrishna, Anil Wed-O-2-5-5 201
Ramakrishnan, Ganesh Thu-P-1-2-7 240
Ramanarayanan, Vikram Mon-S&T-2-1-2 127
Tue-P-2-1-1 166
Wed-S&T-1-1-2 185
Wed-SS-2-1-6 202
Ramanathan, Prabha Tue-S&T-2-1-1 165
Ramasubramanian, V Tue-SS-1-1-2 146
Tue-SS-1-1-7 147
Ramoji, Shreyas Tue-P-1-1-7 149
Rana, Rajib Mon-P-1-2-4 115
Thu-O-1-3-6 232
Randhawa, Shan Tue-SS-1-1-3 146
Rao MV, Achuth Tue-P-2-1-10 168
Wed-P-3-4-7 228
Rao, Bhaskar D. Tue-P-1-2-10 153
Rao, Ch V Rama Tue-P-1-2-6 152
Rao, K. Sreenivasa Mon-P-2-3-8 135
Mon-P-2-5-3 138
Tue-SS-1-1-2 146
Wed-P-3-4-6 228
Rao, Preeti Tue-O-2-1-4 159
Tue-P-1-3-7 155
Tue-S&T-2-1-2 165
Wed-SS-1-2-1 183
Rao, Wei Thu-O-2-2-2 248
Räsänen, Okko Mon-P-2-3-2 133
Tue-P-1-3-1 154
Rasooli, Mohammad S. Mon-O-2-1-5 121
Rastogi, Abhinav Thu-O-2-1-6 247
Rastrow, Ariya Thu-P-1-3-1 241
Tue-P-1-3-6 154
Rath, Shakti P Thu-P-2-3-8 258
Rathner, Eva-Maria Mon-P-1-2-6 116
Mon-SS-1-1-1 109
Mon-SS-1-1-2 109
Mon-SS-1-1-4 109
Mon-SS-1-1-8 110
Mon-SS-2-1-1 125
Mon-SS-2-1-10 127
Thu-P-1-4-1 243
Ravanelli, Mirco Thu-P-2-3-4 257
Ravi, Abishek Wed-P-1-2-6 189
Ravi, Nithya Thu-SS-1-2-7 236
Ravi, Vijay Tue-SS-2-2-2 173
Ray, Avik Thu-O-2-1-3 247
Raymond, Christian Thu-O-2-1-2 247
Rayner, Manny Mon-S&T-1-1-5 112
Wed-SS-2-1-1 201
Raza, Agha Ali Tue-SS-1-1-3 146
Reale, Nathan Tue-P-1-3-9 155
Reed, Lena Thu-O-1-2-2 230
Reidy, Patrick Tue-P-2-1-2 166
Ren, Zhao Mon-SS-1-1-1 109
Mon-SS-1-1-2 109
Mon-SS-1-1-4 109
Mon-SS-1-1-8 110
Mon-SS-2-1-1 125
Mon-SS-2-1-10 127
Thu-O-1-3-1 231
Renduchintala, Adithya Wed-O-2-1-1 196
Wed-P-2-1-1 204
Renkens, Vincent Mon-P-2-1-10 130
Tue-SS-2-1-3 164
Reuber, Markus Wed-SS-1-1-2 182
Reynolds, Douglas Wed-O-1-2-1 178
Riad, Rachid Wed-O-3-1-4 215
Wed-P-2-4-13 214
Ribeiro, Jorge Wed-S&T-1-1-3 185
Ribeiro, Manuel Sam Wed-SS-1-1-1 182
Riccardi, Giuseppe Tue-O-1-5-6 146
Richards, Hywel Wed-P-1-2-13 190
Richburg, Brian Tue-SS-2-2-9 175
Richey, Colleen Tue-P-1-1-10 150
Tue-P-2-2-2 169
Richmond, Korin Wed-SS-1-1-1 182
Ridouane, Rachid Mon-P-1-1-5 113
Wed-O-1-3-4 180
Riebling, Eric Tue-O-2-3-4 161
Riedhammer, Korbinian Tue-P-1-4-11 157
Riley, Michael Wed-O-2-1-4 196
Rilliard, Albert Mon-O-2-2-3 121
Ringeval, Fabien Thu-O-1-3-1 231
Rishoni, Shay Mon-P-1-4-5 120
Roberts, Angela Wed-SS-1-1-3 182
Robinson, Amanda Thu-S&T-1-1-1 236
Rochet-Capellan, Amélie Tue-O-1-4-1 144
Roddy, Matthew Mon-P-2-1-7 129
Roe, Paul Wed-P-1-3-3 191
Rohdin, Johan Wed-SS-3-1-3 220
Roman, Stephane Wed-P-3-4-1 227
Romanenko, Aleksei Wed-P-3-2-5 224
Ronanki, Srikanth Mon-O-1-2-1 105
Rong, Yu Wed-P-2-3-2 209
Rose, Phil Wed-O-3-3-2 217
Rosec, Olivier Wed-P-2-2-6 207
Rosenberg, Andrew Wed-P-2-1-4 204
Rosenfeld, Elizabeth Thu-P-2-2-4 255
Rosenfeld, Roni Tue-SS-1-1-3 146
Rosenstock, Todd Wed-P-2-4-6 213
Ross, Sula Wed-O-3-3-5 217
Rossato, Solange Tue-P-1-1-1 148
Roth, Joseph Tue-P-1-3-9 155
Rouas, Jean-Luc Wed-P-1-4-10 195
Rouhe, Aku Tue-S&T-1-1-1 148
Roxburgh, Zoe Wed-SS-1-1-1 182
Roy, Rajib Tue-S&T-2-1-3 165
Roy, Somnath Wed-P-3-1-2 221
Rozenberg, Shai Mon-S&T-2-1-3 127
Rozgic, Viktor Tue-O-2-2-6 160
Ruan, Huabin Tue-O-1-3-2 142
Rudolph, Johanna Tue-SS-2-2-3 173
Rudresh, Sunil Mon-P-2-3-10 135
Russell, Martin Thu-P-2-3-11 259
Tue-SS-2-1-4 165
Wed-P-2-3-9 211
Wed-SS-2-1-1 201
Wed-SS-2-1-5 202
Rybach, David Wed-O-2-1-5 196

## S

S, Chandana Thu-O-1-4-4 233
S, Kumaran Thu-O-2-3-5 249
S.R., Nirmala Thu-S&T-1-1-2 237
Thu-SS-1-1-6 147
Sabu, Kamini Tue-S&T-2-1-2 165
Wed-SS-1-2-1 183
Sachdev, Umesh Tue-S&T-1-1-2 148
Tue-S&T-1-1-4 148
Sad, Gonzalo Wed-O-3-4-5 219
Sadasivan, Jishnu Tue-P-1-2-2 151
Sadeghi, Ahmad-Reza Tue-P-1-4-11 157
Sadjadi, Seyed Omid Wed-O-1-2-1 178
Sadoughi, Najmeh Thu-S&T-1-1-1 236
Safarik, Radek Wed-O-1-2-2 179
Sagisaka, Yoshinori Wed-P-2-3-4 210
Saha, Goutam Thu-P-1-2-10 241
Saha, Pramit Tue-P-1-3-11 155
Sahidullah, Md Mon-O-1-3-4 107
Sahu, Saurabh Thu-P-2-2-12 256
Saif, Umar Tue-SS-1-1-3 146
Sailor, Hardik B. Mon-P-2-3-3 134
Thu-SS-1-2-4 235
Sailor, Hardik Mon-P-2-2-9 132
Sainath, Tara Mon-O-1-1-6 105
Tue-O-1-1-6 141
Wed-O-2-1-5 196
Sainathan, Abhilash Mon-P-2-3-10 135
Saito, Daisuke Tue-O-1-5-4 145
Saito, Daisuke Tue-P-2-3-7 172
Sak, Hasim Wed-P-3-3-8 226
Sakai, Naomi Tue-P-2-1-11 168
Sakai, Shinsuke Wed-O-2-1-6 196
Wed-P-2-1-7 205
Sakakibara, Ken-Ichi Wed-P-1-1-13 188
Sakshi, Surbhi Mon-P-2-3-13 136
Sakti, Sakriani Mon-P-2-4-10 137
Tue-O-1-1-5 141
Tue-O-1-2-2 141
Salamon, Gudrun Thu-P-1-4-1 243
Saleem, Muhammad Bilal Tue-SS-1-1-3 146
Sameti, Hossein Mon-O-1-1-3 104
Samudravijaya, K Thu-S&T-1-1-2 237
San Segundo F., E. Mon-P-1-1-11 114
Sanabria, Ramon Mon-O-2-1-3 120
Sanders, Michiel Tue-P-2-3-10 173
Sangwan, Abhijeet Thu-P-2-1-9 253
Wed-O-3-4-1 218
Sankar, Ananth Wed-P-3-3-8 226
Sar Shalom, Oren Wed-P-3-1-10 222

