CONFERENCE PROGRAM & ABSTRACT BOOK

INTERSPEECH

Understanding Speech Processing in Humans and Machines

September 8-12, 2016

The Hyatt Regency | San Francisco, California

INTERSPEECH 2016 SPONSORS

PLATINUM

www.apple.com

DIAMOND

amazon alexa
www.amazon.jobs/
interspeech2016

Google
http://research.google.com

Microsoft
www.microsoft.com/en-us

ebay™
www.ebay.com

GOLD

facebook
https://research.facebook.com

YAHOO! JAPAN
www.yahoo.co.jp

Baidu Research
http://research.baidu.com

IBM Research
www.research.ibm.com

CIRRUS LOGIC®
www.cirrus.com/en

SILVER

DATATANG
http://datatang.com/en

NUANCE®
http://research.nuance.com

BRONZE

Speechocean Limited

Yandex

Raytheon Technologies

SUPPORTER

Disney Research

EML European Media
Laboratory GmbH

University of Washington's
Master of Science in
Computational Linguistics

EXHIBITORS

Appen
Apple
Amazon Alexa
Beijing Huiting Technology Co., Ltd.
Cobalt Speech and Language
Datatang Technology Inc.
eBay
ELSEVIER
Furhat Robotics
Globalme Language & Technology
Google
innoetics
ISCA
Interspeech 2017
Interspeech 2018
Linguistic Data Consortium
Microsoft
Oben
Speechocean Limited
Voxygen
Yahoo! JAPAN

INTERSPEECH

Understanding Speech Processing in Humans and Machines

San Francisco | September 8–12, 2016

Hyatt Regency San Francisco
San Francisco, California

www.interspeech2016.org

http://www.isca-speech.org
secretariat@isca-speech.org

ISSN 2308-457X Proceedings USB
ISSN 1990-9770 Proceedings in ISCA Archive
Editors: Nelson Morgan, Panayiotis Georgiou, Shrikanth Narayanan, Florian Metze

Abstracts and Proceedings USB Production by
Causal Productions Pty Ltd
PO Box 100, Rundle Mall
SA 5000, Australia
http://www.causalproductions.com
info@causalproductions.com

TABLE OF CONTENTS

INTERNATIONAL SPEECH COMMUNICATION ASSOCIATION (ISCA)

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 organization to a fully-independent and self-supporting international association.

At the General Assembly 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.

MESSAGE FROM THE ISCA PRESIDENT

Welcome to San Francisco — where INTERSPEECH makes a return to North America!

I am indeed honored to pen the first words as a welcome message. In the past year, research in speech communication science and technology has continued to thrive in the ISCA community and all over the world. On this occasion, I share with you the excitement of our community as we gather again at our annual conference.

INTERSPEECH 2016 is special. We embrace a theme of 'understanding speech processing in humans and machines' in San Francisco, the City that Knows How. During the conference, the community will honour Dr. John Makhoul as the recipient of the 2016 ISCA Medal for Scientific Achievement for leadership and extensive contributions to speech and language processing. We will also celebrate our members' achievements by recognizing six ISCA Fellows 2016, who include Mary Harper, Helen Meng, Shri Narayanan, Steve Renals, Tanja Schultz, and Chiu-yu Tseng. Please join me in giving them the warmest congratulations!

INTERSPEECH 2016 marks the 17th Annual Conference of ISCA, which continues the success of recent events. I feel privileged to be part of conference preparation as the ISCA President in the first year of my term and as one of the 67 area chairs. We received a record number of 1585 paper submissions. The technical program committee, led by the Technical Chairs Panayiotis (Panos) Georgiou and Shrikanth (Shri) Narayanan, deserve 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, INTERSPEECH 2016 has increased the number of reviews per paper from three to four on average. A total of 1291 reviewers contributed to the review process this year. A big thanks to all of you!

Organizing an INTERSPEECH event takes enormous courage, endurance and dedication, I would like to express my gratitude and appreciation to the General Chair, Nelson Morgan, who led an experienced organization team to bring INTERSPEECH to San Francisco for the first time.

Finally, I do hope that you have an enjoyable and productive time in San Francisco, and that you will leave with fond memories of INTERSPEECH 2016. With my best wishes for a successful conference!

Haizhou Li
ISCA President

MESSAGE FROM THE GENERAL CHAIR

Welcome to Interspeech 2016 in San Francisco!

2016 marks the occasion of the 17th annual conference of the International Speech Communication Association (ISCA), and we're proud to hold it in San Francisco! The Bay Area has long been an international focal point for technology development, with Silicon Valley and its two great universities "in the neighborhood." As speech processing becomes a part of the computing mainstream, it has an even more central role in developments at the Bay Area's many related companies, large and small. Many of these companies are represented at this conference.

But even as work on speech processing becomes increasingly focused on applications, our community's need to understand the nature of speech increases. Why, after decades of work, are aspects of machine speech processing so difficult, and in some circumstances, so fragile? This question, along with our community's deep interest in understanding human communication, leads us to push for answers to more basic questions. For these reasons, we have chosen the theme, "Understanding speech processing in humans and machines" for this conference.

The main conference is augmented by an exciting group of special sessions and special events, including a repeat of the successful forensics special event. Overall, there are 60 different oral sessions and 40 poster sessions, with close to 800 refereed and accepted papers, which on average received almost 4 reviews each.

The Interspeech 2016 venue is the primary conference hotel, the Hyatt Regency at the Embarcadero in San Francisco. Here you are steps from the San Francisco Bay, and a short walk (or cable car ride) from many restaurants, ranging from simple fare to some of the best in the City (a city that the magazine "Bon Appetit" referred to as the "best food city in the country now"). Starting from the international terminal at San Francisco Airport, you can reach the hotel directly using the local railway system, BART (get off at the Embarcadero station). From the hotel, it is a short walk to transportation that can take you many other places in the area; for instance, from the ferry terminal near the hotel you can take a boat to nearby beautiful Marin County, and you can take BART to the downtown Berkeley station, which is just downhill from the campus of the University of California at Berkeley (and a block from ICSI). For those with some time to spend in the area, you'll also find deserts, forests, mountains, and ocean within a few hours' drive.

I want to close this brief welcome by thanking the entire organizing committee for their long efforts to create an event that we hope will be memorable for you. I particularly want to thank the Technical Chairs, Shri and Panos of the University of Southern California, who not only took on the toughest jobs in one of the events, but who through their own network of past and current colleagues, provided many of the members of the organizing committee. I also very much thank Kate Porter and her team and Conference Solutions, without whom this event would not have been possible – trying to stage a multi-day event for 1000+ people using only researchers to pull it together would be almost as big a challenge as solving robust speech recognition! And finally, I thank the International Computer Science Institute (ICSI) and its former directors, Deborah Crawford and Roberto Pieraccini, for its support over the years that it took to pull this together. And of course to ISCA itself for making these conferences possible.

Welcome to California, to San Francisco, and to our conference!

Nelson Morgan
International Computer Science Institute (ICSI), University of California, Berkeley, and UpRise Campaigns
General Chair

MESSAGE FROM THE TECHNICAL CHAIRS

Welcome to California, to San Francisco and to Interspeech 2016!

When we signed up to organize the Technical Program of Interspeech 2016 we knew it was hard work, but we also knew that we would be working with many individuals who have over the years been pillars of our community and that we would be doing our little piece to give back to this wonderful interdisciplinary community. This was precisely the experience: hard work, excellent team work, and supporting colleagues.

We cannot emphasize and thank enough the area chairs, reviewers, ISCA board members, and authors for their many contributions all along the various stages of creating an exciting and high quality technical program.

Panayiotis (Panos) Georgiou

Shrikanth (Shri) Narayanan

The first step as we set out to organize the program for IS2016 was to learn from those who have been there before. Bernd Möbius & Elmar Nöth (IS016) and Lori Lamel (IS2013) were exceptionally helpful in guiding us including in aspects of forming our committees to learning the conference management software. We also gained valuable experience in 2015 by being Area Chairs and participating in the meetings where final decisions were made and the program was created.

One of the first tasks was defining the Scientific Areas for the program. We decided on 13 scientific areas, adding the 13th area of "Speech and Spoken-Language Based Multimodal Processing and Systems" to encourage a more holistic, multimodal, view of speech communication. With the Scientific Areas in place, and several months prior to Interspeech 2015 we already had the core team of the technical program committee in place. The choice of a large and diverse team of scholars and technical experts was made to ensure a broad oversight of each and every paper by multiple experts, to ensure fairness, and increase quality of the final technical program. The team was comprised of the two Technical Program Chairs, two special session chairs (Abeer Alwan and Dilek Hakkani-Tur) and 67 Area Chairs including the ISCA president (Haizhou Li) and other ISCA Board representatives (Martin Cooke, John Hansen, Mark Hasegawa-Johnson, Douglas O'Shaughnessy, Lori Lamel, Keikichi Hirose, Gérard Bailly, Hynek Hermansky) and the Interspeech 2017 Technical Program Chairs (Alan Black, Alexandros Potamianos, Angeliki Metallinou, Bastiaan Kleijn, Bin Ma, Brian Kingsbury, Brian Strope, Carlos Busso, Chi-Chun (Jeremy) Lee, David Traum, Denis Jouvet, Dimitrios Dimitriadis, Engin Erzin, Florian Metze, Geoff Zweig, Heiga Zen, Helen Meng, Helmer Strik, Hema Murthy, Hermann Ney, Horacio Franco, Ingmar Steiner, Isabel Trancoso , Jean-François Bonastre, Joakim Gustafson, Joseph Tepperman, Julia Hirschberg, Julien Epps, Kartik Audhkhasi, Keiichi Tokuda, Keikichi Hirose, Khalil Iskarous, Koichi Shinoda, Liz Shriberg, Malcom Slaney, Mari Ostendorf, Mary Beckman, Mattias Heldner, Ming Li, Murat Akbacak, Murat Saraclar, Najim Dehak, Odette Scharenborg, Olov Engwall, Ozlem Kalinli, Pascale Fung, Prasanta Ghosh, Ramabhadran Bhuvana, Rick Rose, Roland Kuhn, Rolf Carlson, Tan Lee, Tanaya Guha, Tanja Schultz, Tatsuya Kawahara, Umesh Srinivasan, Vikram Ramanarayanan)..

One of the comments we have heard over the many years of participating in Interspeech and other conferences is the desire for more constructive reviewer feedback. From over a year ago we set out to systematically improve that. Our first task was to increase the reviewer pool. We setup a recommendation portal for selecting new qualified reviewers. From combining these and past Interspeech reviewers we constructed a database of 1959 reviewers who were invited to participate in the peer review process; a record 1291 reviewers eventually accepted to review (an impressive 40% increase from past year).

Further to a record number of reviewers, we also had a record number of submissions. From the initial 1,644 submissions 1,585 remained active at the review phase after pruning duplicate submissions and withdrawals.

The increased reviewer pool allowed us to assign 4 reviewers per paper, and through constant monitoring and active re-assignments by the amazing teams of Area Chairs we achieved an average of 3.90 first-level reviews per paper. Further, by encouraging and guiding reviewers to provide detailed comments along specific dimensions, the average reviewer feedback more than doubled to 109 words (from 51).

After the first stage of the review process, each area chair team carefully scrutinized the reviewer recommendations and did a first pass of recommendations. The area chairs initiated discussion (many times in writing) for all borderline papers to guide their

decisions. As a guidance we explicitly requested that Area Chairs strive for recognizing good research and not be constrained by any preset quotas: *"We do not want any good paper to go unpublished for any reason. Hence, we do not want to impose thresholds or quotas!"*

The final decisions were made in a face-to-face meeting of the TPC team held at the campus of the University of Southern California (USC) in Los Angeles on May 18–19, 2016. At least one area chair from each team was present. The ISCA president (also an area chair), 4 other ISCA members and the General Chair and Technical Program Chairs of 2017 were also present. During the meeting we finalized the decisions for all papers and discussed early session assignments.

The detailed reviewer feedback and the large number of very engaged and dedicated Area Chairs resulted in a relatively easy, thoughtful and accurate decision process. In the end, we had 1,585 submissions (after removal of duplicates & withdrawals) and 779 accepted papers corresponding to an overall about 50% acceptance rate. This resulted in 98 technical sessions: 58 Oral and 40 Poster sessions (we note that there was no quality distinction between these two formats). The technical program also includes 13 Special Sessions (Abeer Alwan & Dilek Hakkani-Tur) with an acceptance rate of about 60%. Further, Interspeech 2016 has 8 Tutorials (Eric Fosler-Lussier), 6 Special Events (David Suendermann-Oeft), 23 Show & Tell events (Shiva Sundaram & Nicolas Scheffer) and 6 Satellite Events (Abhinav Sethy & Vikram Ramanarayanan).

Given the complexity of organizing an event like Interspeech, it wouldn't have been possible without the help of many, many people. First we would like to thank the ISCA board members and especially Professors Haizhou Li, Tanja Schultz, Alan Black, and Lori Lamel. In addition to their support and advise, Profs. Schultz and Black were instrumental in convincing us to undertake this responsibility during our interactions in Portland 2012 (Thank you Tanja & Alan!). Nicole Santamaria and Kate Porter from Conference Solutions Inc. have been amazing in their professional support. There is no easy way to describe their help — suffice to say that without them this event wouldn't be possible. Further to handling all local arrangements and logistics, they gently kept us on track on all tasks and deadlines... We also want to thank our lab's incredibly talented administrator at USC, Tanya Acevedo-Lam, for helping us in so many ways including in ensuring a smooth TPC meeting in May. It would also be remiss if we didn't thank our families for putting up with us and our lack-of-mental-presence during the second quarter of this year. Thank you. This list is far from exhaustive and we are almost certainly forgetting many— we are grateful to all!

But, last, but not least, we want to thank the authors and reviewers. Without you this conference wouldn't take place. The heart of the conference is not in the organization, but in the hard work in labs and offices (and more often than not in dark rooms at 2am) by all of you! Thank you for sharing your work with everyone else, for advancing the field, and for joining us in San Francisco in creating an incredible conference of minds.

We are sure that you will find the technical program intellectually rich and stimulating and equally sure that you will love San Francisco and California — and if you have the time, a few days in nature at the Sequoias and a relaxed drive along the incredible California coast on the PCH is highly recommended!

Enjoy!

Panayiotis (Panos) Georgiou and Shrikanth (Shri) Narayanan
Interspeech 2016 Technical Program Chairs

INTERSPEECH 2016 CONFERENCE COMMITTEE

GENERAL CHAIR

Nelson Morgan
International Computer Science Institute (ISCI),
University of California, Berkeley, & UpRise Campaigns, USA

TECHNICAL CHAIRS

Panayiotis (Panos) Georgiou
University of Southern California, USA

Shrikanth (Shri) Narayanan
University of Southern California, USA

PLENARY SESSIONS

Dan Jurafsky
Stanford, USA

Andreas Stolcke
Microsoft Research, USA

SATELLITE WORKSHOPS

Abhinav Sethy
International Business Machines Corporation, USA

Vikram Ramanarayanan
Educational Testing Service, USA

SPECIAL SESSIONS

Abeer Alwan
University of California, Los Angeles, USA

Dilek Hakkani-Tur
Microsoft Research, USA

SPECIAL EVENTS

David Suendermann-Oeft
Educational Testing Service, USA

SHOW & TELL

Nicolas Scheffer
Facebook, USA

Shiva Sundaram
Amazon, USA

TUTORIALS

Eric Fosler-Lussier
Ohio State University, USA

ISCA-SAC LIASON, INTERDISCIPLINARY OUTREACH

Elizabeth Shriberg
SRI International, USA

STUDENT ADVISORY COMMITTEE

Catherine Oertel Genannt Bierbach
KTH Stockholm, Sweden

Lori Lamel
CNRS/LIMSI, France

PUBLICATIONS

Florian Metze
Carnegie Mellon University, USA

PUBLICITY

Carlos Busso
University of Texas at Dallas, USA

Emily Mower Provost
University of Michigan, USA

SPONSORSHIP

Jordan Cohen
Semantic Machines, Spelamode, Kextil, USA

Prem Natarajan
University of Southern California, Information Sciences Institute, USA

Michael Picheny
International Business Machines Corporation, USA

EXHIBITIONS

Sankaranarayanan Ananthakrishnan
Amazon.com, LLC, USA

Bhuvana Ramabhadran
International Business Machines Corporation, USA

LOCAL ARRANGEMENTS

Anita Bounds Morgan
Consultant, USA

Nicolas Scheffer
Facebook, USA

Conference Management

Conference Solutions

COMMITTEES

AREA CHAIRS

Area 1. SPEECH PERCEPTION, PRODUCTION AND ACQUISITION

- Martin Cooke, *University of the Basque Country, Spain*
- Olov Engwall, *KTH Stockholm, Sweden*
- Engin Erzin, *Koc University, Turkey*
- Khalil Iskarous, *University of Southern California, USA*
- Tan Lee, *The Chinese University of Hong Kong, China*
- Odette Scharenborg, *Radboud Universiteit Nijmegen, Netherlands*

Helmer Strik, *Radboud Universiteit Nijmegen, Netherlands*

Area 2. PHONETICS, PHONOLOGY, AND PROSODY

- Mary Beckman, *Ohio State University, USA*
- Mattias Heldner, *Stockholm University, Sweden*
- Keikichi Hirose, *University of Tokyo, Japan*
- Julia Hirschberg, *Columbia University, USA*

Area 3. ANALYSIS OF PARALINGUISTICS IN SPEECH AND LANGUAGE

- Carlos Busso, *University of Texas, Dallas, USA*
- Julien Epps, *University New South Wales, Australia*
- Chi-Chun (Jeremy) Lee, *National Tsing Hua University, Taiwan, Province of China*
- Elizabeth Shriberg, *SRI International, USA*
- Joseph Tepperman, *Sensory, Inc., USA*

Area 4. SPEAKER AND LANGUAGE IDENTIFICATION

- Jean-François Bonastre, *University Avignon, France*
- Najim Dehak, *Massachusetts Institute of Technology, USA*
- Haizhou Li, *Institute for Infocomm Research, Singapore*
- Ming Li, *Sun Yat-sen University and Carnegie Mellon University, China and USA*

Area 5. ANALYSIS OF SPEECH AND AUDIO SIGNALS

- Prasanta Ghosh, *Indian Institute of Science, India*
- Mark Hasegawa-Johnson, *University of Illinois, Urbana–Champaign, USA*
- Denis Jouvet, *INRIA Nancy, France*
- Koichi Shinoda, *Tokyo Institute of Technology, Japan*
- Malcom Slaney, *Google, USA*

Area 6. SPEECH CODING AND ENHANCEMENT

- Dimitrios Dimitriadis, *IBM, USA*
- Horacio Franco, *SRI International, USA*
- John Hansen, *University of Texas, Dallas, USA*
- Hynek Hermansky, *Johns Hopkins University, USA*
- Bastiaan Kleijn, *Victoria University of Wellington, New Zealand*

Area 7. SPEECH SYNTHESIS AND SPOKEN LANGUAGE GENERATION

- Alan Black, *Carnegie Mellon University, USA*
- Hema Murthy, *Indian Institute of Technology, Madras, India*
- Ingmar Steiner, *Saarland University, Germany*
- Keiichi Tokuda, *Nagoya Institute of Technology, Japan*
- Heiga Zen, *Google, UK*

Area 8. SPEECH RECOGNITION – SIGNAL PROCESSING, ACOUSTIC MODELING, ROBUSTNESS, AND ADAPTATION

- Ozlem Kalinli, *SONY, USA*
- Brian Kingsbury, *IBM, USA*
- Bhuvana Ramabhadran, *IBM, USA*
- Rick Rose, *Google, USA*
- Umesh Srinivasan, *Indian Institute of Technology, Madras, India*

Area 9. SPEECH RECOGNITION – ARCHITECTURE, SEARCH, AND LINGUISTIC COMPONENTS

- Murat Akbacak, *Microsoft, USA*
- Lori Lamel, *LIMSI, France*
- Mari Ostendorf, *University of Washington, USA*
- Murat Saraclar, *Bogazici University, Turkey*
- Geoff Zweig, *Microsoft, USA*