Saranu, Kinnera Tue-S&T-2-1-5 166
Sari, Leda Thu-O-2-3-5 249
Sarikaya, Ruhi Mon-P-2-1-9 129
Sarma, Biswajit Dev Tue-SS-1-1-6 147
Sarma, Kandarpa Kumar Thu-O-1-3-4 232
Sarma, Mousmita Thu-O-1-3-4 232
Wed-S&T-2-1-1 203
Sarmah, Priyankoo Mon-P-1-1-13 114
Thu-S&T-1-1-2 237
Tue-SS-1-1-6 147
Wed-S&T-2-1-5 203
Sarria-Paja, Milton Tue-P-2-1-4 166
Satheesh, Sanjeev Mon-O-2-1-1 120
Sathyendra, K. M. Mon-P-2-1-4 129
Sato, Masaaki Wed-P-2-2-1 206
Satyanarayana, JV Wed-S&T-2-1-2 203
Saurous, Rif A. Tue-O-1-3-3 143
Scaff, Camila Tue-O-2-3-4 161
Scharenborg, Odette Tue-SS-2-1-6 165
Wed-O-2-3-4 198
Wed-P-1-3-10 192
Schatz, Thomas Wed-P-2-4-13 214
Scheiner, Justin Wed-O-2-1-4 196
Schiller, Dominik Mon-SS-1-1-9 111
Schlüter, Ralf Mon-O-1-1-2 104
Mon-O-2-1-2 120
Mon-P-2-4-2 136
Thu-P-1-3-6 242
Tue-O-1-1-3 140
Schmid, Stephan Wed-O-3-3-3 217
Schmidhuber, Jürgen Thu-SS-1-1-7 234
Schmitt, Maximilian Mon-SS-1-1-1 109
Mon-SS-1-1-2 109
Mon-SS-1-1-4 109
Mon-SS-1-1-8 110
Mon-SS-2-1-1 125
Mon-SS-2-1-10 127
Thu-O-1-3-1 231
Schneider, Thomas Tue-P-1-4-11 157
Schnell, Bastian Thu-SS-1-1-3 234
Schnupp, Jan W. H. Thu-P-2-4-2 259
Schroeter, Julien Mon-SS-2-1-3 125
Schuller, Björn Mon-P-1-2-6 116
Mon-SS-1-1-1 109
Mon-SS-1-1-2 109
Mon-SS-1-1-4 109
Mon-SS-1-1-7 110
Mon-SS-1-1-8 110
Mon-SS-2-1-1 125
Mon-SS-2-1-10 127
Thu-O-1-3-1 231
Thu-O-2-3-1 249
Thu-P-1-4-1 243
Thu-P-2-2-1 254
Tue-O-1-3-2 143
Wed-O-2-5-3 201
Wed-P-2-3-12 211
Wed-P-3-1-9 222
Schultz, Tanja Thu-SS-1-1-6 234
Thu-SS-1-1-7 234
Wed-O-2-5-1 200
Schwab, Sandra Thu-P-2-4-8 260
Schwartz, Jean-Luc Thu-P-2-4-6 260
Tue-O-1-4-1 143
Scobbie, James M Wed-SS-1-1-1 182
Sebastian, Jilt Mon-P-1-2-11 116
Mon-SS-2-1-6 126
Sèbe, Frédéric Wed-O-2-3-5 199
Seelamantula, Chandra S. Mon-P-2-3-10 135
Mon-P-2-3-9 135
Tue-P-1-2-2 151
Seeram, Tejaswi Thu-SS-1-2-7 236
Segura, Carlos Wed-P-3-2-1 223
Seiderer, Andreas Mon-SS-1-1-9 111
Sell, Gregory Wed-SS-3-1-5 220
Seltzer, Michael Wed-P-3-2-7 224
Seneviratne, Nadee Thu-O-1-4-6 233
Serdyuk, Dmitriy Thu-P-2-3-4 257
Serigos, Jacqueline Wed-P-2-3-1 209
Serra, Xavier Mon-P-2-3-5 134
Sethu, Vidhyasaharan Mon-P-2-2-10 132
Sethu, Vidhyasaharan Mon-P-2-2-14 133
Thu-P-2-2-7 255
Wed-O-1-2-5 179
Sezgin, Metin Tue-P-2-5-5 176
Shafaei-Bajestan, Elnaz Tue-O-1-4-3 144
Shah, Neil Thu-SS-1-1-5 234
Shah, Nirmesh Mon-P-2-3-4 134
Thu-SS-1-1-5 234
Wed-P-1-1-1 186
Shah, Nisar Tue-SS-1-1-5 147
Shah, Nisar Wed-S&T-2-1-3 203
Shahin, Mostafa Tue-SS-2-2-1 173
Shahnawazuddin, Syed Tue-P-1-2-5 152
Tue-P-2-2-9 170
Shamir, Asaph Tue-P-1-2-8 152
Shan, Changhao Wed-P-1-2-2 188
Shankar, Ravi Mon-O-1-4-6 108
Shantirani Devi, Senjam Wed-S&T-2-1-5 203
Shao, Yiwen Wed-P-1-3-1 191
Shapiro, Avi Mon-SS-2-1-2 125
Sharma, Bhamini Thu-P-2-4-4 260
Sharma, Dravyansh Thu-P-2-3-7 258
Wed-P-3-1-8 222
Sharma, Neeraj Mon-P-2-3-9 135
Sharma, Pulkit Thu-P-1-2-5 240
Wed-P-1-3-7 192
Sharma, Saketh Tue-P-1-2-9 153
Sharma, Tanay Wed-O-1-4-5 181
Sharon, Rini A Thu-SS-1-2-7 236
Tue-O-1-1-4 140
Shchemelinin, Vadim Wed-O-2-2-2 197
Shechtman, Slava Mon-S&T-2-1-3 127
Wed-P-3-1-10 222
Shen, Binbin Mon-P-1-3-4 118
Shen, John Thu-P-1-3-7 242
Shen, Peng Thu-P-2-3-2 257
Tue-O-2-2-4 160
Wed-O-1-2-4 179
Shen, Yi Tue-O-1-4-2 144
Shen, Yilin Thu-O-2-1-1 246
Thu-O-2-1-3 247
Thu-O-2-1-5 247
Thu-P-1-3-5 242
Shetty, Vishwas M. Thu-SS-1-2-7 236
Shi, Kaiyu Tue-P-1-4-1 156
Shi, Yang Thu-P-2-3-3 257
Shi, Ziqiang Mon-O-1-3-2 106
Tue-P-1-1-4 149
Tue-P-1-1-5 149
Shiga, Yoshinori Wed-P-3-1-1 221
Shim, Hye-jin Thu-P-2-1-5 252
Shin, Youhyun Wed-P-1-2-11 190
Shinoda, Koichi Thu-P-2-1-11 253
Tue-SS-2-2-8 174
Wed-O-2-2-4 197
Shinohara, Yusuke Tue-P-1-4-4 156
Wed-P-2-1-2 204
Wed-P-2-1-7 205
Shlykov, Andrei Tue-P-1-1-8 149
Shochi, Takaaki Wed-P-1-4-10 195
Shokouhi, Navid Wed-P-1-3-5 191
Shreyas, Seshadri Tue-P-1-3-1 154
Shulipa, Andrey Wed-O-2-2-2 197
Sidorov, Kirill Mon-SS-2-1-3 125
Sigtia, Siddharth Wed-P-1-2-13 190
Silnova, Anna Mon-O-1-3-3 107
Wed-SS-3-1-3 220
Silpachai, Alif Wed-O-3-4-6 219
Silva, Diego Augusto Wed-SS-3-1-7 220
Sim, Khe Chai Tue-O-1-1-6 141
Simantiraki, Olympia Wed-O-2-3-1 198
Šimko, Juraj Tue-P-2-1-6 167
Šimko, Juraj Wed-O-1-3-3 180
Singer, Elliot Wed-O-1-2-1 178
Singh, Astha Thu-O-1-4-4 233
Wed-P-3-4-4 227
Singh, Chirag Thu-P-2-2-11 256
Singh, Madhusudan Mon-P-2-3-12 135
Singh, Mittul Thu-O-2-3-4 249
Singla, Karan Thu-P-1-4-6 245
Sinha, Rohit Mon-P-2-2-2 131
Thu-S&T-1-1-2 237
Tue-SS-1-1-6 147
Wed-SS-1-2-8 185
Sinyavskaya, Yadviga Tue-P-1-1-8 149
Sisman, Berrak Mon-O-1-2-5 106
Sisman, Berrak Wed-P-1-1-3 186
Sitaram, Dinkar Tue-SS-1-1-4 146
Sitaram, Sunayana Tue-SS-1-1-4 146
Wed-SS-1-2-6 184
Sivaraman, Ganesh Thu-O-1-4-6 233
Thu-S&T-1-1-4 237
Sivasankaran, Sunit Wed-O-3-2-2 216
Skantze, Gabriel Mon-P-2-1-7 129
Sloboda, Jennifer Tue-P-2-5-8 176
Smalt, Christopher Tue-P-2-5-8 176
Smit, Peter Tue-S&T-1-1-1 148
Smolander, Anna-Riikka Tue-S&T-1-1-1 148
Snyder, David Mon-O-1-3-3 107
Thu-P-2-1-3 251
Wed-SS-3-1-5 220
Socher, Richard Thu-P-2-3-12 259
Somandepalli, Krishna Mon-P-1-2-9 116
Sone, Kentaro Wed-P-2-2-12 208
Song, Eunwoo Tue-O-1-2-4 142
Tue-O-1-2-5 142
Song, Hongwei Thu-P-1-2-4 240
Song, Huan Thu-P-2-1-10 253
Song, Yan Mon-O-1-3-1 106
Thu-O-1-3-2 231
Thu-P-1-2-8 241
Thu-P-2-1-4 252
Wed-O-1-1-4 178
Soni, Akshay Thu-O-2-2-5 248
Sonsaat, Sinem Wed-O-3-4-6 219
Soong, Frank K. Wed-P-1-1-11 188
Soong, Frank Tue-P-2-3-3 171
Soraghan, John Mon-O-2-3-1 122
Sorin, Alexander Mon-S&T-2-1-3 127
Wed-P-2-1-4 204
Sorokin, Ivan Wed-P-3-2-5 224
Soto, Victor Wed-SS-1-2-5 184
Spanias, Andreas Thu-P-2-1-10 253
Sperber, Matthias Thu-P-2-3-5 257
Tue-P-1-4-9 157
Spreafico, Lorenzo Tue-P-2-1-5 167
Sreeram, Victor Mon-P-2-2-1 130
Sridhar, Kusha Tue-O-1-3-4 143
Sridharan, Sridha Thu-P-2-1-7 252
Srinivas, Nagapuri Tue-P-1-2-5 152
Srinivasamurthy, Ajay Thu-O-2-3-4 249
Srinivasan, Umesh Tue-O-1-1-4 140
Sriram, Anuroop Mon-O-2-1-1 120
Sriskandaraja, Kaavya Mon-P-2-2-10 132
Srivastava, Brij Mohan Lal Wed-SS-1-2-6 184
Srungarapu, Praneeth Tue-P-1-3-11 155
Stanek, Jan Mon-S&T-1-1-5 112
Stauffer, Allen Tue-P-1-1-10 150
Tue-P-2-2-2 169
Steidl, Stefan Mon-SS-1-1-1 109
Mon-SS-1-1-2 109
Mon-SS-1-1-4 109
Mon-SS-1-1-8 110
Mon-SS-2-1-1 125
Mon-SS-2-1-10 127
Stemmer, Georg Thu-O-2-3-5 249
Stemp, Christoph Thu-O-2-3-1 249
Stepanov, Evgeny Tue-O-1-5-6 146
Stephenson, Cory Thu-O-2-2-1 247
Tue-P-2-2-2 169
Stern, Richard Thu-P-1-1-12 239
Stöter, Fabian-Robert Tue-O-2-1-2 158
Strik, Helmer Wed-SS-2-1-1 201
Strimel, Grant P. Mon-P-2-1-4 129
Studer-Joho, Dieter Wed-O-3-3-3 217
Stüker, Sebastian Thu-P-2-3-5 257
Stüker, Sebastian Wed-P-1-2-9 190
Wed-P-2-1-6 205
Stylianou, Yannis Tue-O-1-5-1 145
Tue-P-2-2-7 170
Wed-O-1-4-3 181
Su, Bo-Hao Mon-SS-2-1-8 126
Su, Dan Mon-P-1-3-1 117
Mon-P-1-3-2 117
Mon-P-2-4-1 136
Thu-O-1-2-5 231
Wed-O-2-2-6 198
Wed-P-2-4-4 212
Su, Ming-Hsiang Tue-O-1-5-5 145
Su, Rongfeng Thu-O-2-3-2 249
Su, Rongfeng Wed-O-1-1-3 177
Su, Zhihua Thu-O-1-4-2 232
Subramanian, Aswin S. Thu-P-1-1-7 238
Tue-P-2-2-3 169

Suendermann-Oeft, D. Mon-S&T-2-1-2 127
Thu-S&T-1-1-1 236
Wed-S&T-1-1-2 185
Sun, Jian Wed-P-3-1-7 222
Sun, Lei Wed-SS-3-1-2 219
Sun, Lifa Mon-O-2-4-5 124
Sun, Ming Tue-O-2-2-5 160
Sun, Sining Thu-O-1-1-6 230
Tue-P-2-2-5 169
Wed-P-2-1-3 204
Sun, Yue Wed-P-2-3-4 210
Sundaram, Shiva Tue-O-2-2-6 160
Sung, Wonyong Mon-O-2-1-6 121
Thu-P-2-3-6 258
Suni, Antti Wed-O-1-3-3 180
Sur, Mriganka Tue-S&T-1-1-5 148
Suresh, Parth Mon-P-2-5-9 139
Suter, Heidy Tue-O-2-4-5 162
Suthokumar, Gajan Mon-P-2-2-14 133
Suzuki, Masayuki Wed-P-3-2-3 223
Syed, Ali Raza Tue-P-1-2-13 153
Syed, Zafi Sherhan Mon-SS-2-1-3 125
Synnaeve, Gabriel Mon-P-2-4-5 136
Szakay, Anita Tue-O-2-4-6 162
Szalóki, Szilvia Thu-P-1-4-5 244
Szaszák, György Wed-P-2-4-7 213
Szendi, István Thu-P-1-4-5 244
Szöke, Igor Wed-P-2-4-9 213
Wed-P-3-2-1 223
Szurley, Joseph Thu-P-1-2-1 239
Szymański, Piotr Wed-P-2-4-8 213
Szymczak, Adrian Wed-P-2-4-8 213

**T**

T.S., Sharath Thu-O-1-2-2 230
Tabain, Marija Wed-P-1-4-6 194
Taguchi, Fumiaki Wed-P-2-2-8 208
Tai, Ching-Lun Wed-P-1-2-1 188
Taira, Kazuya Wed-O-1-4-1 181
Taitelbaum, Hagai Wed-O-1-2-3 179
Tak, Hemlata Mon-P-2-2-4 131
Mon-P-2-3-7 134
Takahashi, Naoya Wed-O-3-2-4 216
Takanashi, Katsuya Mon-P-2-1-13 130
Tue-O-1-5-2 145
Takashima, Ryoichi Thu-P-2-3-2 257
Takeda, Kazuya Mon-P-2-4-9 137
Tamura, Satoshi Wed-P-2-2-2 207
Tan, Ke Thu-O-2-2-3 248
Thu-P-1-1-3 238
Tan, Zheng-Hua Wed-O-1-2-6 179
Tanaka, Hiroki Tue-SS-2-2-5 174
Tanaka, Kazuyo Wed-P-1-2-7 189
Tanaka, Tomohiro Mon-O-2-1-4 121
Tue-P-1-4-2 156
Tandon, Shubhangi Thu-O-1-2-2 230
Tang, Dengke Mon-SS-1-1-12 111
Tang, Hao Tue-P-2-2-4 169
Wed-P-3-2-10 224
Tang, Jian Wed-O-1-1-4 178
Tanihara, Keisuke Wed-P-2-2-9 208
Tansuwan, Justin Wed-P-3-3-8 226
Tao, Fei Tue-P-1-3-10 155
Tao, Jianhua Mon-O-1-2-4 106
Mon-P-2-5-5 139
Thu-P-1-1-1 237
Thu-P-2-2-8 255
Tue-O-1-2-3 141
Tue-O-1-2-6 142
Wed-P-2-2-11 208
Tapkir, Prasad Mon-P-2-3-6 134
Tardieu, Julien Wed-P-3-3-4 225
Tariq, Zain Tue-SS-1-1-3 146
Tavarone, Raffaele Tue-P-1-4-5 156
Taylor, Kye Mon-SS-2-1-2 125
Taylor, Sarah Wed-P-2-2-4 207
Teixeira, Francisco Thu-P-1-4-9 245
Wed-P-3-3-6 226
ten Bosch, Louis Thu-P-2-4-3 259
Tue-O-1-4-4 144
Tue-O-2-5-6 164
ten Bosch, Louis Tue-SS-2-1-1 164
Teng, Zhang Thu-P-1-2-3 240
Teplansky, Kristin Tue-SS-1-1-8 147
Wed-SS-1-1-6 183
Terhorst, Yannik Mon-P-1-2-6 116
Thu-P-1-4-1 243
Terissi, Lucas D. Wed-O-3-4-5 219
Thakur, Anshul Thu-P-1-2-5 240
Wed-P-1-3-6 192
Wed-P-1-3-7 192
Thanda, Abhinav Wed-O-1-4-5 181
Thangthai, Kwanchiva Wed-P-2-4-11 214
Thapar, Daksh Wed-P-1-3-6 192
Thiagarajan, Jayaraman J. Thu-P-2-1-10 253
Thomas, Anju Leela Wed-SS-1-2-7 184
Thomas, Samuel Wed-P-2-1-4 204
Wed-P-3-2-3 223
Tian, Guangjian Wed-P-3-3-7 226
Tian, Li Wed-O-3-4-3 218
Tiesmeyer, Sebastian Tue-SS-2-1-6 165
Tihelka, Daniel Wed-P-1-3-4 191
Timpe-Laughlin, V. Mon-S&T-2-1-2 127
Tiwari, Nitya Tue-P-1-2-9 153
Tjandra, Andros Mon-P-2-4-10 137
Tue-O-1-1-5 141
Tobin, Stephen Tue-O-2-4-2 162
Tobing, Patrick Lumban Wed-P-1-1-5 186
Toda, Tomoki Mon-P-2-4-9 137
Thu-SS-1-1-2 234
Wed-P-1-1-13 188
Wed-P-1-1-5 186
Wed-P-2-2-2 207
Todisco, Massimiliano Mon-O-1-3-4 107
Tue-P-1-2-11 153
Wed-O-3-4-4 219
Toffy, Anoop Tue-SS-1-1-7 147
Togneri, Roberto Mon-P-2-2-1 130
Toivola, Minnaleena Tue-S&T-1-1-1 148
Tom, Francis Mon-P-2-2-12 133
Toman, Markus Wed-P-3-1-12 223
Tomar, Vikrant Singh Tue-P-1-4-3 156
Wed-O-2-4-3 199
Tomashenko, Natalia Wed-P-2-1-5 204
Wed-P-3-2-5 224
Tomassini, Alice Mon-P-2-1-2 128
Tong, Sibo Wed-P-2-1-14 206
Toribio, Almeida J. Wed-P-2-3-1 209
Tóth, László Mon-SS-2-1-7 126
Thu-SS-1-1-8 235
Towsey, Michael Wed-P-1-3-3 191
Trabelsi, Chiheb Mon-O-1-1-5 105
Trancoso, Isabel Thu-P-1-4-9 245
Wed-P-1-4-5 194
Wed-P-3-3-6 226
Trinh, Viet Anh Tue-P-1-2-13 153
Tue-P-2-2-12 171
Tripathi, Anshuman Tue-O-1-1-6 141
Wed-P-3-3-8 226
Tripathi, Ayush Tue-SS-2-2-7 174
Tripathi, Kumud Mon-P-2-3-8 135
Tripathi, Suraj Thu-P-2-2-11 256
Trmal, Jan Tue-P-2-2-1 168
Wed-P-1-2-5 189
Truong, Quy-Thao Wed-P-1-4-7 194
Tsai, Fu-Sheng Mon-O-2-2-5 122
Tsao, Yu Mon-O-2-4-3 124
Tue-O-2-2-4 160
Wed-O-1-4-4 181
Tseng, Shao-Yen Thu-P-1-2-1 239
Tsiaras, Vassilis Wed-O-1-4-3 181
Wed-P-1-1-10 187
Tsourakis, Nikos Mon-S&T-1-1-5 112
Tsukada, Kimiko Wed-P-2-3-2 209
Tsuprun, Eugene Mon-S&T-2-1-2 127
Wed-S&T-1-1-2 185
Tu, Ming Mon-P-2-1-6 129
Tue-P-2-3-4 171
Tucker, Benjamin V. Thu-P-2-4-3 259
Tue-P-1-3-2 154
Tündik, Máté Ákos Wed-P-2-4-7 213
Tung, Kuan Wed-P-1-2-1 188
Tuo, Deyi Wed-O-2-2-6 198
Turan, Mehmet Ali T. Mon-SS-1-1-5 110
Turco, Giuseppina Wed-O-1-3-4 180
Türker, Bekir Berker Tue-P-2-5-5 176
Tüske, Zoltán Thu-P-1-3-6 242
Tzinis, Efthymios Tue-O-1-3-1 142
Tzirakis, Panagiotis Mon-SS-1-1-1 109
Mon-SS-1-1-2 109
Mon-SS-1-1-4 110
Mon-SS-1-1-8 110
Mon-SS-2-1-1 125
Mon-SS-2-1-10 127

**U**

Ubale, Rutuja Wed-S&T-1-1-2 185
Wed-SS-2-1-6 202
Ueno, Sei Wed-P-2-1-2 204
Ueno, Sei Wed-P-2-1-7 205
Ujiro, Tsuyoki Tue-SS-2-2-5 174
Umesh, S. Thu-SS-1-2-7 236
Umesh, Srinivasan Wed-O-2-5-1 200
Ungureanu, Victor Wed-P-1-1-8 187
Unno, Yuya Wed-O-2-1-1 196
Upreti, Minali Wed-P-2-1-13 206
Usova, Maria Tue-P-1-1-8 149
Usunier, Nicolas Mon-P-2-4-5 136