Area 10. SPEECH RECOGNITION – TECHNOLOGIES AND SYSTEMS FOR NEW APPLICATIONS

- Pascale Fung, *Hong Kong University of Science and Technology, China*
- Tatsuya Kawahara, *Kyoto University Sakyo-ku, Japan*
- Douglas O'Shaughnessy, *Université du Québec, Canada*
- Brian Strope, *Google, USA*

Area 11. SPOKEN LANGUAGE – DIALOG, SUMMARIZATION, UNDERSTANDING

- Rolf Carlson, *KTH Stockholm, Sweden*
- Joakim Gustafson, *KTH Stockholm, Sweden*
- Helen Meng, *Chinese University Hong Kong, China*
- Alexandros Potamianos, *National Technical University, Athens, Greece*
- David Traum, *University of Southern California, Institute for Creative Technologies, USA*

Area 12. SPOKEN LANGUAGE PROCESSING – TRANSLATION, INFORMATION RETRIEVAL, AND RESOURCES

- Roland Kuhn, *National Research Council, Canada*
- Bin Ma, *Institute for Infocomm Research A*STAR, Singapore*
- Hermann Ney, *RWTH Aachen University, Germany*
- Tanja Schultz, *Universität Bremen, Germany*
- Isabel Trancosco, *INESC-ID, Portugal*

Area 13. SPEECH AND SPOKEN-LANGUAGE BASED MULTIMODAL PROCESSING AND SYSTEMS

- Gérard Bailly, *CNRS/Grenoble University, France*
- Tanaya Guha, *Indian Institute of Technology, India*
- Angeliki Metallinou, *Amazon, USA*
- Florian Metze, *Carnegie Mellon University, USA*

Area 14. SPECIAL SESSIONS

- Abeer Alwan, *University of California, Los Angeles, USA*
- Dilek Hakkani-Tur, *MSR, USA*

Area 15. CROSS-CUTTING AREA CHAIRS

- Kartik Audhkhasi, *IBM, United States*
- Vikram Ramanarayanan, *Educational Testing Service, R&D Connections, USA*

SCIENTIFIC REVIEW COMMITTEE

Alberto Abad
Ossama Abdel-Hamid
Nassima Abdelli-Beruh
Alex Acero
Andre Adami
Gilles Adda
Martine Adda-Decker
Jordi Adell
Mohamed Afify
Byron Ahn
Manu Airaksinen
Masato Akagi
Masami Akamine
Murat Akbacak
Yuya Akita
Jahangir Alam
Felix Albu
Jan Alexandersson
Paavo Alku
Alexandre Allauzen
Fil Alleva
Jens Allwood
Jesús B. Alonso
Tanel Alumäe
Eliathamby Ambikairajah
Angélique Amelot
Noam Amir
Tim Anderson
Bistra Andreeva
Walter Andrews
Jorn Anemuller
Pongtep Angkititrakul
Xavier Anguera
Jan-Niklas Antons
Gopala Krishna Anumanchipalli
Takayuki Arai
Masahiro Araki
Shoko Araki
Julian David Arias Londoño
Ebru Arisoy
Sebastian Arndt
Hagai Aronowitz
Harish Arsikere
Levent Arslan
Peter Assmann
Ramón Astudillo
Bishnu Atal
Vincent Aubanel
Kartik Audhkhasi
Cinzia Avesani
Michiel Bacchiani
Tom Bäckström
Pierre Badin
Paolo Baggia
Gerard Bailly
Raimo Bakis
Plinio Barbosa
Nelly Barbot
Jon Barker
Dante Barone
Claude Barras

Vincent Barriac
Nikoletta Basiou
Fernando Batista
Anton Batliner
Stefan Baumann
Timo Baumann
Ali Orkan Bayer
Frederic Bechet
Mary Beckman
Steve Beet
Homayoon Beigi
Peter Bell
Jerome Bellegarda
Patrice Bellot
Ashwin Bellur
Mohamed Ben Jannet
Atef Ben Youssef
Jose Miguel Benedi
Carmen Benítez Ortúzar
Štefan Beňuš
Mohamed Faouzi Benzeghiba
Nicole Beringer
Kay Berkling
Jared Bernstein
Frederic Berthommier
Nicola Bertoldi
Laurent Besacier
Jonas Beskow
Steven Bird
Peter Birkholz
Jason Bishop
Judith Bishop
Maria Paola Bissiri
Alan W Black
Jose Luis Blanco Murillo
Tobias Bocklet
Louis-Jean Boe
Antonio Bonafonte
Jean-François Bonastre
Zinny Bond
Nandini Bondale
Daniel Bone
Francesca Bonin
Anne Bonneau
Hynek Boril
Tomáš Boňl
Philippe Boula de Mareüil
Gilles Boulianne
Herve Bourlard
Rachel Bouserhal
Pierre-Michel Bousquet
Suzanne Boyce
Michael Brady
Daniela Braga
Thomas Brand
Angelika Braun
Bettina Braun
Hervé Bredin
Andrew Breen
John Bridle
Mirjam Broersma
Alejna Brugos

Niko Brummer
Alessio Brutti
Murtaza Bulut
H Timothy Bunnell
Harry Bunt
Susanne Burger
Lukas Burget
Felix Burkhardt
Carlos Busso
Joao Cabral
Peter Cahill
Luis Caldas de Oliveira
Zoraida Callejas
Jose Ramon Calvo de Lara
Joseph Campbell
Nick Campbell
William Campbell
Valentín Cardeñoso-Payo
Patrick Cardinal
Christopher Carignan
Rolf Carlson
Francisco Casacuberta
Diamantino Caseiro
Diego Castan
Maria Jose Castro-Bleda
Lawrence Cavedon
Christophe Cerisara
Jan řernocký
Loredana Cerrato
Rupayan Chakraborty
Senthilkumar Chandramohan
Delphine Charlet
Ciprian Chelba
Chandra Sekhar Chellu
Berlin Chen
Fei Chen
Guoguo Chen
I-Fan Chen
Kuan-Yu Chen
Ling-Hui Chen
Liping Chen
Nancy Chen
Yun-Nung Chen
Jian Cheng
You-Chi Cheng
Rathinavelu Chengalvarayan
Mohamed Chetouani
Jonathan Chevelu
Jen-Tzung Chien
K.K. Chin
Eng Siong Chng
Taehong Cho
Gérard Chollet
Khalid Choukri
Herbert Clark
Robert Clark
Luísa Coheur
Jennifer Cole
Martin Cooke
Robin Cooper
Ricardo Cordoba
Martin Corley

Piero Cosi
Irina Cotanis
Alex Cristia
Olivier Crouzet
Jia Cui
Xiaodong Cui
Sandro Cumani
Fred Cummins
Nicholas Cummins
Francesco Cutugno
Deborah Dahl
Christophe d'Alessandro
Geraldine Damnati
Jianwu Dang
Falavigna Daniele
Giacobello Daniele
Khalid Daoudi
Tran-Huy Dat
Marelie Davel
Chris Davis
Carme de la Mota
Jose Mario De Martino
Renato de Mori
Bert de Vries
David Dean
Salil Deena
Gilles Degottex
Najim Dehak
Michael Deisher
Phillip DeLeon
Héctor Delgado
Arnaud Delhay
Veronique Delvaux
Grazyna Demenko
Kris Demuynck
Yasuharu Den
Bruce Denby
Huiqun Deng
Li Deng
Anoop Deoras
Olivier Deroo
Om Deshmukh
Laurence Devillers
Jacob Devlin
Luis Fernando D'Haro
Maria-Gabriella Di Benedetto
Christian DiCanio
Mireia Diez
Vassilis Digalakis
Dimitrios Dimitriadis
Mariapaola D'Imperio
Hongwei Ding
Sascha Disch
Paul Dixon
Simon Dobnik
Gerry Docherty
Laura Docio-Fernandez
Rama Sanand Doddipatla
Marion Dohen
Hans Dolfing
Minghui Dong
Carlo Drioli

COMMITTEES

SCIENTIFIC REVIEW COMMITTEE

Jasha Droppo
Thomas Drugman
Andrzej Drygajlo
Jacques Duchateau
Richard Dufour
Sophie Dufour
Stéphane Dupont
Thierry Dutoit
Camille Dutrey
Jens Edlund
Robert Eklund
Mounya Elhilali
Benjamin Elie
Ahmad Emami
Olov Engwall
Julien Epps
Hakan Erdogan
Donna Erickson
Anders Eriksson
Daniel Erro
Engin Erzin
David Escudero
Maxine Eskenazi
Christina Esposito
Carol Espy-Wilson
Yannick Estève
Georgios Evangelopoulos
Keelan Evanini
Nicholas Evans
Mauro Falcone
Isabel Falé
Tiago Falk
Xing Fan
Jérôme Farinas
Kevin Farrell
Mireia Farrús
Friedrich Faubel
Benoit Fauve
Benoit Favre
Marcello Federico
Tibor Fegyó
Klaus Fellbaum
Sidney Fels
Junlan Feng
Raquel Fernandez
Emmanuel Ferragne
Isabelle Ferrané
Marc Ferras
Javier Ferreiros
Carlos Ferrer
Luciana Ferrer
Lionel Feugère
Markus Fiedler
Tim Fingscheidt
Volker Fischer
Janet Fletcher
José A. R. Fonollosa
Eric Fosler-Lussier
George Foster
Horacio Franco
Pasi Fränti
Corinne Fredouille
Gerald Friedland
Daniel Friedrichs
Guillaume Fuchs
Robert Fuchs
Susanne Fuchs
Mark Fuhs
Masakiyo Fujimoto
Takashi Fukuda
Takahiro Fukumori
Pascale Fung
Sadaoki Furui
Mark Gales
Olivier Galibert
Ascension Gallardo-Antolin
Sriram Ganapathy
Suryakanth V Gangashetty
Sharon Gannot
Fernando García
Maria Luisa Garcia Lecumberri
Daniel Garcia-Romero
Philip N. Garner
Maeva Garnier
Harinath Garudadri
Roberto Gemello
Jort F. Gemmeke
Panayiotis Georgiou
Branislav Gerazov
Houman Ghaemmaghami
Prasanta Ghosh
Sayan Ghosh
Arnab Ghoshal
Dafydd Gibbon
Jonathan Ginzburg
Laurent Girin
Jim Glass
Ondrej Glembek
Herve Glotin
Juan Ignacio Godino llorente
Roland Goecke
Vaibhava Goel
Stefan Goetze
Louis Goldstein

SCIENTIFIC REVIEW COMMITTEE

Pavel Golik
Christian Gollan
Angel Gomez
Pedro Gómez-Vilda
Yifan Gong
Jose A. Gonzalez
Jesús González-Rubio
Jeff Good
Ananthakrishnan Gopal
Allen Gorin
Kyle Gorman
Mária Gósy
Yoshi Gotoh
Martijn Goudbeek
Philippe Gournay
Evandro Gouvea
Vince Gracco
Martin Graciarena
Calbert Graham
Björn Granström
Agustin Gravano
Guillaume Gravier
David Grayden
Phil Green
Steven Greenberg
Adele Gregory
Frantisek Grezl
Gintare Grigonyte
David Griol
Alex Gruenstein
Wentao Gu
Jon Gudnason
Tanaya Guha
Rodrigo Guido
Vishwa Gupta
Joakim Gustafson
Tino Haderlein
Reinhold Haeb-Umbach
Christina Hagedorn
Seongjun Hahm
Stefan Hahn
Thomas Hahn
Akmal Haidar
Thomas Hain
Eva Hajicova
Dilek Hakkani-Tur
Simon Hammond
Kyu Han
John H.L. Hansen
Amir Hossein Harati Nejad Torbati
Philip Harding
Jonathan Harrington
Naomi Harte
William Hartmann
Madina Hasan
Taufiq Hasan
Mark Hasegawa-Johnson
Kei Hashimoto
Helen Hastie
Ville Hautamaki
Rachel Hayes-Harb

T. J. Hazen
Martin Heckmann
Rajesh Hegde
Paul Heisterkamp
Mattias Heldner
John Henderson
Nathalie Henrich
Gustav Eje Henter
Caroline Henton
Christian Herbst
Christian Herff
Hynek Hermansky
Inma Hernaez
Luis Hernandez-Gomez
Gabriel Hernandez-Sierra
Javier Hernando
John Hershey
Ingo Hertrich
Dirk Heylen
Ryuichiro Higashinaka
Ivan Himawan
Florian Hintz
Yusuke Hioka
Keikichi Hirose
Hans-Guenter Hirsch
Julia Hirschberg
Daniel Hirst
Michel Hoen
Volker Hohmann
Wendy Holmes
Qingyang Hong
Florian Hönig
Ron Hoory
Chiori Hori
Takaaki Hori
Julian Hough
David House
Ian Howard
Chien-Lin Huang
Dongyan Huang
Po-Sen Huang
Qiang Huang
Rongqing Huang
Mark Huckvale
Thomas Hueber
David Huggins Daines
Qiang Huo
Lluís-F. Hurtado
Ahmed Hussen Abdelaziz
Brian Hutchinson
Mei-Yuh Hwang
Osamu Ichikawa
Yusuke Ijima
Shajith Ikbal
Irina Illina
Satoshi Imaizumi
David Imseng
Toshio Irino
Markus Iseli
Carlos Ishi
Takeshi Ishihara
Masato Ishizaki

Khalil Iskarous
Ken-ichi Iso
Akinori Ito
Alexei V. Ivanov
Koji Iwano
Rukmini Iyer
Toshiko Jaakkola
Bassam Jabaian
Adam Janin
David Janiszek
Stefanie Jannedy
Esther Janse
Aren Jansen
Jarek Krajewski Jarek Krajewski
Javier Perez Javier Perez
Peter Jax
Jesper Jensen
Alexandra Jesse
Luis M.T. Jesus
Ming Ji
Hui Jiang
Minho Jin
Qin Jin
Cheolwoo Jo
Michael Johnson
Emma Jokinen
Kristiina Jokinen
Oliver Jokisch
Arne Jonsson
Szu-Chen Stan Jou
Denis Jouvet
Tim Juergens
Sun-Ah Jun
Dan Jurafsky
Preethi Jyothi
Tokihiko Kaburagi
Abdellah Kacha
Zdravko Kacic
Takehiko Kagoshima
Juliette Kahn
Alexander Kain
Kaustubh Kalgaonkar
Ozlem Kalinli
Yutaka Kamamoto
Hirokazu Kameoka
Naoyuki Kanda
John Kane
Hong-Goo Kang
Stephan Kanthak
Arthur Kantor
Martin Karafiat
Alexey Karpov
Hiroaki Kato
Athanasios Katsamanis
Hideki Kawahara
Tatsuya Kawahara
Shinichi Kawamoto
Heysem Kaya
Patricia Keating
Christian Kell
Finnian Kelly

Casey Kennington
Patrick Kenny
Bilal Khaliq
Elie Khoury
Genichiro Kikui
Doh-Suk Kim
DoYeong Kim
Hoi Rin Kim
Hong Kook Kim
Hyung Soon Kim
Jeesun Kim
Jong-mi Kim
Kee-Ho Kim
Nam Soo Kim
Samuel Kim
Seokhwan Kim
Wooil Kim
Simon King
Brian Kingsbury
Tomi Kinnunen
Keisuke Kinoshita
Irina Kipyatkova
Norihide Kitaoka
Esther Klabbers
Dietrich Klakow
Felicitas Kleber
Bastiaan Kleijn
Thomas Kleinbauer
Katarzyna Klessa
Neil Kleynhans
Kate Knill
Hanseok Ko
Takao Kobayashi
Alexei Kochetov
Marcel Kockmann
Sri Rama Murty Kodukula
Tina Kohler
Daniel Kohlsdorf
Jachym Kolar
Thomas Kollar
Kazunori Komatani
Mariko Kondo
Myoung-Wan Koo
Shashidhar G Koolagudi
Tomoki Koriyama
Takafumi Koshinaka
Maria Koutsogiannaki
Ivan Kraljevski
Jelena Krivokapic
Bernd Kroeger
Christian Kroos
Gernot Kubin
Oleg Kudashev
Roland Kuhn
Kshitiz Kumar
Jimmy Kunzmann
Grace Kuo
Gakuto Kurata
Mikko Kurimo
Frank Kurth
Chul Hong Kwon
Oh-Wook Kwon

SCIENTIFIC REVIEW COMMITTEE

Rafael Laboissière
Francisco Lacerda
Pietro Laface
Catherine Lai
Unto K. Laine
Lori Lamel
Pierre Lanchantin
Ian Lane
Brian Langner
Itshak Lapidot
Yves Laprie
Anthony Larcher
Romain Laroche
Martha Larson
Eva Lasarcyk
Kornel Laskowski
Lukas Latacz
Javier Latorre
Aaron Lawson
Phu Le
Sébastien Le Maguer
Jonathan Le Roux
Margaret Lech
Jeremie Lecomte
Gwénolé Lecorvé
Benjamin Lecouteux
Akinobu Lee
Chi-Chun Lee
Chin-Hui Lee
Hung-yi Lee
Jaewon Lee
Kong Aik Lee
Lin-shan Lee
Siu-Wa Lee
Sungbok Lee
Sungjin Lee
Tan Lee
Adrian Leemann
Fabrice Lefevre
Milan Legát
Yun Lei
Kevin Lenzo
Cheung-Chi Leung
Gary Leung
Michael Levit
Rivka Levitan
Aijun Li
Bo Li
Feipeng Li
Haizhou Li
Jinyu Li
Junfeng Li
Ming Li
Qi (Peter) Li
Hank Liao
Robin Lickley
Jean-Sylvain Lienard
Amaro Lima
Carlos Lima
Jen-Chun Lin
Georges Linares
Børge Lindberg
Anders Lindström
Zhen-Hua Ling
Chaojun Liu
Gang Liu
Jia Liu
Pengfei Liu
Wenju Liu
Xunying Liu
Eduardo Lleida Solano
Joaquim Llisterri
Deborah Loakes
Anders Lofqvist
Damien Lolive
Yanhua Long
José Lopes
Ramon Lopez-Cozar
Paula Lopez-Otero
Teresa Lopez-Soto
Torrey Loucks
Anastassia Loukina
Heng Lu
Liang Lu
Xugang Lu
Jorge Lucero
Steven Lulich
Susann Luperfoy
Jordi Luque
Athanasios Lykartsis
Bin Ma
Changxue Ma
Jeff Ma
Ning Ma
Roland Maas
Ewen MacDonald
Javier Macias-Guarasa
Ian Maddieson
Srikanth Madikeri
Hari Krishna Maganti
Mathew Magimai Doss
Shakuntala Mahanta
Ranniery Maia
Brian Mak