**V**

V, Vishnu Vidyadhara Raju Thu-SS-1-2-5 236
Vachhani, Bhavik Mon-O-2-3-2 123
Mon-O-2-3-6 123
Vaissière, Jacqueline Tue-Plenary-1 140
Valentini-Botinhao, Cassia Wed-P-1-1-12 188
van den Heuvel, Henk Wed-SS-1-2-4 184
van der Westhuizen, Ewald Wed-P-2-4-2 212
Wed-SS-1-2-2 184
van Hamme, Hugo Mon-P-2-1-10 130
Tue-SS-2-1-3 164
Tue-SS-2-1-5 164
van Hout, Julien Thu-P-2-1-2 251
Tue-P-1-1-10 150
Tue-P-2-2-2 169
van Leeuwen, David Wed-SS-1-2-4 184
van Malderen, Sophie Mon-P-1-4-2 119
van Ommen, Sandrien Tue-O-2-5-2 163
van Rij, Jacolien Tue-O-2-5-2 163
van Santen, Jan Thu-SS-1-1-4 234
van Son, Rob Mon-P-1-4-1 119
Variani, Ehsan Thu-O-1-1-2 229
Vasco, Cristina Wed-P-1-1-2 186
Vasilescu, Ioana Wed-O-3-3-1 217
Vásquez Correa, Juan C. Mon-O-2-3-3 123
Wed-O-2-5-6 201
Vaz, Colin Tue-P-1-2-3 152
Velikovich, Leonid Wed-O-2-1-4 196
Velmurugan, Rajbabu Tue-O-2-1-4 159
Venkata Siva Krishna, M. Thu-SS-1-2-4 235
Venkatesan, Shankar Wed-O-1-4-5 181
Venkatesh, Anu Thu-P-1-3-1 241
Venkateswara, Hemanth Mon-O-2-3-5 123
Venneri, Annalena Wed-SS-1-1-2 182
Vepa, Jithendra Thu-P-2-2-11 256
Thu-P-2-3-8 258
Tue-S&T-2-1-5 166
Wed-S&T-1-1-1 185
Verhulst, Sarah Thu-P-1-1-10 239
Verkholyak, Oxana Mon-SS-2-1-5 126
Verma, Prateek Wed-O-2-4-1 199
Verwimp, Lyan Tue-SS-2-1-3 164
Veselý, Karel Wed-P-2-4-9 213
Wed-P-3-2-1 223
Wed-SS-3-1-3 220
Viegas, Carla Mon-SS-2-1-4 125
Vieru, Bianca Wed-O-3-3-1 217
Vijayakumar, Ashwin Wed-P-1-2-3 189
Vijayasenan, Deepu Wed-P-1-2-3 189
Viksnin, Ilya Tue-P-1-1-8 149
Villalba, Jesús Mon-O-1-3-6 107
Mon-P-1-2-2 115
Mon-P-1-2-8 116
Tue-P-1-1-11 150
Wed-O-1-2-6 179
Wed-SS-3-1-5 220
Villavicencio, Aline Wed-O-3-1-3 215
Villemoes, Lars Thu-O-2-4-4 250
Viñals, Ignacio Wed-SS-3-1-4 220
Vincent, Emmanuel Tue-P-2-2-1 168
Wed-O-3-2-2 216
Virkkunen, Päivi Wed-O-1-3-3 180
Virtanen, Tuomas Mon-P-1-3-7 118
Virtanen, Tuomas Mon-P-2-5-8 139
Viswanathan, Ramya Wed-S&T-1-1-1 185
Vlasenko, Bogdan Mon-SS-2-1-6 126

Vogel, Carl Tue-P-2-5-4 176
Voix, Jérémie Tue-P-2-1-4 166
Vozila, Paul Thu-P-1-3-8 243
Vuppala, Anil Kumar Thu-SS-1-2-5 236
Vythelingum, Kévin Wed-P-2-2-6 207

**W**

Wagner, Johannes Mon-SS-1-1-9 111
Waibel, Alex Mon-O-2-1-3 120
Thu-P-2-3-5 257
Tue-P-1-4-9 157
Wed-P-1-2-9 190
Wed-P-2-1-6 205
Waldekar, Shefali Thu-P-1-2-10 241
Walker, Marilyn Thu-O-1-2-2 230
Walker, Traci Wed-SS-1-1-2 182
Wambacq, Patrick Tue-SS-2-1-3 164
Wan, Moquan Tue-O-1-2-1 141
Wan, Nathan Wed-P-3-3-8 226
Wand, Michael Thu-SS-1-1-7 234
Wang, Chao Tue-O-2-2-2 159
Tue-O-2-2-5 160
Tue-O-2-2-6 160
Wang, Chuan Wed-SS-2-1-3 202
Wang, DeLiang Mon-P-1-3-5 118
Thu-O-2-2-3 248
Thu-P-1-1-3 238
Thu-P-1-1-4 238
Thu-P-1-1-5 238
Tue-P-1-2-1 151
Wed-O-3-2-3 216
Wed-O-3-2-5 216
Wang, Disong Mon-P-2-5-1 138
Wang, Guangsen Mon-P-2-4-1 136
Wang, Haishuai Mon-P-1-2-7 116
Wang, Hsin-Min Mon-O-2-4-3 124
Wed-O-1-4-4 181
Wang, Jianrong Thu-O-1-4-2 232
Wang, Jun R. Thu-SS-1-1-4 234
Wang, Jun Mon-P-1-3-2 117
Mon-P-2-4-1 136
Thu-SS-1-1-4 234
Tue-SS-1-1-8 147
Wed-P-3-3-3 225
Wed-SS-1-1-6 183
Wang, Ke Tue-P-2-2-5 169
Wed-P-2-1-8 205
Wang, Lan Thu-O-2-3-2 249
Tue-SS-2-2-6 174
Wang, Lei Wed-P-2-3-6 210
Wang, Lixin Tue-P-2-3-8 172
Wang, Longbiao Mon-P-2-2-7 132
Tue-O-2-5-1 163
Tue-P-2-2-11 170
Wang, Mu Thu-O-1-2-5 231
Wang, Rui Thu-P-2-4-5 260
Wang, Shuai Thu-P-2-1-13 253
Wang, Tianqi Tue-SS-2-2-6 174
Wang, Tieqiang Tue-O-2-1-1 158
Wang, Wei Thu-P-2-4-1 259
Wang, Weiran Tue-O-2-2-2 159
Tue-O-2-2-5 160
Wang, Xi Wed-P-1-1-11 188
Wang, Xiaofei Thu-O-1-1-3 229
Wang, Xiaorui Wed-O-1-1-5 178
Wang, Xin Mon-O-1-2-2 105
Wang, Ye Tue-P-2-1-3 166
Wang, Yih-Ru Mon-O-1-2-3 105
Wang, Yiming Thu-P-1-3-9 243
Thu-P-2-3-9 258
Wed-O-2-1-2 196
Wang, Yonghe Mon-O-1-2-6 106
Wang, Yu Thu-O-2-1-1 246
Thu-O-2-1-5 247
Tue-O-1-1-2 140
Tue-P-2-3-5 172
Wang, Yujun Tue-P-2-2-5 169
Wed-P-1-2-2 188
Wed-P-2-1-8 205
Wang, Yun Thu-P-1-2-1 239
Wang, Yun Tue-O-2-2-1 159
Tue-O-2-3-4 161
Wang, Yu-Wun Tue-P-2-5-1 175
Wang, Zhenyu Thu-O-1-2-6 231
Wang, Zhixiang Tue-O-1-3-2 142
Wang, Zhong-Qiu Mon-P-1-3-5 118
Thu-P-1-1-4 238
Wed-O-3-2-3 216
Wed-O-3-2-5 216
Wang, Ziteng Mon-P-2-5-7 139
Warlaumont, Anne Wed-P-2-3-12 211
Warnita, Tifani Tue-SS-2-2-8 174
Watanabe, Kazuho Tue-P-1-4-4 156
Watanabe, Shinji Mon-O-1-1-1 104
Mon-P-2-4-9 137
Thu-P-1-1-7 238
Tue-P-2-2-1 168
Tue-P-2-2-3 169
Wed-O-1-2-6 179
Wed-O-2-1-1 196
Wed-P-2-1-1 204
Wed-P-2-1-11 206
Wed-SS-3-1-5 220
Watts, Oliver Mon-O-1-2-1 105
Wed-P-1-1-12 188
Wayland, Ratree Mon-P-1-1-13 114
Weber, Cornelius Tue-O-1-5-3 145
Weber, Philip Tue-SS-2-1-4 165
Websdale, Danny Wed-P-2-2-4 207
Wei, Jianguo Thu-O-1-4-2 232
Wei, Wei Thu-P-2-4-1 259
Wei, Xizi Wed-SS-2-1-1 201
Wed-SS-2-1-5 202
Weiner, Jochen Wed-O-2-5-1 200
Weinert, Christian Tue-P-1-4-11 157
Wen, Zhengqi Mon-O-1-2-4 106
Tue-O-1-2-3 141
Tue-O-1-2-6 142
Wed-P-2-2-11 208
Weng, Chao Mon-P-2-4-1 136
Mon-P-2-4-6 137
Weng, Yi-Ming Thu-P-1-4-11 246
Wenjie, Li Wed-P-3-2-2 223
Wermter, Stefan Tue-O-1-5-3 145
Wickramasinghe, Buddhi Mon-P-2-2-3 131
Mon-P-2-2-8 132
Wiesner, Matthew Wed-O-2-1-1 196
Wed-P-1-2-5 189
Wed-P-2-1-1 204
Wijenayake, Chamith Mon-P-2-2-14 133
Willi, Megan Mon-P-2-1-6 129
Thu-P-2-1-10 253
Williams, Daniel Tue-O-2-5-4 163
Williams, Ian Wed-O-2-1-4 196
Wed-O-2-1-5 196
Williamson, James Mon-P-1-2-12 117
Tue-P-2-5-8 176
Wilson, Kevin Tue-P-1-3-9 155
Witzlack-Makarevich, A. Mon-P-2-1-14 130
Wlodarczak, Marcin Tue-O-2-4-3 162
Tue-P-2-1-6 167
Woisard, Virginie Wed-P-3-3-2 225
Wong, Francis C. K. Thu-P-2-4-7 260
Wong, Ka Ho Wed-P-3-3-1 225
Wong, Patrick C. M. Thu-P-2-4-7 260
Tue-O-2-5-3 163
Woodland, Philip Tue-O-1-1-2 140
Wed-O-1-1-2 177
Wed-P-3-2-8 224
Wottawa, Jane Wed-O-3-3-6 218
Wrench, Alan Wed-SS-1-1-1 182
Wright, Jonathan Mon-P-1-1-4 113
Wu, Chung-Hsien Tue-O-1-5-5 145
Wu, Fei Wed-P-2-3-13 211
Wu, Haoran Tue-P-2-5-6 176
Wu, Huiyi Mon-O-2-3-1 122
Wu, Ji Thu-P-1-2-11 241
Thu-P-1-2-3 240
Tue-O-2-2-3 160
Wu, Szu-Lin Thu-O-2-1-4 247
Wu, Xihong Tue-O-1-4-2 144
Wu, Xixin Mon-O-2-4-5 124
Thu-O-1-2-5 231
Wed-P-2-3-5 210
Wed-P-3-3-1 225
Wu, Yichiao Mon-O-2-4-3 124
Wu, Yi-Chiao Wed-P-1-1-5 186
Wu, Yonghui Mon-O-1-1-6 105
Wed-P-3-3-8 226
Wu, Zhiyong Mon-O-1-4-3 108
Mon-P-1-3-4 118
Thu-O-1-2-5 231
Thu-P-2-2-10 256

**X**

Xi, Zhonghua Tue-P-1-3-9 155
Xia, Linzhong Tue-P-2-3-8 172
Xia, Risheng Mon-P-2-5-7 139
Xia, Wei Mon-P-2-3-11 135
Xia, Yangyang Thu-P-1-1-12 239
Xiang, Fei Tue-P-2-2-5 169
Xiao, Bo Wed-P-1-2-6 189
Xiao, Xiong Thu-O-1-1-4 229
Xiao, Yujia Tue-P-2-3-3 171
Xie, Hui Thu-P-2-4-5 260
Xie, Lei Mon-O-1-4-2 108
Tue-P-2-2-5 169
Wed-P-1-2-2 188
Wed-P-2-1-3 204
Wed-P-2-1-8 205
Wed-SS-1-2-3 184
Xie, Xurong Wed-O-1-1-3 177
Wed-P-3-3-1 225
Xie, Yanlu Mon-P-1-3-11 119
Thu-P-2-4-1 259
Wed-P-3-4-9 228
Xiong, Caiming Thu-P-2-3-12 259
Xiong, Shengwu Thu-O-2-4-5 251
Thu-O-2-4-6 251
Xu, Bo Mon-P-2-4-12 138
Mon-P-2-4-7 137
Tue-O-2-1-1 158
Xu, Chenglin Thu-O-2-2-2 248
Xu, Haihua Mon-S&T-2-1-5 128
Wed-SS-1-2-3 184
Xu, Hainan Thu-P-1-3-9 243
Thu-P-2-3-9 258
Tue-P-2-2-3 169
Wed-O-2-1-2 196
Xu, Ji Wed-O-1-1-6 178
Xu, Kaiyuan Thu-P-2-3-3 257
Xu, Longting Tue-P-1-1-2 148
Xu, Mingxing Thu-P-2-2-10 256
Thu-P-2-2-9 256
Xu, Ning Thu-P-2-3-3 257
Xu, Shan Thu-O-2-4-6 251
Xu, Shuang Mon-P-2-4-7 137
Tue-O-2-1-1 158
Xu, Weiqun Mon-P-2-1-3 128
Xu, Yi Wed-P-2-2-13 208
Xu, Yijia Mon-P-1-2-1 115
Xuanda, Chen Tue-P-2-1-13 168

**Y**

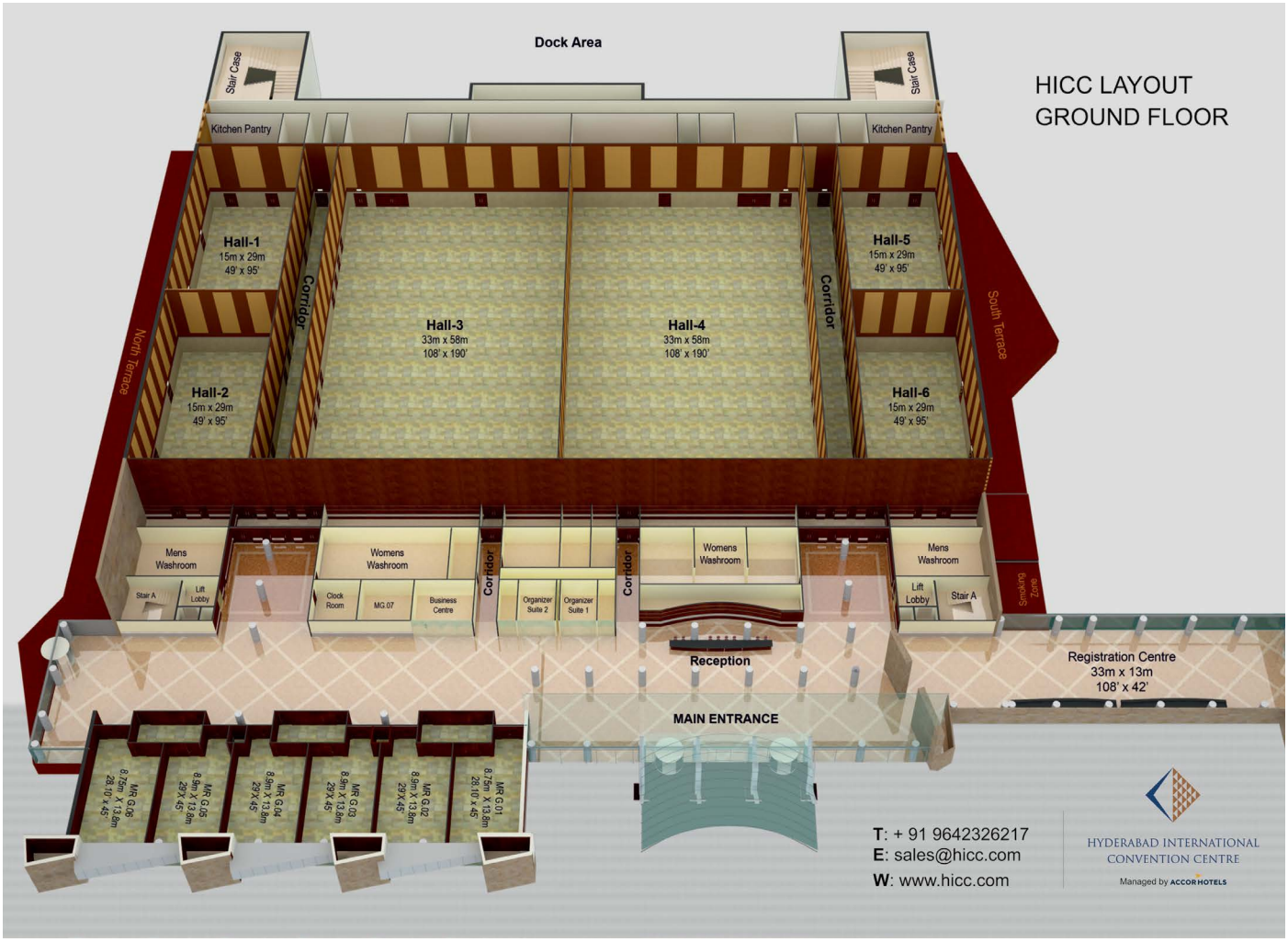
Yadav, Ishwar Chandra Tue-P-2-2-9 170
Yadav, Sarthak Wed-O-2-2-1 197
Yamagishi, Junichi Mon-O-1-2-2 105
Mon-O-1-3-4 107
Thu-O-1-2-3 230
Wed-P-1-1-10 187
Wed-P-2-2-7 207
Yamaguchi, Yoshikazu Tue-P-1-4-4 156
Wed-P-2-1-2 204
Wed-P-2-1-7 205
Yamamoto, Katsuhiko Wed-O-1-4-2 181
Yamamoto, Seiichi Wed-P-1-4-7 194
Yamauchi, Yutaka Tue-P-2-3-7 172
Yan, Nan Tue-SS-2-2-6 174
Yan, Yonghong Mon-P-2-1-3 128
Mon-P-2-5-7 139
Thu-P-1-2-6 240
Thu-P-1-3-4 242
Wed-O-1-1-6 178
Wed-P-3-2-2 223
Yanagita, Tomoya Tue-O-1-2-2 141
Yang, Bing Mon-P-2-5-2 138
Yang, Feng Tue-SS-2-2-6 174
Yang, Hao-Chun Mon-O-2-2-5 122
Yang, Hongying Tue-O-1-4-2 144
Yang, IL-ho Thu-P-2-1-5 252
Yang, Jichen Mon-P-2-2-6 131
Yang, Longfei Mon-P-1-3-11 119
Yang, Shuai Mon-P-1-3-4 118
Yang, Xiaofei Wed-P-2-3-13 211
Yang, Xuesong Wed-P-2-4-5 212
Yang, Yike Tue-O-2-4-4 142

Yang, Zhen Tue-P-1-1-2 148
Yang, Zixiaofan Thu-O-1-3-3 232
Yanushevskaya, Irena Mon-P-1-1-10 114
Mon-P-1-1-9 114
Yao, Haitao Wed-P-2-4-14 214
Yarmohammadi, M. Thu-P-2-3-9 258
Yarra, Chiranjeevi Mon-S&T-2-1-1 127
Thu-O-1-4-4 233
Wed-S&T-2-1-4 203
Ye, Hong Wed-P-3-2-4 223
Yegnanarayana, B. Mon-O-2-2-6 122
Mon-P-1-1-1 112
Mon-P-1-1-12 114
Mon-P-1-1-3 112
Wed-O-2-4-2 199
Wed-O-2-4-6 200
Wed-S&T-2-1-2 203
Yeh, Cheng-chieh Mon-O-2-4-6 125
Yeh, Ching-Feng Wed-P-2-1-3 204
Yeh, Sung-Lin Mon-SS-2-1-8 126
Yemez, Yücel Tue-P-2-5-5 176
Yenigalla, Promod Thu-P-2-2-11 256
Yeruva, Vaishnavi Tue-P-1-1-13 151
Yeşilkanat, Ali Mon-SS-2-1-5 126
Yeung, Gary Tue-P-2-3-9 172
Yılmaz, Emre Wed-P-2-4-2 212
Wed-P-3-3-5 226
Wed-SS-1-2-2 184
Wed-SS-1-2-4 184
Yin, Bing Wed-SS-3-1-2 219
Yin, Lu Mon-P-2-5-7 139
Yin, Ruiqing Tue-O-2-3-6 161
YingjianFu, YingjianFu Tue-O-2-5-1 163
Yokoyama, Masaki Thu-O-1-2-1 230
Yonekura, Shogo Wed-P-2-2-9 208
Yoo, Kang Min Wed-P-1-2-11 190
Yoshida, Tsukasa Tue-P-1-4-4 156
Yoshioka, Takuya Thu-O-1-1-4 229
Thu-O-2-2-6 248
You, Changhuai Mon-P-2-2-6 131
You, Lanhua Mon-O-1-3-1 106
Wed-P-3-1-7 222
You, Yongbin Thu-P-2-3-1 257
Younis, Shahzad Mon-P-1-2-4 115
Yousefi, Midia Wed-P-1-3-5 191
Yu, Chengzhu Mon-P-2-4-1 136
Mon-P-2-4-6 137
Wed-O-3-4-1 218
Yu, Dong Mon-P-1-3-1 117
Mon-P-1-3-2 117
Mon-P-2-4-1 136
Mon-P-2-4-6 137
Thu-O-1-2-5 231
Tue-P-2-2-6 169
Wed-O-2-2-6 198
Wed-P-2-4-4 212
Yu, Ha-jin Thu-P-2-1-5 252
Yu, Hongjiang Thu-P-1-1-2 237
Yu, Jenny Wed-O-1-3-2 180
Yu, Jianwei Wed-O-1-1-3 177
Wed-P-3-3-1 225
Yu, Kai Thu-P-2-1-13 253
Thu-P-2-3-1 257
Tue-P-1-4-1 156
Wed-P-1-1-6 187
Yu, Meng Mon-P-1-3-1 117
Mon-P-1-3-2 117
Wed-P-2-4-4 212
Yuan, Jiahong Mon-P-1-1-4 113
Wed-P-2-3-13 211
Yuan, Qingsheng Thu-P-1-2-6 240
Yuan, Yougen Mon-O-1-4-2 108
Yue, Yang Mon-P-1-1-7 113
Yunusova, Yana Tue-SS-2-2-9 175
Wed-SS-1-1-6 183
Yvon, François Wed-O-3-1-3 215