SCIENTIFIC REVIEW COMMITTEE

Manwai Mak
Zofia Malisz
Sri Harish Mallidi
Nicolas Malyska
Michael Mandel
Claudia Manfredi
Lidia Mangu
Kazunori Mano
Krzysztof Marasek
Etienne Marcheret
Erik Marchi
Stefania Marin
Soroosh Mariooryad
Goran Markovic
David Marks
Jean-Pierre Martens
Rainer Martin
David Martínez González
Carlos Martínez-Hinarejos
David Martins de Matos
Ricard Marxer
Dom Massaro
Hinako Masuda
Takashi Masuko
Ana Isabel Mata
Marco Matassoni
Pavel Matejka
Driss Matrouf
Tomoko Matsui
Yuri Matveev
Julie Mauclair
Ludo Max
Alan McCree
Erik McDermott
Mitchell McLaren
Michael McTear
Daryush Mehta
Sylvain Meignier
Hugo Meinedo
Alexsandro Meireles
Lucie Ménard
Helen Meng
Angeliki Metallinou
Marie Meteer
Florian Metze
Fanny Meunier
Bernd T. Meyer
Yohann Meynadier
Lei Miao
Yajie Miao
Antonio Miguel
Ben Milner
Yasuhiro Minami
Nobuaki Minematsu
Majid Mirbagheri
Ananya Misra
Vikramjit Mitra
Hansjörg Mixdorff
Bernd Möbius
Abdelrahman Mohamed
Saif Mohammad
Parham Mokhtari
Sebastian Möller
Helena Moniz
Juan M Montero
Seung-Jae Moon
Roger Moore
Sylvia Moosmueller
Juan A. Morales-Cordovilla
Asuncion Moreno
Masanori Morise
Alessandro Moschitti
Petr Motlicek
Athanasios Mouchtaris
Emily Mower Provost
Pejman Mowlaee
Karen Mulak
Ludek Muller
Philippe Muller
Markus Müller
Benjamin Munson
Hema Murthy
Narendra N P
Climent Nadeu
Venki Nagesha
Tofigh Naghibi
Devang Naik
Maryam Najafian
Kazuhiro Nakadai
Seiichi Nakagawa
Satoshi Nakamura
Tomohiro Nakatani
Arun Narayanan
Shrikanth Narayanan
Marina Nastasenko
Eva Navas
Géza Németh
Ani Nenkova
Friedrich Neubarth
Graham Neubig
Hermann Ney
Raymond W. M. Ng
Noel Nguyen
Patrick Nguyen
Trung Hieu Nguyen
Chongjia Ni
Mauro Nicolao
Oliver Niebuhr
Jan Niehues
Kuniko Nielsen
Kristina Nilsson Björkenstam
Masafumi Nishida
Masayuki Nishiguchi
Masafumi Nishimura
Takanobu Nishiura
Nobuyuki Nishizawa
Mohamed Noamany
Elmar Noeth
Jan Nouza
Mirek Novak
Sergey Novoselov
Markus Nussbaum-Thom
Nicolas Obin
Yasunari Obuchi
Ferda Ofli
Atsunori Ogawa
Tetsuji Ogawa
John Ohala
Yamato Ohtani
Hiroshi G. Okuno
Mohamed Omar
Maurizio Omologo
Nobutaka Ono
Ilya Oparin
Roeland Ordelman
Antonio Origlia
Juan Rafael Orozco-Arroyave
Rosemary Orr
Alfonso Ortega
Douglas O'Shaughnessy
Mari Ostendorf
Slim Ouni
Keiichiro Oura
Mukund Padmanabhan
Vincent Pagel
Yi-Cheng Pan
Yue Pan
Sankaran Panchapagesan
Ashish Panda
Prem C. Pandey
Aasish Pappu
José Pardo
Naveen Parihar
Alok Parlikar
Vijay Parsa
SHK (Hari) Parthasarathi
Sarangarajan Parthasarathy
Rajesh Patel
Hemant Patil
Kailash Patil
Sanjay Patil
Matthias Paulik
Vijayaditya Peddinti
Robert Peharz
Antonio M. Peinado
Catherine Pelachaud
Carmen Peláez-Moreno
Jason Pelecanos
Thomas Pellegrini
Bryan Pellom
Mikel Penagarikano
Fuchun Peng
Fernando Perdigão
José L. Pérez-Córdoba
Franz Pernkopf
Pascal Perrier
Patrick Perrot
Olivier Perrotin
Sandra Peters
Caterina Petrone
Beat Pfister
Michael Picheny
Olivier Pietquin
Julien Pinquier
John Pitrelli
Paul Piwek
Ferran Pla
Christian Plahl
Oldrich Plchot
Jouni Pohjalainen
Christopher Poletto
Joseph Polifroni
Tim Polzehl
François Portet
Alexandros Potamianos
Gerasimos Potamianos
Marianne Pouplier
Dan Povey
Rohit Prabhavalkar
Kishore Prahallad
S R Mahadeva Prasanna
Kristin Precoda
Simon Preuß
Patti Price
Ryan Price
Michael Proctor
Michael Pucher
Manfred Pützer
Yanmin Qian
Yao Qian
Thomas Quatieri
Carl Quillen
Tuomo Raitio
Padmanabhan Rajan
Nitendra Rajput
Bhuvana Ramabhadran
Vikram Ramanarayanan
V Ramasubramanian
Daniel Ramos
Vivek Kumar Rangarajan Sridhar
K Sreenivasa Rao
Kanishka Rao
Preeti Rao
Wei Rao
Ramya Rasipuram
Shakti Rath
Anabela Rato
Andreia Rauber
Christian Raymond
Manny Rayner
Melissa A. Redford
Henning Reetz
Mario Refice
Uwe Reichel
Patrick Reidy
Norbert Reithinger
Steve Renals
Steven Rennie
Fernando Gil Vianna Resende Junior
Douglas Reynolds
Dayana Ribas Gonzalez
Carlos Ribeiro
Ricardo Ribeiro
Giuseppe Riccardi
Gaël Richard
Fred Richardson

SCIENTIFIC REVIEW COMMITTEE

Korin Richmond
Korbinian Riedhammer
Luca Rigazio
Michael Riley
Fabien Ringeval
Christian Ritz
Tony Robinson
Amelie Rochet-Capellan
Eduardo Rodriguez Banga
Luis Javier Rodriguez-Fuentes
Rick Rose
Andrew Rosenberg
Solange Rossato
Sophie Rosset
Antti-Veikko Rosti
Jean-Luc Rouas
Mickael Rouvier
Viktor Rozgic
Alexander Rudnicky
Vesa Ruoppila
Martin Russell
David Rybach
Rahim Saeidi
Saeid Safavi
Yoshinori Sagisaka
Lakshmi Saheer
Tara Sainath
Daisuke Saito
Sakriani Sakti
Elliot Saltzman
Nele Salveste
Giampiero Salvi
K Samudravijaya
Rubén San Segundo Hernández
Victoria Sanchez
Joan Andreu Sánchez
Emilio Sanchis
Germán Sanchis-Trilles
Bonny Sands
Abhijeet Sangwan
Joao Felipe Santos
George Saon
Murat Saraclar
Ibon Saratxaga
Ruhi Sarikaya
Priyankoo Sarmah
Milton Sarria-Paja
Akira Sasou
Antonio Satue Vilar
Michelina Savino
Oscar Saz
Thomas Schaaf
Odette Scharenborg
Nicolas Scheffer
Stefan Scherer
David Schlangen
Ralf Schlüter
Joerg Schmalenstroeer
Sven Schmeier
Jean Schoentgen
Karl Schuchmann
Björn Schuller
Tanja Schultz
Mike Schuster
Antje Schweitzer
Chandra Sekhar Seelamantula
Encarna Segarra
Frank Seide
Michael Seltzer
Deep Sen
Christine Senac
Gregory Senay
Cheol Jae Seong
Willy Serniclaes
Guruprasad Seshadri
Vidhyasaharan Sethu
Abhinav Sethy
Turaj Shabestary
Izhak Shafran
Jason Shaw
Slava Shechtman
Sven Shepstone
Yoshinori Shiga
Tetsuya Shimamura
Koichi Shinoda
Takahiro Shinozaki
Sayaka Shiota
Elizabeth Shriberg
Vered Silber-Varod
Jan Silovsky
Kim Silverman
Michel Simard
Juraj Šimko
Konstantin Simonchik
Adrian Simpson
Elliot Singer
Rohit Sinha
Sabato Marco Siniscalchi
Olivier Siohan
Sunayana Sitaram
Man-hung Siu
Sunil Sivadas
Gabriel Skantze
Jan Skoglund

COMMITTEES

SCIENTIFIC REVIEW COMMITTEE

Malcolm Slaney
Raymond Slyh
Kamel Smaili
Paris Smaragdis
Rudolph Sock
Maria Josep Solé
Rubén Solera-Ureña
Hagen Soltau
Mitchell Sommers
Frank Soong
Victor Sorokin
Richard Sproat
Thippur Sreenivas
Jacek Stachurski
Themos Stafylakis
Ian Stavness
Stefan Steidl
Ingmar Steiner
Amanda Stent
Evgeny Stepanov
Richard Stern
Andreas Stolcke
Brad Story
Svetlana Stoyanchev
Stephanie Strassel
Helmer Strik
Sofia Strömbergsson
Brian Strope
Sebastian Stüker
Yannis Stylianou
Yi Su Su
Aswin Shanmugam Subramanian
David Suendermann-Oeft
Jun-Won Suh
Ming Sun
Xie Sun
Harshavardhan Sundar
Shiva Sundaram
Masayuki Suzuki
Piergiorgio Svaizer
Torbjørn Svendsen
Marc Swerts
Pawel Swietojanski
Ann Syrdal
Michael Syskind Pedersen
Eva Szekely
Igor Szoke
Marija Tabain
Martha Yifiru Tachbelie
Yuuki Tachioka
Toru Takahashi
Shinji Takaki
Tetsuya Takiguchi
Yik-Cheung Tam
Fabio Tamburini
Zheng-Hua Tan
Hiroki Tanaka
Kazuyo Tanaka
Kevin Tang
Yun Tang
Jianhua Tao

Naohiro Tawara
António Teixeira
Carlos Teixeira
João Paulo Teixeira
Dominic Telaar
Louis ten Bosch
Joseph Tepperman
Fabio Tesser
Veena Thenkanidiyoor
Barry-John Theobald
Samuel Thomas
William Thorpe
Jill Thorson
Sam Tilsen
Michael Tjalve
Tomoki Toda
Massimiliano Todisco
Roberto Togneri
Keiichi Tokuda
Shinichi Tokuma
Kanako Tomaru
Laura Tomokiyo
Rong Tong
Pedro Torres-Carrasquillo
László Tóth
Asterios Toutios
Dung Tran
Isabel Trancoso
David Traum
Juergen Trouvain
Khiet Truong
Stavros Tsakalidis
Yu Tsao
Chiu-yu Tseng
Shu-Chuan Tseng
Andreas Tsiartas
Kimiko Tsukada
Gokhan Tur
Oytun Turk
Markku Turunen
Zoltán Tüske
Michael Tyler
Nicola Ueffing
Srinivasan Umesh
Masashi Unoki
Maria Uther
Michel Vacher
Martti Vainio
Claudio Vair
Cassia Valentini-Botinhao
Francisco J Valverde-Albacete
Dirk Van Compernolle
Rogier van Dalen
Henk van den Heuvel
Laurens van der Werff
Hugo Van hamme
Charl van Heerden
David van Leeuwen
Daniel Van Niekerk
Jan van Santen
Maarten Van Segbroeck
Rob van Son

Amparo Varona
Adriana Vasilache
Ioana Vasilescu
Eric Vatikiotis-Bateson
Nanette Veilleux
Dimitra Vergyri
Pieter Vermeulen
Klara Vicsi
Marina Vigario
Coriandre Vilain
Jesus Villalba
Fernando Villavicencio
Emmanuel Vincent
Ravichander Vipperla
Tuomas Virtanen
Michael Vitevitch
Carlos Vivaracho-Pascual
Adam Vogel
Carl Vogel
Stephen Voran
Ngoc Thang Vu
Anil Kumar Vuppala
Petra Wagner
Marilyn Walker
Michael Walsh
Patrick Wambacq
Guangsen Wang
Hsiao-Chuan Wang
Hsin-Min Wang
Lijuan Wang
William Yang Wang
Xinhao Wang
Yongqiang Wang
Nigel Ward
Wayne Ward
Paul Warren
Shinji Watanabe
Catherine Watson
Jianguo Wei
Benjamin Weiss
Pauline Welby
Christian Wellekens
Stanley Wenndt
Stefan Werner
Mirjam Wester
Lorin Wilde
Daniel Willett
Jason D Williams
Mats Wirén
Guillaume Wisniewski
Silke Witt-Ehsani
Marcin Wlodarczak
Maria Wolters
Phil Woodland
Chuck Wooters
Johan Wouters
Chung-Hsien Wu
Zhiyong Wu
Zhizheng Wu
Chai Wutiwiwatchai
Rui Xia
Bing Xiang

Xiong Xiao
Shaofei Xue
Junichi Yamagishi
Yoichi Yamashita
Umit Yapanel
Mahsa Yarmohammadi
Keiichi Yasu
Guoli Ye
Bayya Yegnanarayana
Ching-Feng Yeh
Serdar Yildirim
Emre Yilmaz
Nestor Becerra Yoma
Chang Yoo
Dongsuk Yook
Su-Youn Yoon
Koichiro Yoshino
Steve Young
Chengzhu Yu
Dong Yu
Dong Yu
Kai Yu
Jiahong Yuan
Young-Sun Yun
François Yvon
Stephen Zahorian
Milos Zelezny
Margaret Zellers
Heiga Zen
Andrej Zgank
Chi (Leo) Zhang
Pengyuan Zhang
Wei Zhang
Yu Zhang
Zhengchen Zhang
Rui Zhao
Yunxin Zhao
Thomas Fang Zheng
Xinhui Zhou
Xiaodan Zhu
Xiaodan Zhuang
Ali Ziaei
Wolfram Ziegler
Frank Zimmerer
Imed Zitouni
Udo Zoelzer
Yuexian Zou
Enrico Zovato
Geoffrey Zweig
Marzena Zygis

ISCA Medal for Scientific Achievement

The ISCA Medal for Scientific Achievement 2016 will be awarded to John Makhoul by the President of ISCA during the opening ceremony.

ISCA Best Student Paper Award

Each year, ISCA awards 3 best student papers at Interspeech based on anonymous reviewing and presentation at the conference. This year, 12 papers are shortlisted for best student paper:

583 *Tanner Sorensen, Asterios Toutios, Louis Goldstein and Shrikanth Narayanan*
Characterizing Vocal Tract Dynamics Across Speakers Using Real-Time MRI
Oral Session 2-3: Articulatory Measurements and Analysis
Friday, 09 September 2016, 14:30–16:30

876 *Bo-Hsiang Tseng, Sheng-Syun Shen, Hung-Yi Lee and Lin-Shan Lee*
Towards Machine Comprehension of Spoken Content: Initial TOEFL Listening Comprehension Test by Machine
Poster Session 7-4: Dialogue Systems and Analysis of Dialogue
Sunday, 11 September 2016, 13:30–15:30

873 *Jia Yu, Xiong Xiao, Lei Xie, Eng Siong Chng and Haizhou Li*
A DNN-HMM Approach to Story Segmentation
Poster Session 4-4: Resources and Annotation of Resources
Saturday, 10 September 2016, 10:00–12:00

419 *Najmeh Sadoughi and Carlos Busso*
Head Motion Generation with Synthetic Speech: A Data Driven Approach
Oral Session 1-2: Special Session: Auditory-Visual Expressive Speech and Gesture in Humans and Machines
Friday, 09 September 2016, 11:00–13:00

607 *Wei Lai, Jiahong Yuan, Ya Li, Xiaoying Xu and Mark Liberman*
The Rhythmic Constraint on Prosodic Boundaries in Mandarin Chinese Based on Corpora of Silent Reading and Speech Perception
Oral Session 1-3: Prosody
Friday, 09 September 2016, 11:00–13:00

565 *Shahin Amiriparian, Jouni Pohjalainen, Erik Marchi, Sergey Pugachevskiy and Björn Schuller*
Is Deception Emotional? An Emotion-Driven Predictive Approach
Oral Session 6-2: Special Session: Interspeech 2016 Computational Paralinguistics Challenge (ComParE): Deception, Sincerity & Native Language
Sunday, 11 September 2016, 10:00–12:00

1292 *Waad Ben Kheder, Driss Matrouf, Ajili Moez and Jean-François Bonastre*
Probabilistic Approach Using Joint Clean and Noisy I-Vectors Modeling for Speaker Recognition
Oral Session 10-5: Speaker Recognition
Monday, 12 September 2016, 13:30–15:30

735 *Lauri Juvela, Hirokazu Kameoka, Manu Airaksinen, Junichi Yamagishi and Paavo Alku*
Majorisation-Minimisation Based Optimisation of the Composite Autoregressive System with Application to Glottal Inverse Filtering
Oral Session 3-6: Co-Inference of Production and Acoustics
Friday, 09 September 2016, 17:00–19:00

586 *Kwang Myung Jeon and Hong Kook Kim*
Local Sparsity Based Online Dictionary Learning for Environment-Adaptive Speech Enhancement with Nonnegative Matrix Factorization
Oral Session 8-5: Speech Enhancement
Sunday, 11 September 2016, 16:00–18:00

342 *Manu Airaksinen, Bajibabu Bollepalli, Lauri Juvela, Zhizheng Wu, Simon King and Paavo Alku*
GlottDNN - A Full-Band Glottal Vocoder for Statistical Parametric Speech Synthesis
Oral Session 7-4: Speech Synthesis Oral I: Neural Networks
Sunday, 11 September 2016, 13:30–15:30

580 *Chunyang Wu, Penny Karanasou, Mark Gales and Khe Chai Sim*
Stimulated Deep Neural Network for Speech Recognition
Oral Session 2-1: New Trends in Neural Networks for Speech Recognition
Friday, 09 September 2016, 14:30–16:30

1358 *Yoni Halpern, Keith Hall, Vlad Schogol, Michael Riley, Brian Roark, Gleb Skobeltsyn and Martin Baeuml*
Contextual Prediction Models for Speech Recognition
Poster Session 6-4: Language Model Adaptation
Sunday, 11 September 2016, 10:00–12:00

Travel Grants

41 ISCA and 20 Interspeech 2016 travel grants have been selected based on the technical quality of the Papers.

The travel award recipients are:

Fahimeh Bahmaninezhad
Waad Ben Kheder
Angel Mario Castro Martinez
Fei Chen
Hongjie Chen
Zhehuai Chen
Yu-An Chung
Nauman Dawalatabad
Yehoshua Dissen
Mortaza Doulaty
Anja Eichenauer
Johannes Fahringer
Arpita Gang
Tian Gao
Aditya Gaonkar
Sahar Ghannay
Yuling Gu
Sri Harsha Dumpala
Saad Irtza
Jeanin Jügler
Mounika Kamsali Veera
Akihiro Kato
Abbas Khosravani
Anna Kruspe
Rohan Kumar Das
Wei Lai
Linchuan Li
Xu Li
Tzu-Hsiang Lin
Mairym Llorens Monteserín
Ciira Maina
Mayuki Matsui
Zhong Meng
Vahid Montazeri
Tasha Nagamine
Shamima Najnin
Catharine Oertel
Divya Prabhakaran
Xiaoke Qi
Gurunath Reddy
Emma Rennie
Dayana Ribas
Jishnu Sadasivan
Suman Samui
Saeed Sarfjoo
Bidisha Sharma
Yiping Song
Kaavya Sriskandaraja
Cornelius Styp von Rekowski
Lifa Sun
Yao Tian
Karthika Vijayan
Lauren Ward
Jochen Weiner
Zhengqi Wen
Feng-Long Xie
Yi Yang
Jia Yu
Yougen Yuan
ShiLiang Zhang
Yimeng Zhuang

SAN FRANCISCO

San Francisco is often called "Everybody's Favorite City," a title earned by its scenic beauty, cultural attractions, diverse communities, and world-class cuisine. Measuring 49 square miles, this very walkable city is dotted with landmarks like the Golden Gate Bridge, cable cars, Alcatraz and the largest Chinatown in the United States. A stroll of the City's streets can lead from Union Square to North Beach to Fisherman's Wharf, with intriguing neighborhoods to explore at every turn. Views of the Pacific Ocean and San Francisco Bay are often laced with fog, creating a romantic mood in this most European of American cities.

A PATCHWORK OF DIVERSE NEIGHBORHOODS:

Fisherman's Wharf: View sea lions, savor fresh seafood and board the ferry to Alcatraz.

North Beach: Home to Italian heritage, cappuccino, cabarets and jazz clubs.

Chinatown: The oldest and among the largest in the U.S., where unique architecture, intriguing alleys and shops abound.

Embarcadero/Financial District: Grab a locally-grown bite and hop a ferry to the east bay.

Union Square: Indulge in luxury shopping and people watching in this designer goods mecca.

SOMA/Yerba Buena: Home to Moscone Center, anchoring a neighborhood where world-class art galleries and museums mingle with sleek nightclubs.

Mission District: Enjoy murals, Latino culture and some of the best weather in the city.

Bayview/Candlestick Point: State park lands, diverse experiences, and one of the largest concentrations of working artists in the U.S.

Castro/Upper Market: Sweeping views grace the "gay capital of the world."

Haight-Ashbury: Well-known residence of Janis Joplin and Jimi Hendrix, home to Alamo Square's "Postcard Row" and the city's tie-dyed roots.