## Z


Zafeiriou, Stefanos Mon-SS-1-1-1 109
Mon-SS-1-1-2 109
Mon-SS-1-1-4 110
Mon-SS-1-1-8 110
Mon-SS-2-1-1 125
Mon-SS-2-1-10 127
Zahner, Katharina Wed-O-1-3-2 180
Zahorian, Stephen Tue-O-2-1-5 159
Zajíc, Zbyněk Wed-SS-3-1-1 219
Zarazinski, Adam Tue-P-1-1-8 149
Zatvornitskiy, Alexander Wed-P-3-2-5 224
Zdansky, Jindrich Wed-O-1-2-2 179
Zegers, Jeroen Tue-SS-2-1-5 165
Zeghidour, Neil Mon-P-2-4-5 136
Wed-P-2-4-13 214
Żelasko, Piotr Wed-P-2-4-8 213
Zelinka, Jan Wed-SS-3-1-1 219
Zellers, Margaret Mon-P-2-1-14 130
Zen, Heiga Tue-P-1-4-7 157
Zeng, Biao Thu-P-2-4-5 260
Zeng, Junlin Mon-SS-1-1-12 111
Zeng, Xiangyu Thu-O-2-1-5 247
Zenkel, Thomas Mon-O-2-1-3 120
Wed-P-1-2-9 190
Zeyer, Albert Mon-O-1-1-2 104
Zhan, Juhong Mon-P-1-1-4 113
Zhan, Yuqing Mon-P-1-1-4 113
Zhang, Chao Tue-O-1-1-2 140
Wed-O-1-1-2 177
Zhang, Chunlei Mon-P-2-4-6 137
Zhang, Chunxiao Tue-P-2-3-8 172
Zhang, Cuiling Tue-O-2-4-4 162
Zhang, Dajie Mon-SS-1-1-1 109
Mon-SS-1-1-2 109
Mon-SS-1-1-4 109
Mon-SS-1-1-8 110
Mon-SS-2-1-1 125
Mon-SS-2-1-10 127
Zhang, Gaoyan Tue-O-2-5-1 163
Zhang, Hao Thu-P-1-1-5 238
Tue-O-2-1-5 159
Zhang, Hong Wed-P-1-3-9 192
Zhang, Hui Mon-O-1-2-6 106
Tue-P-1-2-4 152
Zhang, Jiacen Thu-P-2-1-11 253
Zhang, Jinsong Mon-P-1-3-11 119
Thu-P-2-4-1 259
Wed-P-1-4-9 194
Wed-P-2-3-4 210
Wed-P-3-4-9 228
Zhang, Junbo Tue-P-2-2-5 169
Wed-P-1-2-2 188
Wed-P-2-1-8 205
Zhang, Kailai Thu-P-1-2-11 241
Thu-P-1-2-3 240
Tue-O-2-2-3 160
Zhang, Linjuan Tue-P-2-2-11 170
Zhang, Liwen Thu-P-1-2-2 240
Zhang, Minghui Tue-O-2-4-1 162
Zhang, Mingyang Wed-P-1-1-3 186
Zhang, Pengyuan Thu-P-1-2-6 240
Thu-P-1-3-4 242
Wed-P-3-2-2 223
Zhang, Sheng Mon-O-1-3-1 106
Zhang, ShiLiang Mon-P-2-4-3 136
Wed-P-2-4-14 214
Zhang, Shuyuan Mon-O-1-1-6 105
Zhang, Teng Thu-P-1-2-11 241
Tue-O-2-2-3 160
Zhang, Wang Tue-SS-2-2-6 174
Zhang, Wei Wed-P-3-4-9 228
Zhang, Xiao-Lei Tue-P-1-1-14 151
Zhang, Xiaotong Thu-P-2-2-9 256
Zhang, Xuedong Wed-P-3-3-8 226
Zhang, Xueliang Mon-P-1-3-5 118
Tue-P-1-2-4 152
Zhang, Xueyang Wed-SS-3-1-2 219
Zhang, Yang Mon-SS-2-1-5 126
Zhang, Yaping Thu-P-1-1-1 237
Zhang, Yike Thu-P-1-3-4 242
Zhang, Ying Mon-O-1-1-5 105
Zhang, Yixin Wed-P-1-4-9 194
Zhang, Zixing Mon-P-1-2-7 116
Mon-SS-1-1-7 110
Thu-O-1-3-1 231
Wed-P-2-3-12 211
Zhao, Bin Tue-O-2-5-1 163
Zhao, Guanlong Mon-O-2-4-1 124
Wed-O-3-4-6 219
Zhao, Rui Wed-P-3-2-7 224
Zhao, Xuemin Mon-P-2-1-3 128
Zhao, Yiqin Mon-P-1-2-7 116
Zhao, Yuanjun Mon-P-2-2-1 130
Zhao, Yuanyuan Wed-O-1-1-5 178
Zhao, Yunxin Thu-O-1-1-6 230
Zhao, Ziping Mon-P-1-2-7 116
Zheng, Jimeng Wed-P-2-4-4 212
Zheng, Nengheng Thu-P-2-4-2 259
Zheng, Thomas Fang Thu-P-2-2-9 256
Zheng, Yibin Mon-O-1-2-4 106
Tue-O-1-2-3 141
Tue-O-1-2-6 142
Wed-P-2-2-11 208
Zheng, Yu Mon-P-1-2-7 116
Zhong, Jinghua Mon-O-2-4-5 124
Zhong, Shun-Chang Mon-SS-2-1-8 126
Zhou, Cong Wed-P-1-1-2 186
Zhou, Ming Wed-P-1-1-4 186
Zhou, Shiyu Mon-P-2-4-12 138
Mon-P-2-4-7 137
Zhou, Xiao Wed-P-2-2-10 208
Zhou, Yingbo Thu-P-2-3-12 259
Zhou, Zhi-Ping Wed-P-2-2-10 208
Zhu, Anna Thu-O-2-4-5 251
Thu-O-2-4-6 251
Zhu, Wei-Ping Thu-P-1-1-2 237
Zhu, Yingke Thu-P-2-1-3 251
Zhu, Ziwei Mon-O-1-4-3 108
Zhukova, Yulia Tue-P-1-1-8 149
Zisserman, Andrew Thu-O-2-3-3 249
Thu-P-1-1-6 238
Tue-P-1-1-6 149
Ziyu, Xiong Tue-P-2-1-13 168
Žmolíková, Kateřina Wed-SS-3-1-3 220
Zong, Wenqing Tue-O-2-5-3 163
Zou, Yuexian Mon-P-2-5-1 138
Mon-SS-1-1-10 111

# Venue Maps


HICC LAYOUT
GROUND FLOOR
Dock Area
Stair Case
Stair Case
Kitchen Pantry
Kitchen Pantry
Hall-1
15m x 29m
49' x 95'
Hall-2
15m x 29m
49' x 95'
Hall-3
33m x 58m
108' x 190'
Hall-4
33m x 58m
108' x 190'
Hall-5
15m x 29m
49' x 95'
Hall-6
15m x 29m
49' x 95'
Corridor
Corridor
North Terrace
South Terrace
Smoking Zone
Mens Washroom
Womens Washroom
Womens Washroom
Mens Washroom
Stair A
Lift Lobby
Clock Room
MG.07
Business Centre
Corridor
Organizer Suite 2
Organizer Suite 1
Corridor
Lift Lobby
Stair A
Reception
Registration Centre
33m x 13m
108' x 42'
MAIN ENTRANCE
MR G.06
8.7m X 13.8m
28.10' x 45'
MR G.05
8.9m X 13.8m
29'x 45'
MR G.04
8.9m X 13.8m
29'x 45'
MR G.03
8.9m X 13.8m
29'x 45'
MR G.02
8.9m X 13.8m
29'x 45'
MR G.01
8.75m X 13.8m
28.10' x 45'
T: + 91 9642326217
E: sales@hicc.com
W: www.hicc.com
HYDERABAD INTERNATIONAL CONVENTION CENTRE
Managed by ACCOR HOTELS

HICC LAYOUT - FIRST FLOOR

**T**: + 91 9642326217
**E**: sales@hicc.com
**W**: www.hicc.com

HYDERABAD INTERNATIONAL CONVENTION CENTRE
Managed by ACCOR HOTELS

HICC LAYOUT - HICC SECOND FLOOR

**T**: + 91 9642326217
**E**: sales@hicc.com
**W**: www.hicc.com

HYDERABAD INTERNATIONAL CONVENTION CENTRE
Managed by ACCOR HOTELS

# Daily Schedule

## Tutorials

| INTERSPEECH 2018 - September 2, 2018 - PROGRAM AT A GLANCE | | | | | |
|---|---|---|---|---|---|
| | MR G.01-G.03 | MR G.04-G.06 | MR 1.03 | MR 1.04 | |
| 08:00-08:30 | | | | | REGISTRATION at Ground Floor |
| 08:30-09:00 | | | | | |
| 09:00-09:30 | **Tutorial 1:** Deep Learning Based Speech Separation | **Tutorial 2:** End-To-End Models for Automatic Speech Recognition | **Tutorial 3:** Spoofing Attacks in Automatic Speaker Verification: Analysis and Countermeasures | **Tutorial 4:** Spoken Dialog Technology for Education Domain Applications | |
| 09:30-10:00 | | | | | |
| 10:00-10:30 | | | | | |
| 10:30-11:00 | Refreshment Break (First Floor Foyer) | | | | |
| 11:00-11:30 | **Tutorial 1:** Deep Learning Based Speech Separation | **Tutorial 2:** End-To-End Models for Automatic Speech Recognition | **Tutorial 3:** Spoofing Attacks in Automatic Speaker Verification: Analysis and Countermeasures | **Tutorial 4:** Spoken Dialog Technology for Education Domain Applications | |
| 11:30-12:00 | | | | | |
| 12:00-12:30 | | | | | |
| 12:30:13:00 | Lunch Break | | | | |
| 13:00-13:30 | | | | | |
| 13:30-14:00 | | | | | |
| 14:00-14:30 | **Tutorial 5:** Information Theory of Deep Learning: What do the Layers of Deep Neural Networks Represent ? | **Tutorial 6:** Articulatory Representations: Measurement, Estimation, and Application to State-of-the-Art Automatic Speech Recognition | **Tutorial 7:** Vocal Music Processing for Music Information Retrieval (MIR) | **Tutorial 8:** Multimodal Speech and Audio Processing in Audio-Visual Human-Robot Interaction | |
| 14:30-15:00 | | | | | |
| 15:00-15:30 | | | | | |
| 15:30-16:00 | Refreshment Break (First Floor Foyer) | | | | |
| 16:00-16:30 | **Tutorial 5:** Information Theory of Deep Learning: What do the Layers of Deep Neural Networks Represent ? | **Tutorial 6:** Articulatory Representations: Measurement, Estimation, and Application to State-of-the-Art Automatic Speech Recognition | **Tutorial 7:** Vocal Music Processing for Music Information Retrieval (MIR) | **Tutorial 8:** Multimodal Speech and Audio Processing in Audio-Visual Human-Robot Interaction | |
| 1630-17:00 | | | | | |
| 17:00-17:30 | | | | | |
| 17:30-18:00 | | | | | |