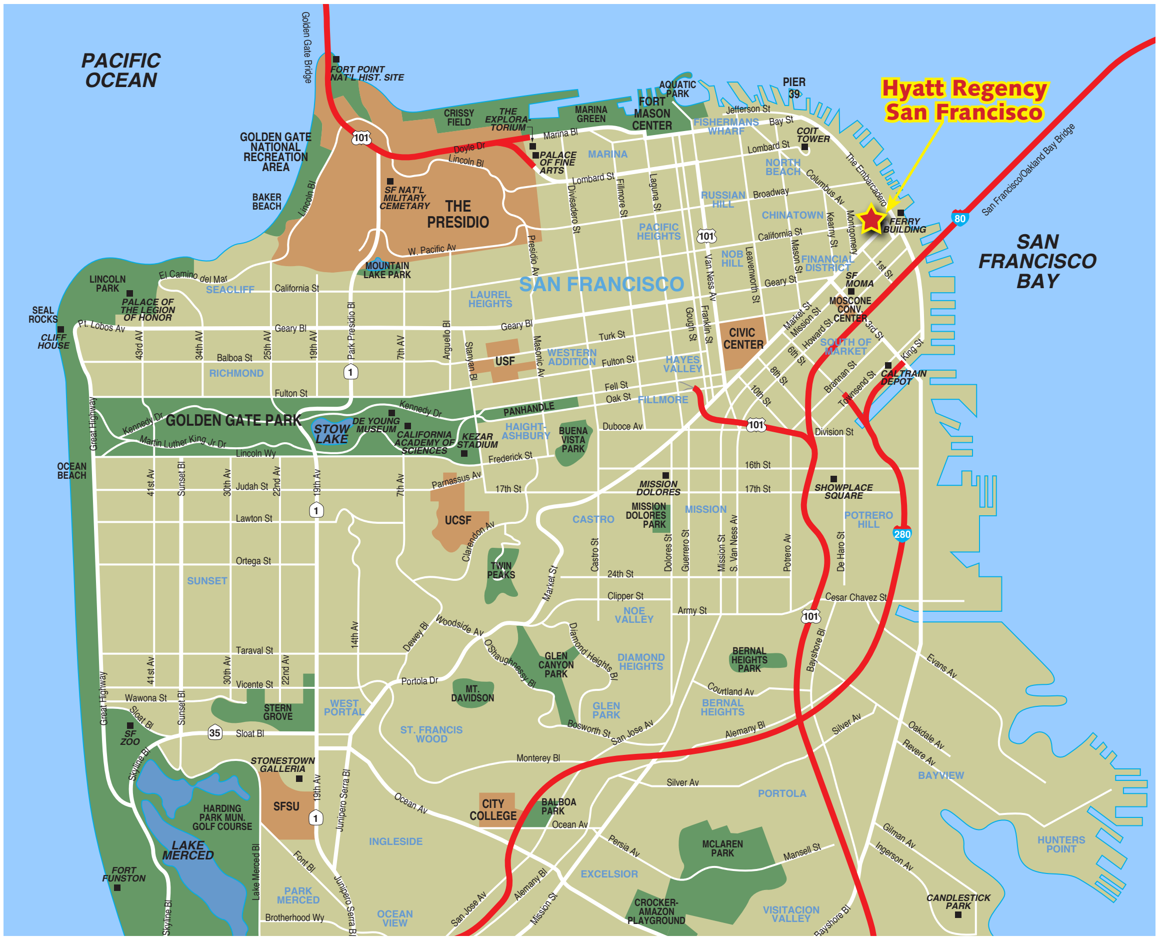
Japantown/Fillmore: The Far East meets the "Harlem of the West."

Civic Center/Hayes Valley: Eclectic shopping, performing arts and Beaux Arts wonder City Hall.

Nob Hill: Features the best view of the bay and the famous cable cars.

Marina/Presidio: An urban National Park with scenic views of the Golden Gate Bridge.

Golden Gate Park/Sunset: Grassy Golden Gate Park gives way to sandy Ocean Beach.



GENERAL INFORMATION

VENUE & ACCOMMODATION

Hyatt Regency San Francisco
5 Embarcadero Center
San Francisco, California, 94111
1-415-788-1234
sanfranciscoregency.hyatt.com

Check-in: 15:00
Check-out: 12:00

MOBILE APP

The Interspeech 2016 mobile app is a native application for tablet and smartphone devices (iPhone and Android).

The mobile app provides easy-to-use interactive capabilities to enhance your experience as an attendee. Features include:

- Agenda: View schedules, explore sessions and find networking events. Create your own personal schedule for easy tracking.
- Update: A quick way to share photos, comments and sessions you're attending.
- Activity Feed: The real-time pulse of the event. See what people are saying, view photos and find sessions and topics.
- Users: See who's here and connect via the app.
- Sponsors/Exhibitors: A complete list of exhibitors/sponsors.

To download the application, visit your app store and search for "Interspeech 2016". Provide the email you used during the online registration process and enter the passcode: IS2016

REGISTRATION

Location: Grand Ballroom Foyer, Street Level, Hyatt Regency San Francisco

Day/Date:	Time:
Thursday, 8 September	08:00 – 19:00
Friday, 9 September	08:00 – 19:30
Saturday, 10 September	08:00 – 18:00
Sunday, 11 September	08:00 – 18:00
Monday, 12 September	08:00 – 17:30

BADGES & MATERIALS

As a registered attendee, you will be issued an Interspeech name badge when you pick up your registration materials onsite. You will be required to display your name badge for admission to all official functions. In the event of a lost badge, you may purchase a replacement badge for a $25.00 fee.

Materials:

- The conference proceedings (a printed abstract book and USB stick)
- A conference bag
- A mobile Android or iOS App
- Access to all official conference sessions
- Coffee breaks Thursday through Monday
- One admission to the Welcome Reception on Friday evening

Electronic PDFs of the abstract book and attendee roster will be available online at: www.interspeech2016.org. Passcode: IS2016

SPEAKER INFORMATION

Speaker Check-in
Location: Regency AB, Street Level, Hyatt Regency San Francisco

Speakers are required to use the computers provided by the conference for their oral presentations. **Personal laptops may not be used.** The conference has arranged for PC laptops equipped with the latest version of operating systems. PowerPoint is the accepted presentation format. Presentations must be submitted on a USB stick at the Speaker Check-In Desk during the designated hours to ensure your talk is accompanied by slides.

Presentation Date	**Speaker Check-In**	
Thursday, 8 September	Wednesday, 7 September	16:00 – 20:00
Friday, 9 September	Thursday, 8 September	07:30 – 17:00
Saturday, 10 September	Friday, 9 September	07:30 – 17:00
Sunday, 11 September	Saturday, 10 September	07:30 – 17:00
Monday, 12 September	Sunday, 11 September	07:30 – 17:00

SPEAKERS TIPS:

- Please arrive to your presentation room 15 minutes prior to the session start time to familiarize yourself with the laser pointer, slide advancer, and stage set-up.
- Please sit towards the front of the room in the session in which you present, a tent card will mark these reserved seats. The Session Chair will introduce your presentation as well as monitor the length of the presentation.
- A laser pointer and slide advancer will be available at the podium for your use.
- All mobile phones must be turned off while you are presenting. Mobile phones on silent will cause feedback with the microphones.

POSTER INFORMATION

Requirements

- Poster presentation space is limited to the dimensions of 92″ x 44″ (234 cm x 112 cm). Material to affix your poster to the poster board will be available in the poster area.
- Presenters are advised to mount their posters during the break before the start of the session, and remove it after the end of the session. Posters have to be removed after the end of the session. Conference organizers will not be responsible for remaining posters.
- At least one of the authors must be present at your poster during the presentation session designated for the poster topic.

INTERSPEECH 2017

Situated interaction

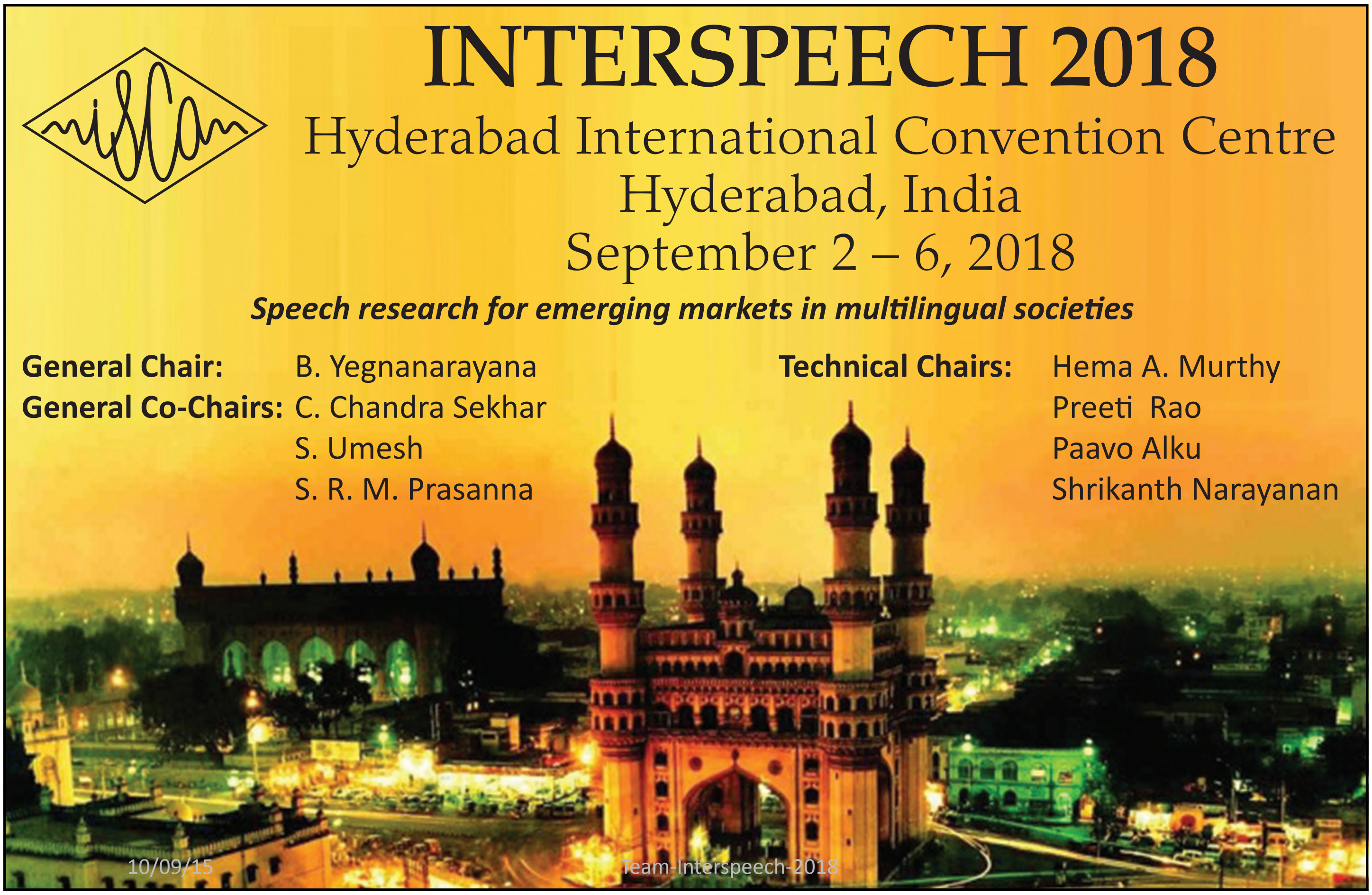
SAVE THE DATE!

August 20-24, 2017 | Stockholm, Sweden
www.interspeech2017.org

General chair:
Francisco Lacerda

Technical program chairs:
Mattias Heldner
Joakim Gustafson
Sofia Strömbergsson

Photograph: Yanan Li/mediabank.visitstockholm.com



INTERSPEECH 2018

Hyderabad International Convention Centre
Hyderabad, India
September 2 – 6, 2018

Speech research for emerging markets in multilingual societies

General Chair: B. Yegnanarayana
General Co-Chairs: C. Chandra Sekhar
S. Umesh
S. R. M. Prasanna

Technical Chairs: Hema A. Murthy
Preeti Rao
Paavo Alku
Shrikanth Narayanan

10/09/15

Team-Interspeech-2018

GENERAL INFORMATION

EMERGENCIES

Please notify the Hyatt Regency San Francisco staff or Conference Solutions for basic medical assistance. The general emergency call number in the United States is 911.

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.

TAXES

San Francisco and the state of California have a sales tax rate of 8.75% on consumer goods and services. Hotel rooms are subject to a 14% lodging tax.

POLICIES

SMOKING POLICY

Smoking is permitted only in designated smoking areas. It is illegal to smoke tobacco products in any public gathering space in California, including parks, restaurants, bars, stores and office buildings.

PHONES

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

LOST & FOUND

Lost & Found will be located at the registration desk in the Grand Ballroom Foyer, Hyatt Regency San Francisco

NAME BADGE

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

DISABILITIES ACT

The Hyatt Regency San Francisco is in compliance with the Americans with Disabilities Act. Special services (e.g., wheelchair-accessible transportation, reserved seating) are available if requested in advance. Should you require assistance onsite, please visit the conference Registration Desk, located in the Grand Ballroom Foyer, Hyatt Regency San Francisco.

SATELLITE DESCRIPTIONS

The following workshops will be organized as satellites of Interspeech 2016:

L1TLT 2016: The 2nd Workshop on Language Teaching, Learning and Technology

Dates: 6-7 September 2016
Location: San Francisco, CA
Website: https://sites.google.com/site/l1teachingandtechnology/

Names and affiliation of organizers:

- Kay Berkling, Baden-Wuerttemberg Cooperative State University, Karlsruhe, Germany
- Keelan Evanini, Educational Testing Service, USA
- David Suendermann-Oeft, Educational Testing Service, USA

Description:
The LTLT workshop intends to join researchers across countries on the topic of language teaching/learning. Papers submitted here do not have to employ any technology yet. We are looking for contributions from users that may not be aware of all the possibilities that the technologies have to offer to solve educational research problems. What these papers bring to the table are problem statements and data collections that the speech and text processing community may in turn not be aware of. Thus we are looking for symbioses between the two disciplines in research about learning/teaching language. It is important for both areas to get to know each other's research questions and potential application for technologies.

WOCCI 2016: The 5th Workshop on Child Computer Interaction

Dates: 6-7 September 2016
Location: San Francisco, CA
Website: http://www.wocci.org/2016/home.html

Names and affiliation of organizers:

- Kay Berkling, Baden-Wuerttemberg Cooperative State University, Karlsruhe, Germany
- Keelan Evanini, Educational Testing Service, USA
- David Suendermann-Oeft, Educational Testing Service, USA

Description:
This workshop aims to join researchers and practitioners from universities and industry working in all aspects of child-machine interaction including computer, robotics and multi-modal interfaces. Children are special both at the acoustic/linguistic level as well as the interaction level. The Workshop provides a unique opportunity for bringing together different research communities from cognitive science, robotics, speech processing, linguistics as well as applied areas such as medical and educational technologies. Various state-of-the-art components can be presented here as key components for the next generation of child-centered computer interaction. Technological advances are increasingly necessary in a world where education and health pose growing challenges to the core wellbeing of our societies. Noticeable examples are remedial treatments for children with or without disabilities and capabilities for providing individualized attention. The Workshop will serve as a venue for presenting recent advancements in core technologies as well as experimental systems and prototypes.

SECNS 2016: The 1st Workshop on Speech Engineering and Computational Neuroscience of Speech

Date: 8 September 2016
Location: San Francisco, CA
Website: https://sites.google.com/site/secns16/

Names and affiliation of organizers:

- Edward F. Chang, University of California, San Francisco, USA
- Gopala Anumanchipalli, University of California, San Francisco, USA

Description:
Several of the goals of speech scientists and neuroscientists working in speech are of mutual relevance and are increasingly converging into each other. The aim of this workshop is to promote exchange of ideas, methods, data and to foster collaborations between researchers working in these fields. This synergy between the neuroscience of speech and computer speech technologies is indispensable for creating the next generation rehabilitative technologies for a range of speech and language disorders and to bring computer based speech technologies closer to human performance in speech recognition, synthesis and understanding.

CHiME 2016: The 4th International Workshop on Speech Processing in Everyday Environments

Dates: 13 September 2016
Location: San Francisco, CA
Website: http://spandh.dcs.shef.ac.uk/chime_workshop/

Names and affiliation of organizers:

- Emmanuel Vincent, Inria, France
- Shinji Watanabe, Mitsubishi Electric Research Laboratories, USA
- Jon Barker, University of Sheffield, UK
- Ricard Marxer, University of Sheffield, UK
- Kean Chin, Google, USA

Description:
This one-day workshop will bring together researchers from the fields of computational hearing, speech enhancement, acoustic modelling and machine learning to discuss the robustness of speech processing in everyday environments, i.e., real-world conditions with acoustic clutter, where the number and nature of the sound sources is unknown and changing over time. As a focus for discussion, the workshop will host CHiME-4 Speech Separation Recognition Challenge.

MLSLP 2016: The 3rd Workshop on Machine Learning in Speech and Language Processing

Date: 13 September 2016
Location: San Francisco, CA
Website: http://ttic.uchicago.edu/~klivescu/MLSLP2016/

Names and affiliation of organizers:

- Karen Livescu, TTI-Chicago, USA
- Mark Hasegawa-Johnson, University of Illinois, USA
- Navdeep Jaitly, Google, USA
- Joseph Keshet, Bar-Ilan University, Israel
- Tara Sainath, Google, USA

Description:
MLSLP is a workshop of SIGML, the ISCA SIG on machine learning in speech and language processing. Prior workshops were held in 2011 and 2012. Speech and language processing is continually mining new ideas from ML and ML, in turn, is devoting more interest to speech and language applications. This workshop aims to be a venue for identifying and incubating the next waves of research directions for interaction and collaboration. In general, the workshop will (1) discuss the emerging research ideas with potential for impact in speech/language and (2) bring together relevant researchers from ML and speech/language who may not regularly interact at conferences.

SLPAT 2016: The 7th Workshop on Speech and Language Processing for Assistive Technologies

Date: 13 September 2016
Location: San Francisco, CA
Website: http://www.slpat.org/slpat2016/

Names and affiliation of organizers:

- Heidi Christensen, University of Sheffield, UK
- François Portet, Laboratoire d'Informatique de Grenoble, France
- Thomas Quatieri, MIT Lincoln Labs, USA
- Frank Rudzicz, University of Toronto, Canada
- Keith Vertanen, Michigan Tech, USA

Description:
Assistive technologies (AT) allow individuals with disabilities to do things that would otherwise be difficult or impossible for them to do. Many examples of assistive technologies involve providing universal access, such as modifications to televisions or telephones to make them accessible to those with vision or hearing impairments. An important sub-discipline within the AT research community is Augmentative and Alternative Communication (AAC), which is focused on communication technologies for those with impairments that interfere with some aspect of human communication, including spoken or written modalities. Speech and natural language processing (NLP) can be used in AT/AAC in a large variety of ways including, for example, improving the intelligibility of unintelligible speech, and providing communicative assistance for frail people or individuals with severe motor impairments. However, there has not been very much interaction in the intersection between researchers of AT/AAC and speech/NLP. This workshop will bring individuals from both of these research communities together with AAC users to share research findings, and to discuss present and future challenges and the potential for collaboration and progress. The workshop has historically had a strong focus on applications and user inclusion.

SIGDIAL 2016: The 17th Annual SIGdial Meeting on Discourse and Dialogue

Dates: 13-15 September 2016
Location: Los Angeles, CA
Website: http://www.sigdial.org/workshops/conference17/

Names and affiliation of organizers:

- Raquel Fernandez, University of Amsterdam, Netherlands
- Wolfgang Minker, Ulm University, Germany
- Jason Williams, Microsoft, USA

Description:
The SIGDIAL venue provides a regular forum for the presentation of cutting edge research in discourse and dialogue to both academic and industry researchers. Continuing with a series of successful sixteen previous meetings, this conference spans the research interest area of discourse and dialogue. The conference is sponsored by the SIGdial organization, which serves as the Special Interest Group in discourse and dialogue for both ACL and ISCA. SIGDIAL 2016 will be co-located with INTERSPEECH 2016 as a satellite event, and also with YRRSDS 2016, the Young Researchers' Roundtable on Spoken Dialog Systems.

SOCIAL PROGRAMS

Welcome Reception

Friday, 9 September 2016
19:00 – 21:00
Atrium Lounge, Hyatt Regency San Francisco

The organizers welcome you to Interspeech 2016! Join us to celebrate the *magic* of San Francisco. Enjoy a sampling of California wine, locally crafted beer and tasty bites in the beautiful open air Atrium of the Hyatt Regency San Francisco. Rendezvous with colleagues and make your plans for the week. We'll see you there!

Reviewer's Reception

Saturday, 10 September 2016
18:30 – 20:00
The City Club of San Francisco, 155 Sansome Street, 10th Floor, San Francisco, California

As a special thank you to the Interspeech 2016 Reviewers, we invite you to the City Club of San Francisco, formerly the Stock Exchange Tower. The 11-story building opened just one year after the stock market crash and housed the offices of brokers who worked on the trading floor of the adjacent San Francisco Stock Exchange. The architect Timothy Pflueger believed that great art should be an integral part of great architecture and commissioned a number of the era's most renowned artists and craftsmen to work on The Stock Exchange Tower. You will find that the building is filled with original art deco furnishings and beautiful art, such as the unique fresco, "Allegory of California" created by Mexican artist Diego Rivera in 1931, located in the grand stairwell.

Hosted beverages and light hors d'oeuvres will be offered.

This event is by invitation only.

Students' Reception

Saturday, 10 September 2016
19:00 – 21:00
Jones, 620 Jones Street, San Francisco, California

One of San Francisco's premier event locations featuring a one-of-a-kind 8,000 square foot courtyard. The perfect place to gather with fellow students for a casual reception under the stars. Hosted beverages and appetizers will be served.

Online pre-registration was required for this event. Onsite purchase will not be available.

Banquet *(Advance purchase required)*

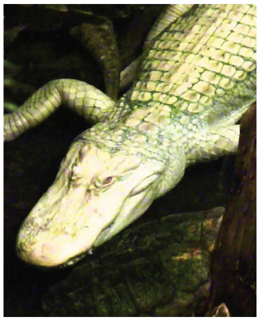
Sunday, 11 September 2016

18:30 – 19:00	Private transfer from Hyatt Regency San Francisco
19:00 – 22:00	Banquet
20:00 – 22:00	Return private transfer from banquet to the Hyatt Regency San Francisco

California Academy of Science, 55 Music Concourse Drive,
San Francisco, California

Attendees will enjoy an evening exploring the museum and noshing on a delectable tasting menu with locations throughout the museum to grab a bite and socialize. Visit the living roof and enjoy the open air with a drink from one of the various hosted bars. Or kick up your heels and dance the night away with some local entertainment. Round trip transfer via school bus from the Hyatt Regency San Francisco will be included with your banquet registration. We hope you will join us, and Claude, the white alligator, the academy's most famous resident, for an evening of revelry!

Online pre-registration was required for this event. Onsite purchase will not be available.



MINDFULNESS SPECIAL EVENT

Saturday, 10 September I 08:00 – 08:25 I Grand Ballroom ABC

Names and affiliation of organizer:

- Nikki Mirghafori, International Computer Science Institute (ICSI), USA

Description:
Mindfulness has entered the cultural mainstream in recent years, with classes and workshops offered on the topic at many universities and companies (including Google, Facebook, etc.). Mindfulness can be thought of as a way to train our mind to be fully present with this moment's experience with curiosity, kindness, and equanimity. The training can serve as a refuge in our busy professional lives and help build resilience. This special event will be in the form of a guided meditation and serve as an introduction for those who are new to this practice, and a chance to practice in community for those who have previous experience. Everyone is welcome.

CLINICAL AND NEUROSCIENCE-INSPIRED VOCAL BIOMARKERS OF NEUROLOGICAL AND PSYCHIATRIC DISORDERS

Saturday, 10 September I 10:00 – 12:00 I Grand Ballroom BC

Names and affiliation of organizers:

- Nicholas Cummins, *Universität Passau, Germany*
- Julien Epps, *University of New South Wales, Australia*
- Emily Mower Provost, *University of Michigan, USA*
- Thomas Quatieri, *MIT Lincoln Laboratory, USA*
- Stefan Scherer, *University of Southern California, Institute for Creative Technologies, USA*

Description:
A variety of neurological and psychiatric conditions can alter a person's behavioral signals. Consequently, research that investigates speech as a way to automatically detect and monitor these conditions has become increasingly popular. This is evident from the growing number of publications in this field over the last five years, including the recent Audio/Visual Emotion Challenge Depression Score Prediction Sub-challenges (AVEC 2013 and AVEC 2014), the recent autism and Parkinson's ComParE challenges at Interspeech 2013 and 2015 respectively, and a depression and suicidality risk assessment tutorial also at Interspeech 2015. However, there is a need to address some key research issues, which are fundamental to characterizing vocal biomarkers for range of neurological and psychiatric conditions.

This combination of a special event with a special session is to increase interactions between speech, neuroscience and clinical communities. The aim of this event is to expose speech processing based neurological and psychiatric research to a wider audience and to foster new interdisciplinary collaborations. Potential topics include: the automatic detection and modelling of depression, post-traumatic stress disorder, traumatic brain injury, suicidality, dementia, Alzheimer's disease, general schizophrenia, Parkinson's disease and autism.

Speech is an attractive signal for use in automated detection of neurological and psychiatric conditions; associated cognitive and physiological alterations influence the process of speech production, affecting the acoustic and linguistic quality of the speech produced in a way that is measurable and possible to objectively assess. However, as speech represents just one potential diagnostic modality, it is important for speech researchers in this field to be conscious of the wide arrange of research into associated biological, physiological and behavioral markers so as to gain an understanding of how speech could be used to augment systems and analysis methods based on these systems. It is also critical that speech researchers gain additional insight into how speech and associated behavioral signals are used in the clinical diagnosis, treatment, and monitoring of these disorders. The special event will provide a focal point for the latest developments within speech-based neurological and psychiatric assessment.

SPEAKER COMPARISON FOR FORENSIC AND INVESTIGATIVE APPLICATIONS II

Saturday, 10 September I 10:00 – 12:00 I Grand Ballroom A

Names and affiliation of organizers:

- Jean-François Bonastre, *LIA, University of Avignon, France*
- Joseph P. Campbell, *MIT Lincoln Laboratory, USA*
- Anders Eriksson, *Stockholm University, Sweden*
- Hiro Nakasone, *Federal Bureau of Investigation, USA*
- Reva Schwartz, *National Institute of Standards and Technology, USA*

Description:
The aim of this special event is to have several structured discussions on speaker comparison for forensic and investigative applications, where many international experts will present their views and participate in the free exchange of ideas. In speaker comparison, speech samples are compared by humans and/or machines for use in investigations or in court to address questions that are of interest to the legal system. Speaker comparison is a high-stakes application that can change people's lives and it demands the best that science has to offer; however, methods, processes, and practices vary widely. These variations are not necessarily for the better and, although recognized, are not generally appreciated and acted upon. Methods, processes, and practices grounded in science are critical for the proper application (and nonapplication) of speaker comparison to a variety of international investigative and forensic applications. This event follows the successful Interspeech 2015 special event of the same name.

SPEECH VENTURES

Monday, 12 September | 10:00 – 12:00 | Grand Ballroom A

Names and affiliation of organizers:

- Korbinian Riedhammer, *Remeeting, USA*
- Nicolas Scheffer, *Facebook, USA*
- Alexandre Lebrun, *Facebook, USA*
- David Suendermann-Oeft, *ETS, USA*

Description:
Interspeech 2016, the world's largest conference on speech technologies to be held in San Francisco, the heart of Silicon Valley, provides a unique opportunity to present the most recent developments and ideas of both academia and industry. Located at the cross-section of the two, startups that are interested in using speech in their products or that want to share their experience in doing so, are invited to participate in the speech venture special event. This event provides a platform for participants to interact with the brightest speech researchers and present and discover new trends in spoken language technology.

The objective of this special event are two-fold:

- Leverage the experience and stories of startups as they adopt speech in their products
- Enable startups to attend a day of the largest conference in speech and meet with researchers, companies, and research institutions

COMPUTATIONAL APPROACHES TO LINGUISTIC CODE SWITCHING

Monday, 12 September | 12:15 – 13:00 | Grand Ballroom A

Names and affiliation of organizers:

- Mona Diab, *George Washington University, USA*
- Pascale Fung, *Hong Kong University of Science and Technology, Hong Kong*
- Julia Hirschberg, *Columbia University, USA*
- Thamar Solorio, *University of Houston, USA*

Description:
Code-switching (CS) is the phenomenon by which multilingual speakers switch back and forth between their common languages in written or spoken communication. CS may occur at the inter-utterance, intra-utterance (mixing of words from multiple languages in the same utterance) and even morphological (mixing of morphemes from different languages) levels. CS presents serious challenges for language technologies such as Automatic Speech Recognition, Language Modeling, Parsing, Machine Translation (MT), Information Retrieval (IR) and Extraction (IE), Keyword Search, and semantic processing. A prime example of this is acoustic modeling and language modeling in automatic speech recognition (ASR): techniques trained on one language quickly break down when there is mixed language input. The lack of basic tools such as language models, part-of-speech (POS) taggers and parsers trained on such mixed language data makes downstream tasks even more challenging. Even for problems that are largely considered solved for monolingual corpora, such as Language Identification, or POS Tagging, performance degrades at a rate proportional to the amount and level of mixed-language present in the data.

This special event is to bring together researchers interested in solving the CS problem, to raise community awareness of the (limited) resources available and the work currently underway for the study of CS, with particular emphasis on work in the speech community. The format will consist of a short introduction from the organizers followed by discussion. We held a workshop in CS in conjunction with EMNLP 2014, developing a shared text-based task for this purpose. We received 18 regular workshop submissions and accepted 8. The goal of this event is to engage the speech processing community now working in this area and to encourage new research by those now working primarily with monolingual corpora.

We will solicit participation from researchers working in speech processing for the analysis and/processing of CS data. Topics of relevance to the event will include the following:

- Methods for improving ASR acoustic and language models in code switched data
- Domain/dialect/genre adaptation techniques applied to CS data processing
- Challenges of language identification in CS data.
- Speech-to-speech translation in CS data
- Keyword search in CS data
- Cross-lingual approaches to CS
- Development of corpora to support research on CS data
- Crowdsourcing approaches for the annotation of code switched data

YOUNG FEMALE RESEARCHERS IN SPEECH SCIENCE & TECHNOLOGY

Thursday, 8 September | 09:00 – 16:00 | 1 Market Street, Suite 400, Spear Tower, San Francisco | Room 1MST-4 CharlesCrocker

Names and affiliation of organizers:

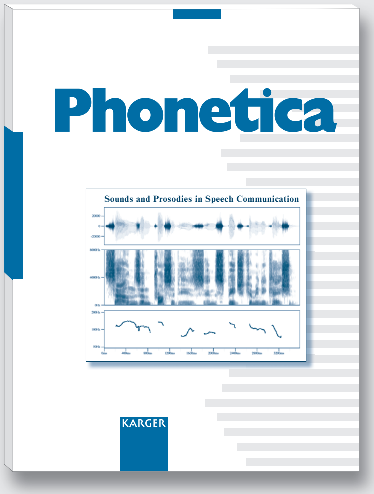
- Co-Chair: Abeer Alwan, *University of California, Los Angeles, USA*
- Co-Chair: Julia Hirschberg, *Columbia University, USA*
- Mary Beckman, *Ohio State University, USA*
- Carol Espy-Wilson, *University of Maryland at College Park, USA*
- Dilek Hakkani-Tur, *Microsoft, USA*
- Pascale Fung, *Hong Kong University of Science and Technology, Hong Kong*
- Esther Judd, *ReadSpeaker, USA*
- Lori Lamel, *LIMSI, France*
- Karen Livescu, *TTI-Chicago, USA*
- Yang Liu, *University of Texas, Dallas, USA*
- Helen Meng, *Chinese University of Hong Kong, Hong Kong*
- Mari Ostendorf, *University Washington, USA*
- Bhuvana Ramabhadran, *IBM, USA*
- Liz Shriberg, *SRI International, USA*
- Isabel Trancoso, *INESC, Portugal*
- Petra Wagner, *University Bielefeld, Germany*

Description:

Women in Science is a workshop for women undergraduate and masters students who are currently working in speech science and technology at Interspeech 2016. This workshop is designed to foster interest in research in our field in women at the undergraduate or masters 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 on individual or group projects. The motivation for the workshop is the realization on the part of many women in our field that there seem to be relatively few younger women at Interspeech conferences in the 'pipeline' to careers in speech. We wish to address this problem by providing a venue where younger students can present their work with more senior women as mentors and where senior women can describe their own research experience to the students.

An interdisciplinary forum for spoken language research

Phonetica



Contemporary research into spoken language employs a wide range of approaches, from instrumental measures to perceptual and neurocognitive measures, to computational models, for investigating the properties and principles of speech in communicative settings across the world's languages. *Phonetica* is an international interdisciplinary forum for phonetic science that covers all aspects of the subject matter, from phonetic and phonological descriptions of segments and prosodies to speech physiology, articulation, acoustics, perception, acquisition, and phonetic variation and change. *Phonetica* thus provides a platform for a comprehensive understanding of speaker-hearer interaction across languages and dialects, and of learning contexts throughout the lifespan. Papers published in this journal report expert original work that deals both with theoretical issues and with new empirical data, as well as with innovative methods and applications that will help to advance the field.

Phonetica
Founded: 1957
Category: Basic Research
Fields of Interest: Phonetics, Communication Disorders

Listed in bibliographic services, including
Journal Citation Reports, PubMed/MEDLINE,
Biological Abstracts, Embase

2016: Volume 73
4 issues per volume
Language: English
ISSN 0031–8388
e-ISSN 1423–0321

Impact Factor: 0.520

Selected contributions

- Welcome Editorial: Change and Continuity in *Phonetica:* **Best, C.** (Bankstown, N.S.W.)
- Accommodation of End-State Comfort Reveals Subphonemic Planning in Speech: **Derrick, D.** (Christchurch); **Gick, B.** (Vancouver, B.C.)
- Perceptual Assimilation and Discrimination of Non-Native Vowel Contrasts: **Tyler, M.D.** (Sydney, N.S.W.); **Best, C.T.** (Sydney, N.S.W./New Haven, Conn.); **Faber, A.** (New Haven, Conn./Middletown, Conn.); **Levitt, A.G.** (New Haven, Conn./Wellesley, Mass.)
- A Comparative Analysis of Media Lengua and Quichua Vowel Production: **Stewart, J.** (Winnipeg, Man.)
- Acoustic Correlates of English Rhythmic Patterns for American versus Japanese Speakers: **Mori, Y.** (Kyoto); **Hori, T.** (Tokyo); **Erickson, D.** (Ishikawa)
- Individual Differences in Learning Talker Categories: The Role of Working Memory: **Levi, S.V.** (New York, N.Y.)
- Gestural Control in the English Past-Tense Suffix: An Articulatory Study Using Real-Time MRI: **Lammert, A.; Goldstein, L.; Ramanarayanan, V.; Narayanan, S.** (Los Angeles, Calif.)
- The 'Whistled' Fricative in Xitsonga: Its Articulation and Acoustics: **Lee-Kim, S.-I.** (New York, N.Y.); **Kawahara, S.** (Tokyo); **Lee, S.J.** (New Britain, Conn.)

More information at **www.karger.com/pho**

KARGER

KF16092_Insp

Student Events Organized by ISCA-SAC

2ND DOCTORAL CONSORTIUM

Thursday, 8 September | ICSI, Berkeley, California

Organized by the Student Advisory Committee of the International Speech Communication Association (ISCA-SAC), the Doctoral Consortium provides students an opportunity to present their PhD projects to experts and peers and receive valuable feedback on research plans as well as interesting new ideas. Each talk will last 20 minutes and be followed by 20 minutes of discussion. This forum provides an invaluable and enriching experience for the students.

STUDENTS MEET EXPERTS

Friday, 9 September | 15:30 - 17:30 | Waterfront AB

The Student Advisory Committee of the International Speech Communication Association (ISCA-SAC) is very proud to announce that this year the Student meet Expert Event will be back to Interspeech in an exciting new format.

ISCA-SAC Contacts:

- Catharine Oertel Genannt Bierbach (ISCA-SAC)
- Angel Mario Castro (ISCA-SAC)
- Lori Lamel (ISCA)
- Elizabeth Shriberg (Local Coordinator)

Thursday, 8 September

Registration	08:00 - 19:00		Grand Ballroom Foyer
Speaker Check-In	07:30 - 17:00		Regency AB
Morning Tutorials | 08:30 - 10:00			
Bayview A	Bayview B	Seacliff A	Seacliff BCD
Attending to Speech and Audio	Machine Learning for Speaker Recognition	Spoken Content Retrieval - Beyond Cascading Speech Recognition with Text Retrieval	Recent Advances in Distant Speech Recognition
Refreshment Break 10:00 - 10:30 | Seacliff Foyer			
Morning Tutorials Continued | 10:30 - 12:00			
Bayview A	Bayview B	Seacliff A	Seacliff BCD
Attending to Speech and Audio	Machine Learning for Speaker Recognition	Spoken Content Retrieval - Beyond Cascading Speech Recognition with Text Retrieval	Recent Advances in Distant Speech Recognition
Lunch Break 12:00 - 13:00			
Afternoon Tutorials | 13:00 - 14:30			
Bayview A	Bayview B	Seacliff A	Seacliff BCD
Singing Synthesis	Pushing the Frontiers of Speech Processing – What Does It Take to Tackle New Languages and Domains?	Hearing Assistive Technologies: Challenges and Opportunities	Data-Driven Approaches to Speech Enhancement and Separation
Refreshment Break 14:30 - 15:00 | Seacliff Foyer			
Afternoon Tutorials Continued | 15:00 - 16:30			
Bayview A	Bayview B	Seacliff A	Seacliff BCD
Singing Synthesis	Pushing the Frontiers of Speech Processing – What Does It Take to Tackle New Languages and Domains?	Hearing Assistive Technologies: Challenges and Opportunities	Data-Driven Approaches to Speech Enhancement and Separation

TUTORIALS

MORNING TUTORIALS

ATTENDING TO SPEECH AND AUDIO

Thursday, 8 September | 08:30 – 12:00 | Bayview A

Names and affiliation of organizer:

- Malcolm Slaney, *Google Machine Hearing Research, USA*

Abstract:
This tutorial brings together a number of topics related to attention and listening effort that are important for speech practitioners and researchers interested in the next generation of speech applications. This tutorial will describe applications, auditory saliency models, topdown vs. bottomup attention, an attentional model of auditory scene analysis, listening effort (which is limited by attention) as a measure of speech quality, using attention for ASR, and finally decoding attention from EEG, MEG and ECoG signals.

MACHINE LEARNING FOR SPEAKER RECOGNITION

Thursday, 8 September | 08:30 – 12:00 | Bayview B

Names and affiliation of organizers:

- Man-Wai Mak, *The Hong Kong Polytechnic University, Hong Kong*
- Jen-Tzung Chien, *National Chiao Tung University, Taiwan*

Abstract:
In this tutorial, we will present state-of-the-art techniques for speaker recognition and some related tasks such as speaker diarization. This tutorial shall cover different components in speaker recognition including front-end feature extraction, back-end modeling and scoring. A range of learning models will be detailed – from GMMs, SVM, PLDA to deep neural networks – along with learning algorithms – from Bayesian learning, unsupervised learning, discriminative learning, transfer learning to deep learning. A series of case studies and modern models based on PLDA and DNN will be addressed. In particular, different variants of deep models and their solutions to different problems in speaker recognition are presented. In addition, we will point out some new trends for speaker recognition including model regularization and deep belief networks.

SPOKEN CONTENT RETRIEVAL – BEYOND CASCADING SPEECH RECOGNITION WITH TEXT RETRIEVAL

Thursday, 8 September | 08:30 – 12:00 | Seacliff A

Names and affiliation of organizers:

- Lin-shan Lee, *National Taiwan University, Taiwan*
- Hung-yi Lee, *National Taiwan University, Taiwan*

Abstract:
Spoken content retrieval refers to retrieving spoken content direcly based on the audio without relying on the text descriptions offered by the context provider. It has been very successful with the basic approach of cascading automatic speech recognition (ASR) with text information retrieval: after the spoken content is transcribed into text or lattice format, a text retrieval engine searches over the ASR output to find desired information. This framework works well when the ASR accuracy is relatively high, but becomes less adequate when more challenging real-world scenarios are considered, since retrieval performance depends heavily on ASR accuracy. This leads to the emergence of another approach to spoken content retrieval: to go beyond the basic framework of cascading ASR with text retrieval, in order to have retrieval performance less dependent on ASR accuracy. This tutorial is intended to provide an overview of the major technical contributions along this second line of investigation.