# Day 1 (Monday)

| INTERSPEECH 2018 - September 3, 2018 - PROGRAM AT A GLANCE | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Hall 1 | Hall 2 | Hall 3 | MR G.01-G.02 | MR G.03-G.04 | MR G.05-G.06 | MR 1.01-1.02 | Hall 4 - 6 | | |
| 08:00-08:30 | | | | | | | | | | |
| 08:30-09:00 | | | | | | | | | | |
| 09:00-09:30 | Welcome Refreshment (Hall 5 - 6) | | | | | | | | REGISTRATION at Ground Floor | Speaker Check-in (MR 1.07) |
| 09:30-10:00 | | | | | | | | | | |
| 10:00-10:30 | | | **Opening Ceremony** | | | | | | | |
| 10:30-11:00 | | | | | | | | | | |
| 11:00-11:30 | | | **ISCA Medal Talk: Bishnu S. Atal** | | | | | | | |
| 11:30-12:00 | | | | | | | | | | |
| 12:00-12:30 | Lunch Break | | | | | | | | | |
| 12:30:13:00 | | | | | | | | | | |
| 13:00-13:30 | | | | | | | | | | |
| 13:30-14:00 | | | | | | | | | | |
| 14:00-14:30 | **O-1-2:** Prosody Modeling and Generation | **O-1-3:** Speaker Verification I | **O-1-1:** End-to-End Speech Recognition | **O-1-4:** Spoken Term Detection | | **S&T-1-1**: Show and Tell 1 | **SS-1-1:** The INTERSPEECH 2018 Computational Paralinguistics ChallengE | **Posters:** P-1-1 P-1-2 P-1-3 P-1-4 | | |
| 14:30-15:00 | | | | | | | | | | |
| 15:00-15:30 | | | | | | | | | | |
| 15:30-16:00 | | | | | | | | | | |
| 16:00-16:30 | Refreshment Break (Hall 5 - 6) | | | | | | | | | |
| 1630-17:00 | **O-2-1:** ASR Systems and Technologies | **O-2-2:** Deception, Personality, and Culture Attribute | | **O-2-3:** Automatic Detection and Recognition of Voice and Speech Disorders | **O-2-4:** Voice Conversion | **S&T-2-1:** Show and Tell 2 | **SS-2-1:** The INTERSPEECH 2018 Computational Paralinguistics ChallengE | **Posters:** P-2-1 P-2-2 P-2-3 P-2-4 P-2-5 | | |
| 17:00-17:30 | | | | | | | | | | |
| 17:30-18:00 | | | | | | | | | | |
| 18:00-18:30 | | | | | | | | | | |
| 18:30-19:00 | | | **Cultural Programme** | | | | | | | |
| 19:00-19:30 | | | | | | | | | | |
| 19:30-20:00 | Welcome Reception (Novotel Gardens) | | | | | | | | | |
| 20:00-20:30 | | | | | | | | | | |
| 20:30-21:00 | | | | | | | | | | |
| 21:00-21:30 | | | | | | | | | | |

**P-1-1:** Speech Segments and Voice Quality
**P-1-2:** Speaker State and Trait
**P-1-3:** Deep Learning for Source Separation and Pitch Tracking
**P-1-4:** Acoustic Analysis-Synthesis of Speech Disorders

**P-2-1:** Spoken Dialogue Systems and Conversational Analysis
**P-2-2:** Spoofing Detection
**P-2-3:** Speech Analysis and Representation
**P-2-4:** Sequence Models for ASR
**P-2-5:** Source Separation and Spatial Analysis

# Day 2 (Tuesday)

| INTERSPEECH 2018 - September 4, 2018 - PROGRAM AT A GLANCE | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Hall 1 | Hall 2 | Hall 3 | MR G.01-G.02 | MR G.03-G.04 | MR G.05-G.06 | MR 1.01-1.02 | Hall 4 - 6 | REGISTRATION at Ground Floor | Speaker Check-in (MR 1.07) |
| 08:00-08:30 | | | | | | | | | | |
| 08:30-09:00 | | | **Plenary Talk-1: Jacqueline Vaissière** | | | | | | | |
| 09:00-09:30 | | | | | | | | | | |
| 09:30-10:00 | **Refreshment Break (Hall 4 - 5)** | | | | | | | | | |
| 10:00-10:30 | **O-1-2:** Statistical Parametric Speech Synthesis | **O-1-3:** Emotion Modeling | **O-1-1:** Acoustic Model Adaptation | **O-1-4:** Models of Speech Perception | **O-1-5:** Multimodal Dialogue Systems | **S&T-1-1:** Show and Tell 3 | **SS-1-1:** Speech Recognition for Indian Languages | **Posters: P-1-1 P-1-2 P-1-3 P-1-4** | | |
| 10:30-11:00 | | | | | | | | | | |
| 11:00-11:30 | | | | | | | | | | |
| 11:30-12:00 | | | | | | | | | | |
| 12:00-12:30 | | | **Perspective Talk-1: Dilek Hakkani-Tur** | | | | | | | |
| 12:30:13:00 | **Industry Presentation-2: JD** | **Industry Presentation-3: Uniphore** | **Industry Presentation-1: Amazon** | | | | | | | |
| 13:00-13:30 | **Lunch Break** | | | | | | | | | |
| 13:30-14:00 | | | | | | | | | | |
| 14:00-14:30 | | | | | | | | | | |
| 14:30-15:00 | **O-2-2:** Audio Events and Acoustic Scenes | **O-2-3:** Speaker Diarization | **O-2-1:** Dereverberation | **O-2-4:** Phonation | **O-2-5:** Cognition and Brain Studies | **S&T-2-1:** Show and Tell 4 | **SS-2-1:** Deep Neural Networks: How Can We Interpret What They Learned? | **Posters: P-2-1 P-2-2 P-2-3 SS-2-2 P-2-5** | | |
| 15:00-15:30 | | | | | | | | | | |
| 15:30-16:00 | | | | | | | | | | |
| 16:00-16:30 | | | | | | | | | | |
| 1630-17:00 | **Refreshment Break (Hall 5 - 6)** | | | | | | | | | |
| 17:00-17:30 | | | **Perspective Talk-2: Bhuvana Ramabhadran** | | | | | | | |
| 17:30-18:00 | | | **ISCA General Assembly** | | | | | | | |
| 18:00-18:30 | | | | | | | | | | |
| 18:30-19:00 | | | | | | | | | | |
| 19:00-19:30 | **Students' Reception (Club House, Jesmin, Ella Hotel, Gachibowli)** | | | | | | | | | |
| 19:30-20:00 | | | | | | | | | | |
| 20:00-20:30 | **Reviewers' Reception (Kebab Pavilion, Ella Hotel, Gachibowli)** | | | | | | | | | |
| 20:30-21:00 | | | | | | | | | | |
| 21:00-21:30 | | | | | | | | | | |

**P-1-1:** Speaker Verification II
**P-1-2:** Novel Approaches to Enhancement
**P-1-3:** Syllabification, Rhythm, and Voice Activity Detection
**P-1-4:** Selected Topics in Neural Speech Processing

**P-2-1:** Speech and Singing Production
**P-2-2:** Robust Speech Recognition
**P-2-3:** Applications in Education and Learning
**SS-2-2:** Integrating Speech Science and Technology for Clinical Applications
**P-2-5:** Speaker Characterization and Analysis

# Day 3 (Wednesday)

| INTERSPEECH 2018 - September 5, 2018 - PROGRAM AT A GLANCE | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Hall 1** | **Hall 2** | **Hall 3** | **MR G.01-G.02** | **MR G.03-G.04** | **MR G.05-G.06** | **MR 1.01-1.02** | **Hall 4 - 6** | | |
| **08:00-08:30** | | | | | | | | | REGISTRATION at Ground Floor | Speaker Check-in (MR 1.07) |
| **08:30-09:00** | | | **Plenary Talk-2: Hervé Bourlard** | | | | | | | |
| **09:00-09:30** | | | | | | | | | | |
| **09:30-10:00** | **Refreshment Break (Hall 5 - 6)** | | | | | | | | | |
| **10:00-10:30** | **O-1-2:** Language Identification | **O-1-3:** Production of Prosody | **O-1-1:** Novel Neural Network Architectures for Acoustic Modelling | **O-1-4:** Speech Intelligibility and Quality | **SS-1-1:** Integrating Speech Science and Technology for Clinical Applications | **S&T-1-1:** Show and Tell 5 | **SS-1-2:** Speech Technologies for Code-Switching in Multilingual Communities | **Posters:** **P-1-1** **P-1-2** **P-1-3** **P-1-4** | | |
| **10:30-11:00** | | | | | | | | | | |
| **11:00-11:30** | | | | | | | | | | |
| **11:30-12:00** | | | | | | | | | | |
| **12:00-12:30** | | | **Perspective Talk-3: Nima Mesgarani** | | | | | | | |
| **12:30:13:00** | **Industry Presentation-5: Xiaomi** | **Industry Presentation -6: MeitY** | **Industry Presentation-4: Microsoft** | | | | | | | |
| **13:00-13:30** | **Lunch Break** | | | | | | | | | |
| **13:30-14:00** | | | | | | | | | | |
| **14:00-14:30** | | | | | | | | | | |
| **14:30-15:00** | **O-2-2:** Speaker Verification Using Neural Network Methods I | **O-2-3:** Speech Perception in Adverse Conditions | **O-2-1:** Recurrent Neural Models for ASR | **O-2-4:** Measuring Pitch and Articulation | **O-2-5:** Speech and Language Analytics for Mental Health | **S&T-1-2:** Show and Tell 6 | **SS-2-1:** Spoken CALL Shared Task, Second Edition | **Posters:** **P-2-1** **P-2-2** **P-2-3** **P-2-4** | | |
| **15:00-15:30** | | | | | | | | | | |
| **15:30-16:00** | | | | | | | | | | |
| **16:00-16:30** | | | | | | | | | | |
| **1630-17:00** | **Refreshment Break (Hall 5 - 6)** | | | | | | | | | |
| **17:00-17:30** | **O-3-1:** Zero-resource Speech Recognition | **O-3-2:** Spatial and Phase Cues for Source Separation and Speech Recognition | | **O-3-3:** Dialectal Variation | **O-3-4:** Spoken Corpora and Annotation | | **SS-3-1:** The First DIHARD Speech Diarization Challenge | **Posters:** **P-3-1** **P-3-2** **P-3-3** **P-3-4** | | |
| **17:30-18:00** | | | | | | | | | | |
| **18:00-18:30** | | | | | | | | | | |
| **18:30-19:00** | | | | | | | | | | |
| **19:00-19:30** | **Cultural Programme** **+ Banquet (Hall 5 - 6)** | | | | | | | | | |
| **19:30-20:00** | | | | | | | | | | |
| **20:00-20:30** | | | | | | | | | | |
| **20:30-21:00** | | | | | | | | | | |
| **21:00-21:30** | | | | | | | | | | |