RECENT ADVANCES IN DISTANT SPEECH RECOGNITION

Thursday, 8 September | 08:30 – 12:00 | Seacliff BCD

Names and affiliation of organizers:

- Marc Delcroix, *NTT Communication Science Laboratories, Japan*
- Shinji Watanabe, *Mitsubishi Electric Research Laboratories, USA*

Abstract:
Automatic speech recognition (ASR) is being deployed successfully more and more in products such as voice search applications for mobile devices. However, it remains challenging to perform recognition when the speaker is distant from the microphone, because of the presence of noise, attenuation, and reverberation. Research on distant ASR has received increased attention, and has progressed rapidly due to the emergence of 1) deep neural network (DNN) based ASR systems, 2) the launch of recent challenges such as CHiME series, REVERB, ASpIRE, and DIRHA, and 3) the development of new products such as the Microsoft Kinect and the AMAZON Echo. This tutorial will review the recent progresses made in the field of distant speech recognition in the DNN era, including single and multi-channel speech enhancement front-ends, and acoustic modeling techniques for robust back-ends. The tutorial will also introduce practical schemes for building distant ASR systems based on the expertise acquired from past challenges.

AFTERNOON TUTORIALS

SINGING SYNTHESIS

Thursday, 8 September | 13:00 – 16:30 | Bayview A

Names and affiliation of organizer:

- Christophe d'Alessandro, *LIMSI-CNRS, France*

Abstract:
Singing synthesis touches on several areas in Speech Communication, like voice quality, vocal emotion and expression, text-to-speech synthesis, voice personality, speech signal modeling and synthesis, voice transformation, and of course music. It is as old as speech synthesis, with a number of challenging research questions, and several successful applications in the music industry, including movies soundtracks, avant-garde contemporary music, or artificial characters in pop music. This tutorial aims at presenting: 1/ a review of the scientific bases, history and open questions in singing synthesis research; 2/ a state of the art in the design, applications and evaluation of text-to-chant system, singing instruments, and singing processing systems.

PUSHING THE FRONTIERS OF SPEECH PROCESSING – WHAT DOES IT TAKE TO TACKLE NEW LANGUAGES AND DOMAINS?

Thursday, 8 September | 13:00 – 16:30 | Bayview B

Names and affiliation of organizers:

- Samuel Thomas, *IBM T.J. Watson Research Center, USA*
- Florian Metze, *Carnegie Mellon University, USA*
- Brian Kingsbury, *IBM T.J. Watson Research Center, USA*
- Bhuvana Ramabhadran, *IBM T.J. Watson Research Center, USA*

Abstract:
Although the realworld impact of speech technology has grown significantly in the past few years, especially in mobile applications, this growth has largely been limited to only well studied languages and domains. Speech technologies must become truly universal by being available in the numerous languages and dialects spoken by people across the globe, even under low resource conditions. The performance of these technologies can also disappoint when they are deployed in new domains other than those available during training, even in well researched languages. This tutorial reviews technological breakthroughs in building speech processing systems in new languages and domains. The tutorial draws on techniques and results from several evaluation campaigns, including MediaEval's QUESST (Spoken Web Search), the IARPA Babel and NIST OpenKWS evaluations, and the IARPA ASpIRE challenge.

HEARING ASSISTIVE TECHNOLOGIES: CHALLENGES AND OPPORTUNITIES

Thursday, 8 September | 13:00 – 16:30 | Seacliff A

Names and affiliation of organizers:

- Oldooz Hazrati, *University of Texas at Dallas, USA*
- Hussnain Ali, *University of Texas at Dallas, USA*
- John H.L. Hansen, *University of Texas at Dallas, USA*
- James M. Kates, *University of Colorado Boulder, USA*

Abstract:
This tutorial will cover an overview of hearing assistive devices (e.g. hearing aids and cochlear implants), challenging listening environments (e.g. noisy, reverberant, whisper/vocal effort, babble noise), current advancements and technologies, as well as future directions (e.g. naturalistic evaluations, next generation spaces).

DATA-DRIVEN APPROACHES TO SPEECH ENHANCEMENT AND SEPARATION

Thursday, 8 September | 13:00 – 16:30 | Seacliff BCD

Names and affiliation of organizer:

- Jonathan Le Roux, *Mitsubishi Electric Research Labs (MERL), USA*
- Emmanuel Vincent, *Inria, France*
- Hakan Erdogan, *Sabanci University, Turkey*

Abstract:
Being able to isolate a target speech signal from background signals is of direct importance for telephony, hands-free communication and audio surveillance, and it is also critical as a pre-processing step in applications such as voice activity detection, automatic speaker identification, and most importantly automatic speech recognition (ASR) in challenging environments. While speech enhancement and separation methods originally did not rely on training, there has recently been an explosion in the use of machine learning based methods that exploit large amounts of training data. This tutorial will present a broad overview of these methods, analyzing the insights that can be gained from the pre-deep-learning era of graphical modeling and NMF approaches, then diving into an in-depth presentation of recent deep learning approaches encompassing single-channel methods, multi-channel methods, and new directions.

KEYNOTES

FRIDAY, 9 SEPTEMBER 2016 I 09:30 – 10:30 I GRAND BALLROOM ABC

John Makhoul
BBN Technologies
Cambridge, Massachusetts, USA

A 50-year Retrospective on Speech and Language Processing

This talk is a retrospective of speech and language processing as witnessed by the speaker during the last 50 years. From exploratory scientific beginnings that emphasized the discovery of how speech is produced and perceived by humans to today's plethora of applications using our technology, our field has witnessed explosive growth. The talk will review the historical development of our community and some of the key technical ideas that have shaped our field. Some of the ideas were influenced by developments in other fields, while some of the developments in our field have been instrumental in key advances in other fields, such as optical character recognition and machine translation. Important developments include the source-filter model, digital signal processing, linear prediction, vector quantization, deep neural networks, and statistical modeling methods, especially hidden Markov models (HMMs), with primary applications to speech analysis, synthesis, coding, and recognition. The talk will be sprinkled with lessons learned in the importance of various factors in performing our research, and will be peppered with interesting tidbits about key moments in the development of our technology. The talk will end with a brief prospective peek at the next 50 years.

Biography

John Makhoul received the B.E. degree from the American University of Beirut in 1964, the M.Sc. degree from the Ohio State University in 1965, and the Ph.D. degree from the Massachusetts Institute of Technology in 1970, all in electrical engineering.

Since 1970, Dr. Makhoul has been with BBN Technologies, Cambridge, MA, where he has been working on various aspects of speech and language processing, including speech analysis and synthesis, speech coding, speech recognition, speech enhancement, digital signal processing, artificial neural networks, human-machine interaction using voice, optical character recognition, and machine translation, including speech-to-speech translation.

Dr. Makhoul is also an Adjunct Professor at Northeastern University. He has served as a member of several panels of the National Research Council and chaired panels in the areas of speech recognition and speech enhancement.

Dr. Makhoul is a Fellow of the IEEE, the Acoustical Society of America, and the International Speech Communication Association. He is the recipient of the 1978 Best Paper Award, the 1982 Technical Achievement Award, and the 1988 Society Award of the IEEE Signal Processing Society. In 2000, he was awarded the IEEE Third Millenium Medal. In 2009, he received the IEEE James L. Flanagan Speech and Audio Processing Award for his "pioneering contributions to speech modeling."

SATURDAY, 10 SEPTEMBER 2016 I 08:30 – 09:30 I GRAND BALLROOM ABC

Edward Chang
University of California, San Francisco
San Francisco, California, USA

The Human Speech Cortex

A unique and defining trait of human behavior is our ability to communicate through speech. The fundamental organizational principles of the neural circuits within speech brain areas are largely unknown. In this talk, I will present new results from our research on the functional organization of the human higher-order auditory cortex, known as Wernickes area. I will focus on how neural populations in the superior temporal lobe encode acoustic-phonetic representations of speech, and also how they integrate influences of linguistic context to achieve perceptual robustness.

Biography

Eddie Chang, MD, a neurosurgeon-scientist, stands at the threshold between novel technology and brain circuitry with his expertise in the development of brain machine interface devices. Chang specializes in operative brain mapping, enabling him to not only carry out the safest possible surgeries but also make the most precise, real-time measurements of human brain activity currently available. His work has led to an unprecedented level of understanding of the brain circuitry underlying speech perception and production. Chang studies the patterns in the brain that orchestrate the lips, tongue, jaw, and larynx for development of speech prosthetic device for patients who have lost their capacity to speak due to brain injury, ALS, or stroke. He has been awarded the 2015 Blavatnik National Laureate in Life Sciences, NIH New Innovator Award, and the Robertson Fellow of the New York Stem Cell Foundation.

SUNDAY, 11 SEPTEMBER 2016 | 08:30 – 09:30 | GRAND BALLROOM ABC

Anne Fernald
Stanford University
Stanford, California, USA

Talking with Kids Really Matters: Early Language Experience Shapes Later Life Chances

The foundation for lifelong literacy is built through a child's experience with language in the first five years. Integrating research from biological, psycholinguistic, and sociocultural perspectives, I will examine why millions of children fail to reach their developmental potential in the early years and enter school without a strong foundation for learning, resulting in enormous loss of human potential.

Biography

Anne Fernald is the Josephine Knotts Knowles Professor of Human Biology and director of the Language Learning Lab in the Department of Psychology at Stanford University. Fernald's early cross-linguistic research on the melodic intonation of caregiver's speech to children provided the first acoustic analyses of prosody in child-directed speech. She and her team then moved on to studies of infant language comprehension using real-time processing measures. Their longitudinal studies with rich and poor children from linguistically, culturally, and economically diverse families reveal the vital role of early language experience in strengthening children's speech processing skills, which in turn facilitate vocabulary growth and language learning. New results from their Habla Conmigo Academy intervention program with low-income Spanish-speaking families in San Jose CA show that when Latino parents increase their verbal engagement with their toddlers, their children show significantly greater gains in language proficiency over the second year. Fernald also studies caregiver-child interactions in rural villages in Senegal, where traditional taboos against talking to infants may influence the course of early language learning. A central goal is of this translational research program is to provide rigorous scientific evidence that parents from diverse sociocultural backgrounds can play a vital role in supporting their children's language and cognitive development.

MONDAY, 12 SEPTEMBER 2016 | 08:30 – 09:30 | GRAND BALLROOM ABC

Dan Jurafsky
Stanford University
Stanford, California, USA

Ketchup, Interdisciplinarity, and the Spread of Innovation in Speech and Language Processing

I show how natural language processing can help model the spread of innovation through scientific communities, with special focus on the history of speech and language processing, and the important role of interdisciplinarity.

Biography

Dan Jurafsky is Professor and Chair of Linguistics and Professor of Computer Science, at Stanford University. He is a computational linguist, with special interests in the automatic extraction of meaning from speech and text in English and Chinese. His most recent work has focused on applying natural language processing to the behavioral and social sciences. Dan is a 2002 MacArthur Fellowship recipient, co-wrote the widely-used textbook "Speech and Language Processing and is also interested in the linguistics of food. His latest book, "The Language of Food: A Linguist Reads the Menu", was nominated for the 2015 James Beard Award.

SPECIAL SESSIONS

AUDITORY-VISUAL EXPRESSIVE SPEECH AND GESTURE IN HUMANS AND MACHINES
(Area 13: Speech and Spoken-language based Multimodal Processing and Systems)

Friday, 9 September | 11:00 – 13:00 | Grand Ballroom BC
Friday, 9 September | 14:30 – 16:30 | Pacific Concourse Poster A

Names and affiliation of organizers:

- Jeesun Kim, *Associate Professor, The MARCS Institute, Western Sydney University, Australia*
- Gérard Bailly, Professor, *GIPSA-Lab/Speech & Cognition dpt., CNRS/Grenoble-Alpes University, France*

Description:
The topic 'Auditory-visual expressive speech and gesture in humans and machines', encompasses many research fields and is relevant to researchers: who investigate the role of the talker's face and head movements in human face-to-face communication; who are interested in the relationship between speech and gesture; and who are working to develop platforms for human-machine communication (e.g., a key topic for sociable humanoid robots).

The proposed session aims to bring together these researchers to create the focus and critical mass for effective interaction, more specifically to enable sharing of techniques and investigative methods and research findings. It will provide a forum for researchers to explore how studies about human communication may be relevant for enabling social machines. Conversely, it will provide an opportunity for researchers working with machines to showcase developments in their field. The feedback between the two communities will be stimulating and rewarding.

THE REDDOTS CHALLENGE: TOWARDS CHARACTERIZING SPEAKERS FROM SHORT UTTERANCES
(Area 4: Speaker and Language Identification)

Friday, 9 September | 14:30 – 16:30 | Grand Ballroom BC

Names and affiliation of organizers:

- Kong Aik Lee, *Institute for Infocomm Research (I2R), A*STAR, Singapore*
- Anthony Larcher, *LIUM, Université du Maine, France*
- Hagai Aronowitz, *IBM Research Haifa, Israel*
- Guangsen Wang, *Institute for Infocomm Research (I2R), A*STAR, Singapore*
- Patrick Kenny, *CRIM, Canada*

Description:
The RedDots project was initiated, with collaboration from multiple sites, as a follow-up to a special session during INTERSPEECH 2014. It was set out to collect speech data through mobile crowd-sourcing, with the benefit of potentially wider population and greater diversity. The project was rolled out in January 29, 2015. At the time of writing, the project has recruited 89 speakers (72 male, 17 female) from 21 countries, with a total of 875 complete sessions.

The purpose of this special session is to gather the research efforts towards a common goal of exploring new directions and better understanding of speaker-channel-phonetic variability modelling for text-dependent and text-prompted speaker verification over short utterances. To get the RedDots database, please contact the organizers or visit the page http://goo.gl/forms/Dpk3OiJkWV.

INTELLIGIBILITY UNDER THE MICROSCOPE
(Area 1: Speech Perception, Production, and Acquisition)

Friday, 9 September | 14:30 – 16:30 | Pacific Concourse Poster B

Names and affiliation of organizers:

- Ricard Marxer, *University of Sheffield, UK*
- Martin Cooke, *Ikerbasque (Basque Foundation for Science), Spain*
- Jon P. Barker, *University of Sheffield, UK*

Description:
Existing models of intelligibility can successfully estimate word identification in broadly stated noise conditions. These predictions may be characterized as 'macroscopic' in that they represent averages -- averages over many listeners and over many speech tokens. This Special Session asks whether we can go beyond macroscopic predictions. We invite work that might contribute to a new breed of 'microscopic' models that are evaluated according to their ability to make precise predictions of what a specific listener might hear in response to a specific noisy speech token. Developing such models will deepen our understanding of speech perception and enable a wealth of new intelligibility modeling applications. To focus the session, we will provide contributers with access to two large corpora that record 'slips of the ear' made by listeners hearing words in complex noise backgrounds. Participants will be encouraged to use this data where it can support the aims of their work.

THE SPEAKERS IN THE WILD (SITW) SPEAKER RECOGNITION CHALLENGE
(Area 4: Speaker and Language Identification)

Friday, 9 September | 17:00 – 19:00 | Grand Ballroom BC

Names and affiliation of organizers:

- Mitchell McLaren, *Speech Technology and Research (STAR) Laboratory at SRI International, Menlo Park, California, USA*
- Aaron Lawson, *Speech Technology and Research (STAR) Laboratory at SRI International, Menlo Park, California, USA*
- Luciana Ferrer, *Departamento de Computacion, FCEN, Universidad de Buenos Aires and CONICET, Argentina*
- Diego Castán, *Speech Technology and Research (STAR) Laboratory at SRI International, Menlo Park, California USA*

Description:
The Speakers in the Wild (SITW) speaker recognition challenge will focus on the challenges of applying current speaker recognition technology to unconstrained conditions of real-world data. The challenge is based on a newly collected database of speech samples from open source media consisting of single and multi-speaker audio acquired across unconstrained or 'wild' conditions. Multiple speech samples from nearly three hundred individuals are represented in the database, with all noise, reverb, compression and other artifacts being natural characteristics of the original audio. Challenge participants will be provided with the SITW database on which they can benchmark current technologies and explore new high-risk techniques for the task of speaker recognition under the conditions exhibited in the data. The special session will be dedicated to the discussion of applied technology, performance thereof and any issues highlighted as a result of the challenge.

CLINICAL AND NEUROSCIENCE-INSPIRED VOCAL BIOMARKERS OF NEUROLOGICAL AND PSYCHIATRIC DISORDERS
(Area 3: Analysis of Paralinguistics in Speech and Language)

Saturday, 10 September | 10:00 – 12:00 | Grand Ballroom BC
Saturday, 10 September | 13:30 – 15:30 | Pacific Concourse Poster D

Names and affiliation of organizers:

- Nicholas Cummins, *Universität Passau, Germany*
- Julien Epps, *University of New South Wales, Australia, Data61, Australia*
- Emily Mower Provost, *University of Michigan, USA*
- Thomas Quatieri, *MIT Lincoln Laboratory, USA*
- Stefan Scherer, *University of Southern California Institute for Creative Technologies, USA*

Description:
This session will focus on the latest developments within speech-based neurological and psychiatric assessment. Topics will include (but not limited to) the automatic detection of depression, PTSD, schizophrenia, traumatic brain injury, dementia, Parkinson's disease and autism. Participants are encouraged to target the following themes: (i) Novel clinically- or neuroscience-motivated analysis methods and vocal features: features designed to capture speech effects specific to one or more conditions, (ii) Nuisance Variability Compensation: removing effects of comorbid conditions or unwanted acoustic variability, (iii) Clinical utility and quantifying uncertainty: considerations of clinical utility or systems that self-determine a level of uncertainty associated with detection, and (iv) Cross-corpus studies; analysing the similarities and differences in speech patterns between different conditions or different recording paradigms.

SINGING SYNTHESIS CHALLENGE: FILL-IN-THE GAP
(Area 7: Speech Synthesis and Spoken Language Generation)

Saturday, 10 September | 10:00 – 12:00 | Bayview A

Names and affiliation of organizers:

- Christophe d'Alessandro, *LIMSI-CNRS, France*
- Axel Roebel, *IRCAM-CNRS-UMPC, France*
- Olivier Deroo, *ACAPELA Group, Belgium*

Description:
The special session "Singing Synthesis Challenge: Fill-In the Gap" aims at bringing together research teams working on singing synthesis from all over the world by means of proposing a common challenge. The challenge will be to fill-in the gap in a well-known song (e.g. "Autumn leaves"), i.e. to synthesize a new, specially written couplet. The new couplet includes new lyrics, and possibly a new melody, to be inserted in the song. The chosen song is a top hit song, so that a large number of interpretations are available on the net and can be used for reference, acoustic analysis, machine learning, comparison etc. All aspects of singing synthesis and all methodologies are welcome, including both off-line (studio) singing synthesis systems, with no limits on time for producing the result, and performative (real-time) singing instruments. While contributors are encouraged to produce a singing synthesis, other aspects like evaluation methodologies are welcome and will be considered as valid contributions. Interspeech 2016 attendants will be given the opportunity to vote for their preferred synthetic song.

SHARING RESEARCH AND EDUCATION RESOURCES FOR UNDERSTANDING SPEECH PROCESSING

(Area 10: Speech Recognition—Technologies and Systems for New Applications)
Saturday, 10 September | 13:30 – 15:30 | Grand Ballroom BC

Names and affiliation of organizers:

- Eric Fosler-Lussier, *The Ohio State University, USA*
- Rebecca Bates, *Minnesota State University, Mankato, USA*
- Florian Metze, *Carnegie Mellon University, USA*

Description:
Speech processing systems have become increasingly complex and difficult to share across sites. Significant time is spent reimplementing published methods; even when software is shared, the lack of common environments between sites means that reproducing results can require significant effort. Open software repositories, virtual machines, and tools for automatically building container environments in the cloud are beginning to facilitate cross-site collaboration.

Featured virtual machines will be shared with session attendees in order to foster community discussion and encourage use after the session.

VOICE CONVERSION CHALLENGE 2016

(Area 7: Speech Synthesis and Spoken Language Generation)
Saturday, 10 September | 13:30 – 15:30 | Bayview A

Names and affiliation of organizers:

- Tomoki Toda, *Nagoya University, Japan*
- Junichi Yamagishi, *National Institute of Informatics, Japan & University of Edinburgh, UK*
- Fernando Villavicencio, *National Institute of Informatics, Japan*
- Zhizheng Wu, *University of Edinburgh, UK*
- Ling-Hui Chen, *University of Science and Technology of China, China*
- Daisuke Saito, *University of Tokyo, Japan*
- Mirjam Wester, *University of Edinburgh, UK*

Description:
The focus of our Special Session is on better understanding and comparing the current performance of various voice conversion techniques on identical speech corpora. It is a Special Session with an incorporated, standard challenge "Voice Conversion Challenge 2016." Authors submitting papers to the Special Session will be encouraged to submit results assessing converted voices using a new, standard database provided in the Challenge. The task of the Challenge focuses on conversion of five source speaker's voices to five different target speaker's voices. In total, 25 conversion cases will be evaluated in terms of perceived naturalness and similarity via listening tests. This will help different research groups working on voice conversion to converge on common tasks under common conditions and it will enable us to share our views about the unsolved problems and challenges behind the technologies.

COMPUTATIONAL PARALINGUISTICS CHALLENGE (ComParE): DECEPTION, SINCERITY & NATIVE LANGUAGE

(Area 3: Analysis of Paralinguistics in Speech and Language)
Sunday, 11 September | 10:00 – 12:00 | Grand Ballroom BC
Sunday, 11 September | 13:30 – 15:30 | Grand Ballroom BC

Names and affiliation of organizers:

- Björn Schuller, *University of Passau, Germany & Imperial College London, UK*
- Stefan Steidl, *FAU Erlangen-Nuremberg, Germany*
- Anton Batliner, *TUM, Germany*
- Julia Hirschberg, *Columbia University, New York, USA*
- Judee K. Burgoon, *University of Arizona, Tucson, USA*
- Eduardo Coutinho, *University of Liverpool, UK & Imperial College London, UK*

Description:
The Interspeech 2016 Computational Paralinguistics ChallengE (ComParE) is an open Challenge dealing with states of speakers as manifested in in their speech characteristics. There have so far been seven consecutive Challenges at INTERSPEECH since 2009, but there still exists a multiplicity of not yet covered, but highly relevant paralinguistic phenomena.

Thus, we introduce two new tasks by the Deception Sub-Challenge and the Sincerity Sub-Challenge. All data, including features that may be used, are provided by the organizers. Based on speech analysis, in the Deception Sub-Challenge, it has to be automatically determined whether speech is deceptive or not, and in the Sincerity Sub-Challenge, the degree of perceived sincerity has to be determined.

Results of the Challenge will be presented at Interspeech 2016 and Prizes will be awarded to the Sub-Challenge winners.

SPEECH, AUDIO, AND LANGUAGE PROCESSING TECHNIQUES APPLIED TO BIRD AND ANIMAL VOCALISATIONS

(Area 5: Analysis of Speech and Audio Signals)

Sunday, 11 September | 10:00 – 12:00 | Bayview A
Sunday, 11 September | 13:30 – 15:30 | Pacific Concourse Poster B

Names and affiliation of organizers:

- Naomi Harte, *Trinity College Dublin, Ireland*
- Peter Jancovic, *University of Birmingham, UK*
- Karl-L. Schuchmann, *Zoological Research Museum Alexander Koenig & University of Bonn, Germany*

Description:
The ability to analyze sounds from animals and birds has important implications for understanding the biodiversity of different regions of the world, finding and tracking populations of rare species, and understanding communication in species other than humans. Knowledge in the speech processing community can inform and transform the analysis, classification and understanding of these vocalisations within the wider scientific community. Numerous collaborations have already developed between researchers in the areas of speech, audio and language and those in the ornithology and zoology community. This special session aims to bring together researchers from both sides to explore state of the art, consider major challenges in this domain, and identify potential areas for collaboration. Our target audience is both those already involved in such research, and any Interspeech attendee who may like to get involved in this exciting area of research.

REALISM IN ROBUST SPEECH PROCESSING

(Area 10: Speech Recognition—Technologies and Systems for New Applications)

Sunday, 11 September | 16:00 – 18:00 | Grand Ballroom BC

Names and affiliation of organizers:

- Dayana Ribas, *CENATAV, Cuba*
- Emmanuel Vincent, *Inria, France*
- John H.L. Hansen, *Univ. of Texas at Dallas, USA*

Description:
One of the challenges currently faced in speech processing is the migration of laboratory results to real applications. As performing evaluations in the targeted scenarios of use is difficult to carry out, researchers have resorted to simulating data in controlled scenarios. However, many datasets include some levels and types of distortion that never happen in real life. This can result in satisfactory performance on scenarios that will never happen in practice, while the performance may be much worse in real scenarios. Furthermore, complex and expensive methods might be obtained that are actually not required.

This session aims to provide a forum for the cross-fertilization of expertise and experimental evidence about "realism" across different areas of robust speech processing. Through the study of the state of the art and the exchange of specialized experiences, we aim to characterize real scenarios by measuring the ranges and combinations of different parameters, and to establish "good practices" regarding which parameter violations are acceptable or not given the task to be solved and the limitations of today's data collection and simulation tools.

SPEECH AND LANGUAGE TECHNOLOGIES FOR HUMAN-MACHINE CONVERSATION-BASED LANGUAGE EDUCATION

(Area 10: Speech Recognition—Technologies and Systems for New Applications)

Monday, 12 September | 10:00 – 12:00 | Grand Ballroom BC

Names and affiliation of organizers:

- Yao Qian, *Educational Testing Service, USA*
- Helen Meng, *The Chinese University of Hong Kong, Hong Kong SAR*
- Frank K. Soong, *Microsoft Research, China*

Description:
This special session aims to promote research on the state-of-the-art speech and language technologies for human-machine conversation-based language learning. Recent advances in deep learning with big data have improved significantly the performance of speech recognition, dialogue management, language understanding and machine translation, which bring conversation-based language learning and assessment supported by machines much closer to reality and commercialization.

We like to invite researchers and engineers who worked actively in computer-aided, audio-visual language learning, including but not limited to the following topics: automatic scoring and assessment, learning error detection and diagnosis, spoken dialogue for tutoring system, speech and language technologies for education, etc. to submit papers. This special session will be a forum to present new R/D results which can support interactive language learning applications between human and machine.

SUB-SAHARAN AFRICAN LANGUAGES: FROM SPEECH FUNDAMENTALS TO APPLICATIONS (Area 10: Speech Recognition—Technologies and Systems for New Applications)

Monday, 12 September | 13:30 – 15:30 | Grand Ballroom BC

Names and affiliation of organizers:

- Martine Adda-Decker, *CNRS – LPP and LIMSI, France*
- Laurent Besacier, *University Grenoble-Alpes - LIG laboratory, France*
- Marelie Davel, *North-West University, Vanderbijlpark, South Africa*
- Larry Hyman, *Department of Linguistics, University of California, Berkeley, USA*
- Martin Jansche, *Google, London, UK*
- Francois Pellegrino, *CNRS – DDL Lyon, France*
- Olivier Rosec, *Voxygen SAS,- Pleumeur-Bodou, France*
- Sebastian Stüker, *Karlsruhe Institute of Technology (KIT), Germany*
- Martha Tachbelie Yifiru, *School of Information Science, Addis Ababa University, Ethiopia*

Description:

This special session aims at gathering researchers in speech technology and researchers in linguistics (working in language documentation and fundamentals of speech science). Such a partnership is particularly important for Sub-Saharan African languages which tend to remain under-resourced, under-documented and often also un-written.

Friday, 9 September

Registration	08:00 - 19:30	Grand Ballroom Foyer
Speaker Check-In	07:30 - 17:00	Regency AB

Opening Session
08:30 - 09:30
Grand Ballroom ABC

Keynote 1: ISCA Medalist: John Makhoul
09:30 - 10:30
Grand Ballroom ABC

Refreshment Break
10:30 - 11:00 | Pacific Concourse

CONCURRENT SESSIONS | 11:00 - 13:00

Grand Ballroom A	Grand Ballroom BC	Bayview Room A	Bayview Room B	Seacliff BCD	Seacliff A	Pacific Concourse	Market Street Foyer
O-1-1 Neural Networks in Speech Recognition	O-1-2 Special Session: Auditory-Visual Expressive Speech and Gesture in Humans and Machines	O-1-3 Prosody	O-1-4 Speech and Language Processing for Clinical Health Applications	O-1-5 Speech Coding and Audio Processing for Noise Reduction	O-1-6 Speech Analysis	Posters 1-1 to 1-4	Show & Tell 1

Lunch Break
13:00 - 14:30

CONCURRENT SESSIONS | 14:30 - 16:30

Grand Ballroom A	Grand Ballroom BC	Bayview Room A	Bayview Room B	Seacliff BCD	Seacliff A	Pacific Concourse	Market Street Foyer
O-2-1 New Trends in Neural Networks for Speech Recognition	O-2-2 Special Session: The RedDots Challenge: Towards Characterizing Speakers from Short Utterances	O-2-3 Articulatory Measurements and Analysis	O-2-4 Automatic Assessment of Emotions	O-2-5 Acoustic and Articulatory Phonetics	O-2-6 Source Separation and Spatial Audio	Posters 2-1 to 2-4	Show & Tell 2

Refreshment Break
16:30 - 17:00 | Pacific Concourse

CONCURRENT SESSIONS | 17:00 - 19:00

Grand Ballroom A	Grand Ballroom BC	Bayview Room A	Bayview Room B	Seacliff BCD	Seacliff A	Pacific Concourse	Market Street Foyer
O-3-1 Feature Extraction and Acoustic Modeling Using Neural Networks for ASR	O-3-2 Special Session: The Speakers in the Wild (SITW) Speaker Recognition Challenge	O-3-3 Non-Native Speech Perception	O-3-4 Behavioral Signal Processing and Speaker State and Traits Analytics	O-3-5 Spoken Term Detection	O-3-6 Co-Inference of Production and Acoustics	Posters 3-1 to 3-4	Show & Tell 3

Welcome Reception
19:00 - 21:00
Hyatt Regency San Francisco Atrium Lounge

Saturday, 10 September

Registration	08:00 - 18:00	Grand Ballroom Foyer
Speaker Check-In	07:30 - 17:00	Regency AB

Special Event: Mindfulness
08:00 - 08:30
Grand Ballroom ABC

Keynote 2: Edward Chang
08:30 - 09:30
Grand Ballroom ABC

Refreshment Break
09:30 - 10:00 | Pacific Concourse

CONCURRENT SESSIONS | 10:00 - 12:00

Grand Ballroom A	Grand Ballroom BC	Bayview Room A	Bayview Room B	Seacliff BCD	Seacliff A	Pacific Concourse	Market Street Foyer
Special Event: Speaker Comparison for Forensic and Investigative Applications II	O-4-2 Special Session: Clinical and /Neuroscience-Inspired Vocal Biomarkers of Neurological and Psychiatric Disorders	O-4-3 Special Session: Singing Synthesis Challenge: Fill-In the Gap	O-4-4 Conversation and Interaction	O-4-5 Automatic Learning of Representations	O-4-6 Language Modeling for Conversational Speech and Confidence Measures	Posters 4-1 to 4-4	Show & Tell 4

Lunch Break
12:00 - 13:30

CONCURRENT SESSIONS | 13:30 - 15:30

Grand Ballroom A	Grand Ballroom BC	Bayview Room A	Bayview Room B	Seacliff BCD	Seacliff A	Pacific Concourse	Market Street Foyer
O-5-1 Acoustic Model Adaptation	O-5-2 Special Session: Sharing Research and Education Resources for Understanding Speech Processing	O-5-3 Special Session: Voice Conversion Challenge	O-5-4 Intelligibility and Masking	O-5-5 Robust Speaker Recognition and Anti-Spoofing	O-5-6 Speech Enhancement and Applications	Posters 5-1 to 5-4	Show & Tell 5

Refreshment Break
15:30 - 16:00 | Pacific Concourse

ISCA General Assembly
16:00 - 17:30
Grand Ballroom A

Reviewer's Reception
18:30 - 20:00
The City Club

Student Reception
19:00 - 21:00
Jones

Sunday, 11 September

Registration	08:00 - 18:00	Grand Ballroom Foyer
Speaker Check-In	07:30 - 17:00	Regency AB

Keynote 3: Anne Fernald
08:30 - 09:30
Grand Ballroom ABC

Refreshment Break
09:30 - 10:00 | Pacific Concourse

CONCURRENT SESSIONS | 10:00 - 12:00

Grand Ballroom A	Grand Ballroom BC	Bayview Room A	Bayview Room B	Seacliff BCD	Seacliff A	Pacific Concourse	Market Street Foyer
O-6-1 Far-Field Speech Processing	O-6-2 Special Session: Interspeech 2016 Computational Paralinguistics Challenge (ComParE): Deception, Sincerity & Native Language	O-6-3 Special Session: Speech, Audio, and Language Processing Techniques Applied to Bird and Animal Vocalizations	O-6-4 Dialogue Systems and Analysis of Dialogue	O-6-5 Interaction between Speech Production and Perception	O-6-6 Multimodal Processing	Posters 6-1 to 6-4	Show & Tell 6

Lunch Break
12:00 - 13:30

CONCURRENT SESSIONS | 13:30 - 15:30

Grand Ballroom A	Grand Ballroom BC	Bayview Room A	Bayview Room B	Seacliff BCD	Seacliff A	Pacific Concourse
O-7-1 Robustness in Speech Processing	O-7-2 Special Session: Interspeech 2016 Computational Paralinguistics Challenge (ComParE): Deception, Sincerity & Native Language	O-7-3 Acoustic and Articulatory Phonetics	O-7-4 Speech Synthesis Oral I: Neural Networks	O-7-5 Speech Quality & Intelligibility	O-7-6 Speech Translation and Metadata for Linguistic/Discourse Structure	Posters 7-1 to 7-4

Refreshment Break
15:30 - 16:00 | Pacific Concourse

CONCURRENT SESSIONS | 16:00 - 18:00

Grand Ballroom A	Grand Ballroom BC	Bayview Room A	Bayview Room B	Seacliff BCD	Seacliff A	Pacific Concourse
O-8-1 Topics in Speech Recognition	O-8-2 Special Session: Realism in Robust Speech Processing	O-8-3 Spoken Word Recognition	O-8-4 Speech Synthesis Oral: High Level Linguistic Features	O-8-5 Speech Enhancement	O-8-6 Dialogue: Backchannels and Turntaking	Posters 8-1 to 8-4

Banquet
19:00 - 22:00
California Academy of Sciences
(Advance purchase required)

18:30
Buses depart from Hyatt Regency San Francisco

Monday, 12 September

Registration	08:00-17:30	Grand Ballroom Foyer
Speaker Check-In	07:30-13:30	Regency AB

Keynote 4: Dan Jurafsky
08:30 - 09:30
Grand Ballroom ABC

Refreshment Break
09:30 - 10:00 | Pacific Concourse

CONCURRENT SESSIONS | 10:00 - 12:00

Grand Ballroom A	Grand Ballroom BC	Bayview Room A	Bayview Room B	Seacliff BCD	Seacliff A	Pacific Concourse
Special Event: Speech Ventures	O-9-2 Special Session: Speech and Language Technologies for Human-Machine Conversation-Based Language Education	O-9-3 Phonation and Voice Quality	O-9-4 Speech Synthesis Oral: Prosody and Expressive Speech	O-9-5 Language Recognition	O-9-6 Spoken Language Understanding Systems	Posters 9-1 to 9-4

Grand Ballroom A	
Special Event: Computational Approaches to Linguistic Code Switching 12:15 - 13:00	Lunch Break 12:00 - 13:30

CONCURRENT SESSIONS | 13:30 - 15:30

Grand Ballroom A	Grand Ballroom BC	Bayview Room A	Bayview Room B	Seacliff BCD	Seacliff A	Pacific Concourse
O-10-1 Neural Networks for Language Modeling	O-10-2 Special Session: Sub-Saharan African Languages: From Speech Fundamentals to Applications	O-10-3 Speech Production Models	O-10-4 Speaker States and Traits	O-10-5 Speaker Recognition	O-10-6 VAD and Audio Events	Posters 10-1 to 10-4

Refreshment Break
15:30 - 16:00 | Pacific Concourse

Closing Session
16:00 - 17:00
Grand Ballroom ABC

Session Index

Friday, 9 September 2016

Saturday, 10 September 2016

Sunday, 11 September 2016

Monday, 12 September 2016

Abstracts

Keynote 1: ISCA Medalist: John Makhoul

Grand Ballroom ABC, 09:30–10:30, Friday, 9 Sept. 2016
Chair: Haizhou Li

A 50-Year Retrospective on Speech and Language Processing

John Makhoul; BBN Technologies, USA

Fri-Keynote-1, Time: 09:30

This talk is a retrospective of speech and language processing as witnessed by the speaker during the last 50 years. From exploratory scientific beginnings that emphasized the discovery of how speech is produced and perceived by humans to today's plethora of applications using our technology, our field has witnessed explosive growth. The talk will review the historical development of our community and some of the key technical ideas that have shaped our field. Some of the ideas were influenced by developments in other fields, while some of the developments in our field have been instrumental in key advances in other fields, such as optical character recognition and machine translation. Important developments include the source-filter model, digital signal processing, linear prediction, vector quantization, deep neural networks, and statistical modeling methods, especially hidden Markov models (HMMs), with primary applications to speech analysis, synthesis, coding, and recognition. The talk will be sprinkled with lessons learned in the importance of various factors in performing our research, and will be peppered with interesting tidbits about key moments in the development of our technology. The talk will end with a brief prospective peek at the next 50 years.

Fri-O-1-1 : Neural Networks in Speech Recognition

Grand Ballroom A, 11:00–13:00, Friday, 9 Sept. 2016
Chairs: Penny Karanasou, Bhuvana Ramabhadran

Improving English Conversational Telephone Speech Recognition

Ivan Medennikov [1], Alexey Prudnikov [2], Alexander Zatvornitskiy [1]; [1] STC-innovations, Russia; [2] ITMO University, Russia

Fri-O-1-1-1, Time: 11:00

The goal of this work is to build a state-of-the-art English conversational telephone speech recognition system. We investigated several techniques to improve acoustic modeling, namely speaker-dependent bottleneck features, deep Bidirectional Long Short-Term Memory (BLSTM) recurrent neural networks, data augmentation and score fusion of DNN and BLSTM models. Training set consisted of the 300 hour Switchboard English speech corpus. We also examined the hypothesis rescoring using language models based on recurrent neural networks. The resulting system achieves a word error rate of 7.8% on the Switchboard part of the HUB5 2000 evaluation set which is the competitive result.

The IBM 2016 English Conversational Telephone Speech Recognition System

George Saon, Tom Sercu, Steven Rennie, Hong-Kwang J. Kuo; IBM, USA

Fri-O-1-1-2, Time: 11:20

We describe a collection of acoustic and language modeling techniques that lowered the word error rate of our English conversational telephone LVCSR system to a record 6.6% on the Switchboard subset of the Hub5 2000 evaluation testset. On the acoustic side, we use a score fusion of three strong models: recurrent nets with maxout activations, very deep convolutional nets with 3×3 kernels, and bidirectional long short-term memory nets which operate on FMLLR and i-vector features. On the language modeling side, we use an updated model "M" and hierarchical neural network LMs.

Small-Footprint Deep Neural Networks with Highway Connections for Speech Recognition

Liang Lu, Steve Renals; University of Edinburgh, UK

Fri-O-1-1-3, Time: 11:40

For speech recognition, deep neural networks (DNNs) have significantly improved the recognition accuracy in most of benchmark datasets and application domains. However, compared to the conventional Gaussian mixture models, DNN-based acoustic models usually have much larger number of model parameters, making it challenging for their applications in resource constrained platforms, e.g., mobile devices. In this paper, we study the application of the recently proposed highway network to train small-footprint DNNs, which are *thinner* and *deeper*, and have significantly smaller number of model parameters compared to conventional DNNs. We investigated this approach on the AMI meeting speech transcription corpus which has around 80 hours of audio data. The highway neural networks constantly outperformed their plain DNN counterparts, and the number of model parameters can be reduced significantly without sacrificing the recognition accuracy.