**P-1-1:** Voice Conversion and Speech Synthesis
**P-1-2:** Extracting Information from Audio
**P-1-3:** Signal Analysis for the Natural, Biological and Social Sciences
**P-1-4:** Speech Prosody

**P-2-1:** Adjusting to Speaker, Accent, and Domain
**P-2-2:** Speech Synthesis Paradigms and Methods
**P-2-3:** Second Language Acquisition and Code-switching
**P-2-4:** Topics in Speech Recognition

**P-3-1:** Text Analysis, Multilingual Issues and Evaluation in Speech Synthesis
**P-3-2:** Neural Network Training Strategies for ASR
**P-3-3:** Application of ASR in Medical Practice
**P-3-4:** Source and Supra-segmentals

# Day 4 (Thursday)

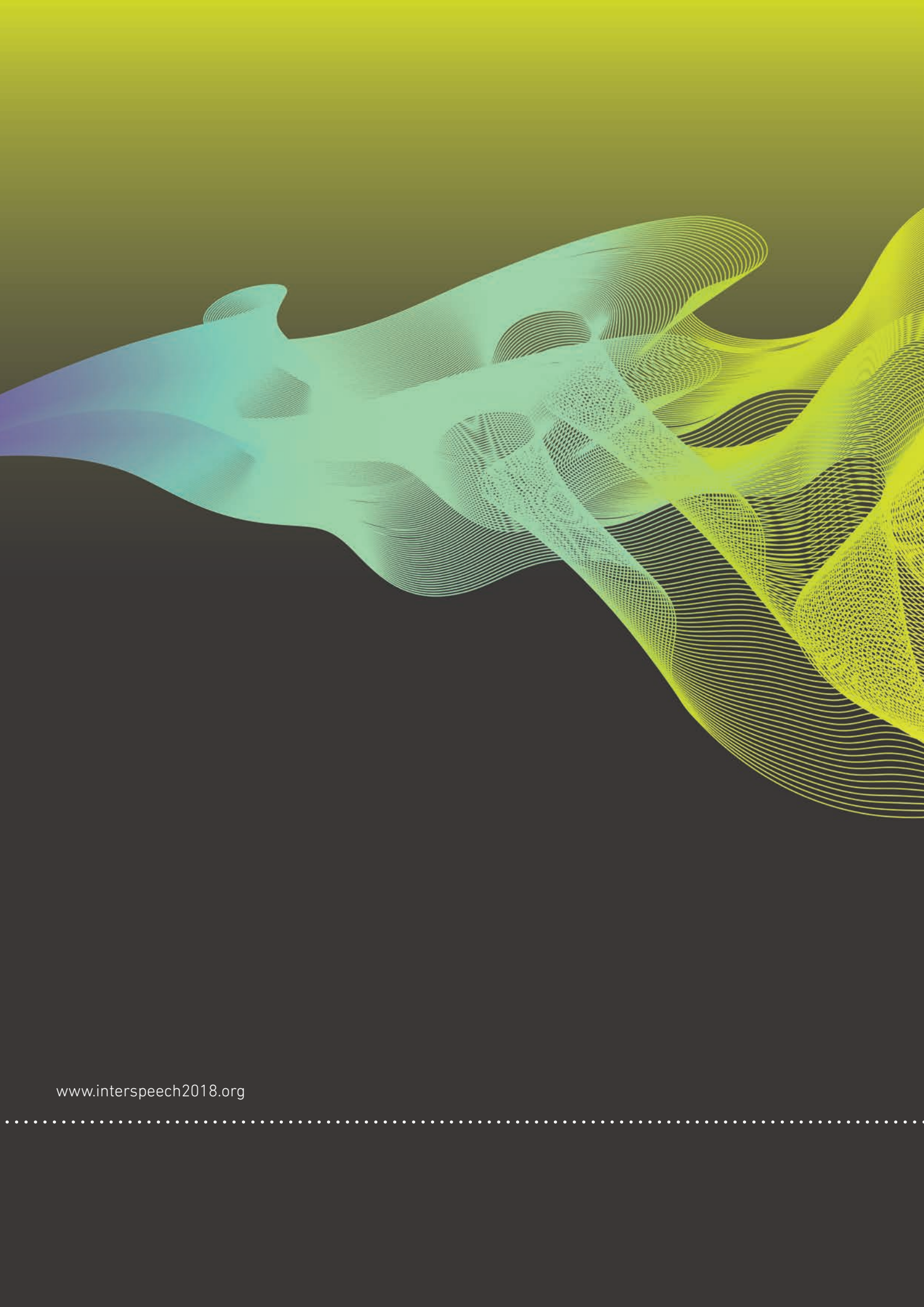
| INTERSPEECH 2018 - September 6, 2018 - PROGRAM AT A GLANCE | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Hall 1** | **Hall 2** | **Hall 3** | **MR G.01-G.02** | **MR G.03-G.04** | **MR G.05-G.06** | **MR 1.01-1.02** | **Hall 4 - 6** | | |
| **08:00-08:30** | | | | | | | | | REGISTRATION at Ground Floor | Speaker Check-in (MR 1.07) |
| **08:30-09:00** | | | **Plenary Talk-3: Helen Meng** | | | | | | | |
| **09:00-09:30** | | | | | | | | | | |
| **09:30-10:00** | **Refreshment Break (Hall 5 - 6)** | | | | | | | | | |
| **10:00-10:30** | **O-1-2:** Expressive Speech Synthesis | **O-1-3:** Representation Learning for Emotion | **O-1-1:** Distant ASR | **O-1-4:** Articulatory Information, Modeling and Inversion | **SS-1-1:** Novel Paradigms for Direct Synthesis Based on Speech-Related Biosignals | **S&T-1-1:** Show and Tell 7 | **SS-1-2:** Low Resource Speech Recognition Challenge for Indian Languages | **Posters: P-1-1 P-1-2 P-1-3 P-1-4** | | |
| **10:30-11:00** | | | | | | | | | | |
| **11:00-11:30** | | | | | | | | | | |
| **11:30-12:00** | | | | | | | | | | |
| **12:00-12:30** | | | **Perspective Talk-4: Sriram Ganapathy** | | | | | | | |
| **12:30:13:00** | **Industry Presentation-8: Baidu** | **Industry Presentation-9: Nvidia** | **Industry Presentation-7: Samsung** | | | | | | | |
| **13:00-13:30** | **Lunch Break** | | | | | | | | | |
| **13:30-14:00** | | | | | | | | | | |
| **14:00-14:30** | | | | | | | | | | |
| **14:30-15:00** | **O-2-2:** Source Separation from Monaural Input | **O-2-3:** Multimodal Systems | **O-2-1:** Spoken Language Understanding | **O-2-4:** Coding | | | | **Posters: P-2-1 P-2-2 P-2-3 P-2-4** | | |
| **15:00-15:30** | | | | | | | | | | |
| **15:30-16:00** | | | | | | | | | | |
| **16:00-16:30** | | | | | | | | | | |
| **1630-17:00** | **Refreshment Break (Hall 5 - 6)** | | | | | | | | | |
| **17:00-17:30** | | | **Closing Ceremony** | | | | | | | |
| **17:30-18:00** | | | | | | | | | | |

**P-1-1:** Deep Enhancement
**P-1-2:** Acoustic Scenes and Rare Events
**P-1-3:** Language Modeling
**P-1-4:** Speech Pathology, Depression, and Medical Applications

**P-2-1:** Speaker Verification Using Neural Network Methods II
**P-2-2:** Emotion Recognition and Analysis
**P-2-3:** Acoustic Modelling
**P-2-4:** Speech and Speaker Perception

## INTERSPEECH 2018 - PROGRAM AT A GLANCE

| | Sep. 2, 2018 | Sep. 3, 2018 | Sep. 4, 2018 | Sep. 5, 2018 | Sep. 6, 2018 |
|---|---|---|---|---|---|
| 08:00-08:30 | Registration | | Registration | Registration | Registration |
| 08:30-09:00 | | | Plenary Talk-1: Jacqueline Vaissière | Plenary Talk-2: Hervé Bourlard | Plenary Talk-3: Helen Meng |
| 09:00-09:30 | Forenoon Tutorial | Registration | | | |
| 09:30-10:00 | | Welcome Refreshment | Refreshment Break | Refreshment Break | Refreshment Break |
| 10:00-10:30 | | Opening Ceremony | Technical Session-3 (36 Orals, 51 Posters and 5 Show-and-Tell) | Technical Session-5 (33 Orals, 53 Posters and 5 Show-and-Tell) | Technical Session-8 (38 Orals, 45 Posters and 4 Show-and-Tell) |
| 10:30-11:00 | Refreshment Break | | | | |
| 11:00-11:30 | | ISCA Medal Talk: Bishnu S. Atal | | | |
| 11:30-12:00 | Forenoon Tutorial | | | | |
| 12:00-12:30 | | Lunch Break | Perspective Talk-1: *Dilek* Hakkani-Tur | Perspective Talk-3: Nima Mesgarani | Perspective Talk-4: Sriram Ganapathy |
| 12:30:13:00 | Lunch Break | | Industry Presentation: Amazon,JD & Uniphore | Industry Presentation: Microsoft, Xiaomi & Meity | Industry Presentation: Samsung, Baidu & Nvidia |
| 13:00-13:30 | | | Lunch Break | Lunch Break | Lunch Break |
| 13:30-14:00 | | | | | |
| 14:00-14:30 | Afternoon Tutorial | Technical Session-1 (33 Orals, 42 Posters and 5 Show-and-Tell) | | | |
| 14:30-15:00 | | | Technical Session-4 (36 Orals, 54 Posters and 5 Show-and-Tell) | Technical Session-6 (36 Orals, 55 Posters and 5 Show-and-Tell) | Technical Session-9 (24 Orals and 48 Posters) |
| 15:00-15:30 | | | | | |
| 15:30-16:00 | Refreshment Break | | | | |
| 16:00-16:30 | | Refreshment Break | | | |
| 1630-17:00 | Afternoon Tutorial | | Refreshment Break | Refreshment Break | Refreshment Break |
| 17:00-17:30 | | Technical Session-2 (32 Orals, 62 Posters and 5 Show-and-Tell) | Perspective Talk-2: Bhuvana Ramabhadran | Technical Session-7 (31 Orals and 40 Posters) | Closing Ceremony |
| 17:30-18:00 | | | ISCA General Assembly | | |
| 18:00-18:30 | | | | | |
| 18:30-19:00 | | Cultural Programme | | | |
| 19:00-19:30 | | | Students' / Reviewers' Reception | Cultural Programme + Banquet | |
| 19:30-20:00 | | Welcome Reception | | | |
| 20:00-20:30 | | | | | |
| 20:30-21:00 | | | | | |
| 21:00-21:30 | | | | | |



www.interspeech2018.org