Deep Convolutional Neural Networks with Layer-Wise Context Expansion and Attention

Dong Yu [1], Wayne Xiong [1], Jasha Droppo [1], Andreas Stolcke [1], Guoli Ye [2], Jinyu Li [1], Geoffrey Zweig [1]; [1] Microsoft, USA; [2] Microsoft, China

Fri-O-1-1-4, Time: 12:00

In this paper, we propose a deep convolutional neural network (CNN) with layer-wise context expansion and location-based attention, for large vocabulary speech recognition. In our model each higher layer uses information from broader contexts, along both the time and frequency dimensions, than its immediate lower layer. We show that both the layer-wise context expansion and the location-based attention can be implemented using the element-wise matrix product and the convolution operation. For this reason, contrary to other CNNs, no pooling operation is used in our model. Experiments on the 309hr Switchboard task and the 375hr short message dictation task indicates that our model outperforms both the DNN and LSTM significantly.

Lower Frame Rate Neural Network Acoustic Models

Golan Pundak, Tara N. Sainath; Google, USA

Fri-O-1-1-5, Time: 12:20

Recently neural network acoustic models trained with Connectionist Temporal Classification (CTC) were proposed as an alternative

NOTES

approach to conventional cross-entropy trained neural network acoustic models which output frame-level decisions every 10ms [1]. As opposed to conventional models, CTC learns an alignment jointly with the acoustic model, and outputs a *blank* symbol in addition to the regular acoustic state units. This allows the CTC model to run with a lower frame rate, outputting decisions every 30ms rather than 10ms as in conventional models, thus improving overall system speed. In this work, we explore how conventional models behave with lower frame rates. On a large vocabulary Voice Search task, we will show that with conventional models, we can slow the frame rate to 40ms while improving WER by 3% relative over a CTC-based model.

Improved Neural Network Initialization by Grouping Context-Dependent Targets for Acoustic Modeling

Gakuto Kurata[1], Brian Kingsbury[2]; [1]IBM, Japan; [2]IBM, USA

Fri-O-1-1-6, Time: 12:40

Neural Network (NN) Acoustic Models (AMs) are usually trained using context-dependent Hidden Markov Model (CD-HMM) states as independent targets. For example, the CD-HMM states of `A-b-2` (second variant of beginning state of `A`) and `A-m-1` (first variant of middle state of `A`) both correspond to the phone `A`, and `A-b-1` and `A-b-2` both correspond to the Context-independent HMM (CI-HMM) state `A-b`, but this relationship is not explicitly modeled. We propose a method that treats some neurons in the final hidden layer just below the output layer as dedicated neurons for phones or CI-HMM states by initializing connections between the dedicated neurons and the corresponding CD-HMM outputs with stronger weights than to other outputs. We obtained 6.5% and 3.6% relative error reductions with a DNN AM and a CNN AM, respectively, on a 50-hour English broadcast news task and 4.6% reduction with a CNN AM on a 500-hour Japanese task, in all cases after Hessian-free sequence training. Our proposed method only changes the NN parameter initialization and requires no additional computation in NN training or speech recognition run-time.

Fri-O-1-2 : Special Session: Auditory-Visual Expressive Speech and Gesture in Humans and Machines

Grand Ballroom BC, 11:00–13:00, Friday, 9 Sept. 2016
Chairs: Jeesun Kim, Gérard Bailly

Automatic Scoring of Monologue Video Interviews Using Multimodal Cues

Lei Chen, Gary Feng, Michelle Martin-Raugh, Chee Wee Leong, Christopher Kitchen, Su-Youn Yoon, Blair Lehman, Harrison Kell, Chong Min Lee; Educational Testing Service, USA

Fri-O-1-2-1, Time: 11:00

Job interviews are an important tool for employee selection. When making hiring decisions, a variety of information from interviewees, such as previous work experience, skills, and their verbal and nonverbal communication, are jointly considered. In recent years, Social Signal Processing (SSP), an emerging research area on enabling computers to sense and understand human social signals, is being used develop systems for the coaching and evaluation of job interview performance. However this research area is still in its infancy and lacks essential resources (e.g., adequate corpora). In this paper, we report on our efforts to create an automatic interview rating system for monologue-style video interviews, which have been widely used in today's job hiring market. We created the first multimodal corpus for such video interviews. Additionally, we conducted manual rating on the interviewee's personality and performance during 12 structured interview questions measuring different types of job-related skills. Finally, focusing on predicting overall interview performance, we explored a set of verbal and nonverbal features and several machine learning models. We found that using both verbal and nonverbal features provides more accurate predictions. Our initial results suggest that it is feasible to continue working in this newly formed area.

The Sound of Disgust: How Facial Expression May Influence Speech Production

Chee Seng Chong, Jeesun Kim, Chris Davis; Western Sydney University, Australia

Fri-O-1-2-2, Time: 11:15

In speech articulation, mouth/lip shapes determine properties of the front part of the vocal tract, and so alter vowel formant frequencies. Mouth and lip shapes also determine facial emotional expressions, e.g., disgust is typically expressed with a distinctive lip and mouth configuration (i.e., closed mouth, pulled back lip corners). This overlap of speech and emotion gestures suggests that expressive speech will have different vowel formant frequencies from neutral speech. This study tested this hypothesis by comparing vowels produced in neutral versus disgust expressions. We used our database of five female native Cantonese talkers each uttering 50 CHINT sentences in both a neutral tone of voice and in disgust to examine five vowels ([ɐ], [ɛː], [iː], [ɔː], [ʊː]). Mean fundamental frequency (F0) and the first two formants (F1 and F2) were calculated and analysed using mixed effects logistic regression. The results showed that the disgust vowels showed a significant reduction in either or both formant values (depending on vowel type) compared to neutral. We discuss the results in terms of how vowel synthesis could be used to alter the recognition of the sound of disgust.

Analyzing Temporal Dynamics of Dyadic Synchrony in Affective Interactions

Zhaojun Yang, Shrikanth S. Narayanan; University of Southern California, USA

Fri-O-1-2-3, Time: 11:30

Human communication is a dynamical and interactive process that naturally induces an active flow of interpersonal coordination, and synchrony, along various behavioral dimensions. Assessing and characterizing the temporal dynamics of synchrony during an interaction is essential for fully understanding the human communication mechanisms. In this work, we focus on uncovering the temporal variability patterns of synchrony in visual gesture and vocal behavior in affectively rich interactions. We propose a statistical scheme to robustly quantify the turn-wise interpersonal synchrony. The analysis of the synchrony dynamics measure relies heavily on functional data analysis techniques. Our analysis results reveal that: 1) the dynamical patterns of interpersonal synchrony differ depending on the global emotions of an interaction dyad; 2) there generally exists a tight dynamical emotion-synchrony coupling over the interaction. These observations corroborate that interpersonal behavioral synchrony is a critical manifestation of

NOTES

the underlying affective processes, shedding light toward improved affective interaction modeling and automatic emotion recognition.

Audiovisual Speech Scene Analysis in the Context of Competing Sources

Attigodu C. Ganesh, Frédéric Berthommier, Jean-Luc Schwartz; GIPSA, France

Fri-O-1-2-4, Time: 11:45

Audiovisual fusion in speech perception is generally conceived as a process independent from scene analysis, which is supposed to occur separately in the auditory and visual domain. On the contrary, we have been proposing in the last years that scene analysis such as what takes place in the cocktail party effect was an audiovisual process. We review here a series of experiments illustrating how audiovisual speech scene analysis occurs in the context of competing sources. Indeed, we show that a short contextual audiovisual stimulus made of competing auditory and visual sources modifies the perception of a following McGurk target. We interpret this in terms of binding, unbinding and rebinding processes, and we show how these processes depend on audiovisual correlations in time, attentional processes and differences between junior and senior participants.

Head Motion Generation with Synthetic Speech: A Data Driven Approach

Najmeh Sadoughi, Carlos Busso; University of Texas at Dallas, USA

Fri-O-1-2-5, Time: 12:00

To have believable head movements for *conversational agents* (CAs), the natural coupling between speech and head movements needs to be preserved, even when the CA uses synthetic speech. To incorporate the relation between speech head movements, studies have learned these couplings from real recordings, where speech is used to derive head movements. However, relying on recorded speech for every sentence that a virtual agent utters constrains the versatility and scalability of the interface, so most practical solutions for CAs use text to speech. While we can generate head motion using rule-based models, the head movements may become repetitive, spanning only a limited range of behaviors. This paper proposes strategies to leverage speech-driven models for head motion generation for cases relying on synthetic speech. The straightforward approach is to drive the speech-based models using synthetic speech, which creates mismatch between the test and train conditions. Instead, we propose to create a parallel corpus of synthetic speech aligned with natural recordings for which we have motion capture recordings. We use this parallel corpus to either retrain or adapt the speech-based models with synthetic speech. Objective and subjective metrics show significant improvements of the proposed approaches over the case with mismatched condition.

The Consistency and Stability of Acoustic and Visual Cues for Different Prosodic Attitudes

Jeesun Kim, Chris Davis; Western Sydney University, Australia

Fri-O-1-2-6, Time: 12:15

Recently it has been argued that speakers use conventionalized forms to express different prosodic attitudes [1]. We examined this by looking at across speaker consistency in the expression of auditory and visual (head and face motion) prosodic attitudes produced on multiple different occasions. Specifically, we examined acoustic and motion profiles of a female and a male speaker expressing six different prosodic attitudes for four within-session repetitions across four different sessions. We used the same acoustic features as [1] and visual prosody was assessed by examining patterns of speaker's mouth, eyebrow and head movements. There was considerable variation in how prosody was realized across speakers, with the productions of one speaker more discriminable than the other. Within-session variation for both the acoustic and movement data was smaller than across-session variation, suggesting that short-term memory plays a role in consistency. The expression of some attitudes was less variable than others and better discrimination was found with the acoustic compared to the visual data, although certain visual features (e.g., eyebrow brow motion) provided better discrimination than others.

Introduction to Poster Presentation of Part II

Jeesun Kim[1], Gérard Bailly[2]; [1]Western Sydney University, Australia; [2]GIPSA, France

Fri-O-1-2-7, Time: 12:30

(No abstract available at the time of publication)

Fri-O-1-3 : Prosody

Bayview A, 11:00–13:00, Friday, 9 Sept. 2016

Chairs: Chiu-Yu Tseng, Štefan Beňuš

The Unit of Speech Encoding: The Case of Romanian

Irene Vogel[1], Laura Spinu[2]; [1]University of Delaware, USA; [2]University of Western Ontario, Canada

Fri-O-1-3-1, Time: 11:00

The number of units in an utterance determines how much time speakers require to physically plan and begin their production [1]–[2]. Previous research proposed that the crucial units are prosodic i.e., Phonological Words (PWs), not syntactic or morphological [3]. Experiments on Dutch using a prepared speech paradigm claimed to support this view [4]–[5]; however, compounds did not conform to predictions and required the introduction of a different way of counting units. Since two PWs in compounds patterned with one PW, with or without clitics, rather than a phrase containing two PWs, a recursive PW' was invoked. Similar results emerged using the same methodology with compounds in Italian [6], and it was thus proposed that the relevant unit for speech encoding is not the PW, but rather the Composite Group (CompG), a constituent of the Prosodic Hierarchy between the PW and Phonological Phrase that comprises both compounds and clitic constructions [7]. We further investigate the relevant unit for speech encoding using the same methodology in Romanian. Similar findings support the CompG as the speech planning unit since, again, compounds with two PWs pattern with single words and clitic constructions, not Phonological Phrases which also contain two PWs.

Notes

The Perceptual Effect of L1 Prosody Transplantation on L2 Speech: The Case of French Accented German

Jeanin Jügler, Frank Zimmerer, Jürgen Trouvain, Bernd Möbius; Universität des Saarlandes, Germany

Fri-O-1-3-2, Time: 11:20

Research has shown that language learners are not only challenged by segmental differences between their native language (L1) and the second language (L2). They also have problems with the correct production of suprasegmental structures, like phone/syllable duration and the realization of pitch. These difficulties often lead to a perceptible foreign accent. This study investigates the influence of prosody transplantation on foreign accent ratings. Syllable duration and pitch contour were transferred from utterances of a male and female German native speaker to utterances of ten French native speakers speaking German. Acoustic measurements show that French learners spoke with a significantly lower speaking rate. As expected, results of a perception experiment judging the accentedness of 1) German native utterances, 2) unmanipulated and 3) manipulated utterances of French learners of German suggest that the transplantation of the prosodic features syllable duration and pitch leads to a decrease in accentedness rating. These findings confirm results found in similar studies investigating prosody transplantation with different L1 and L2 and provide a beneficial technique for (computer-assisted) pronunciation training.

Organizing Syllables into Sandhi Domains — Evidence from F0 and Duration Patterns in Shanghai Chinese

Bijun Ling, Jie Liang; Tongji University, China

Fri-O-1-3-3, Time: 11:40

In this study we investigated grouping-related F0 patterns in Shanghai Chinese by examining the effect of syllable position in a sandhi domain while controlling for tone, number of syllables in a domain, and focus condition. Results showed that F0 alignment had the most consistent grouping-related patterns, and syllable duration was positively related to F0 movement. Focus and word length both increased F0 peak and F0 excursion, but they had opposite influence on F0 slope, which indicated that focus and word length had different mechanisms in affecting F0 implementation, as focus increased articulation strength while word length influenced speaker's pre-planning.

Automatic Analysis of Phonetic Speech Style Dimensions

Neville Ryant, Mark Liberman; Linguistic Data Consortium, USA

Fri-O-1-3-4, Time: 12:00

We apply automated analysis methods to create a multidimensional characterization of the prosodic characteristics of a large variety of speech datasets, with the goal of developing a general framework for comparing prosodic styles. Our datasets span styles including conversation, fluent reading, extemporized narratives, political speech, and advertisements; we compare several different languages including English, Spanish, and Chinese; and the features we extract are based on the joint distributions of F0 and amplitude values and sequences, speech and silence segment durations, syllable durations, and modulation spectra. Rather than focus on the acoustic correlates of a small number of discrete and mutually exclusive categories, we aim to characterize the space in which diverse speech styles live.

The Acoustic Manifestation of Prominence in Stressless Languages

Angeliki Athanasopoulou, Irene Vogel; University of Delaware, USA

Fri-O-1-3-5, Time: 12:20

Languages frequently express focus by enhancing various acoustic attributes of an utterance, but it is widely accepted that the main enhancement appears on stressed syllables. In languages without lexical stress, the question arises as to how focus is acoustically manifested. We thus examine the acoustic properties associated with prominence in three stressless languages, Indonesian, Korean and Vietnamese, comparing real three-syllable words in non-focused and focused contexts. Despite other prosodic differences, our findings confirm that none of the languages exhibits stress in the absence of focus, and under focus, no syllable shows consistent enhancement that could be indirectly interpreted as a manifestation of focus. Instead, a combination of boundary phenomena consistent with the right edge of a major prosodic constituent (Intonational Phrase) appears in each language: increased duration on the final syllable and in Indonesian and Korean, a decrease in F0. Since these properties are also found in languages with stress, we suggest that boundary phenomena signaling a major prosodic constituent break are used universally to indicate focus, regardless of a language's word-prosody; stress languages may use the same boundary properties, but these are most likely to be combined with enhancement of the stressed syllable of a word.

The Rhythmic Constraint on Prosodic Boundaries in Mandarin Chinese Based on Corpora of Silent Reading and Speech Perception

Wei Lai[1], Jiahong Yuan[1], Ya Li[2], Xiaoying Xu[3], Mark Liberman[1]; [1]University of Pennsylvania, USA; [2]Chinese Academy of Sciences, China; [3]Beijing Normal University, China

Fri-O-1-3-6, Time: 12:40

This study investigated the interaction between rhythmic and syntactic constraints on prosodic phrases in Mandarin Chinese. A set of 4000 sentences was annotated twice, once based on silent reading by 130 students assigned 500 sentences each, and a second time by speech perception based on a recording by one professional speaker. In both types of annotation, the general pattern of phrasing was consistent, with short "rhythmic phrases" behaving differently from longer "intonational phrases". The probability of a rhythmic-phrase boundary between two words increased with the total length of those two words, and was also influenced by the nature of the syntactic boundary between them. The resulting rhythmic phrases were mainly 2–5 syllables long, independent of the length of the sentence. In contrast, the length of intonational phrases was not stable, and was heavily affected by sentence length. Intonational-phrase boundaries were also found to be affected by higher-level syntactic features, such as the depth of syntactic tree and the number of IP nodes. However, these syntactic influences on intonational phrases were weakened in long sentences (>20 syllable) and also in short sentences (<10 syllable), where the length effect played the main role.

Notes

Fri-O-1-4 : Speech and Language Processing for Clinical Health Applications

Bayview B, 11:00–13:00, Friday, 9 Sept. 2016
Chairs: Julien Epps, Elmar Nöth

Toward Development and Evaluation of Pain Level-Rating Scale for Emergency Triage based on Vocal Characteristics and Facial Expressions

Fu-Sheng Tsai[1], Ya-Ling Hsu[1], Wei-Chen Chen[1], Yi-Ming Weng[2], Chip-Jin Ng[2], Chi-Chun Lee[1]; [1]National Tsing Hua University, Taiwan; [2]Chang Gung Memorial Hospital, Taiwan

Fri-O-1-4-1, Time: 11:00

In order to allocate the healthcare resource, triage classification system plays an important role in assessing the severity of illness of the boarding patient at emergency department. The self-report pain intensity numerical-rating scale (NRS) is one of the major modifiers of the current triage system based on the Taiwan Triage and Acuity Scale (TTAS). The validity and reliability of *self-report* scheme for pain level assessment is a major concern. In this study, we model the observed expressive behaviors, i.e., facial expressions and vocal characteristics, directly from audio-video recordings in order to measure pain level for patients during triage. This work demonstrates a feasible model, which achieves an accuracy of 72.3% and 51.6% in a binary and ternary pain intensity classification. Moreover, the study result reveals a significant association of current model and analgesic prescription/patient disposition after adjusted for patient-report NRS and triage vital signs.

Predicting Severity of Voice Disorder from DNN-HMM Acoustic Posteriors

Tan Lee, Yuanyuan Liu, Yu Ting Yeung, Thomas K.T. Law, Kathy Y.S. Lee; Chinese University of Hong Kong, China

Fri-O-1-4-2, Time: 11:20

Acoustical analysis of speech is considered a favorable and promising approach to objective assessment of voice disorders. Previous research emphasized on the extraction and classification of voice quality features from sustained vowel sounds. In this paper, an investigation on voice assessment using continuous speech utterances of Cantonese is presented. A DNN-HMM based speech recognition system is trained with speech data of unimpaired voice. The recognition accuracy for pathological utterances is found to decrease significantly with the disorder severity increasing. Average acoustic posterior probabilities are computed for individual phones from the speech recognition output lattices and the DNN soft-max layer. The phone posteriors obtained for continuous speech from the mild, moderate and severe categories are highly distinctive and thus useful to the determination of voice disorder severity. A subset of Cantonese phonemes are identified to be suitable and reliable for voice assessment with continuous speech.

Long-Term Stability of Tracheoesophageal Voices

Klaske E. van Sluis, Michiel W.M. van den Brekel, Frans J.M. Hilgers, Rob J.J.H. van Son; Netherlands Cancer Institute, The Netherlands

Fri-O-1-4-3, Time: 11:40

Long-term voice outcomes of 13 tracheoesophageal speakers are assessed using speech samples that were recorded with at least 7 years in between. Intelligibility and voice quality are perceptually evaluated by 10 experienced speech and language pathologists. In addition, automatic speech evaluations are performed with tools from Ghent University. No significant group effect was found for changes in voice quality and intelligibility. The recordings showed a wide interspeaker variability. It is concluded that intelligibility and voice quality of tracheoesophageal voice is mostly stable over a period of 7 to 18 years.

Detecting Mild Cognitive Impairment from Spontaneous Speech by Correlation-Based Phonetic Feature Selection

Gábor Gosztolya[1], László Tóth[1], Tamás Grósz[2], Veronika Vincze[1], Ildikó Hoffmann[2], Gréta Szatlóczki[2], Magdolna Pákáski[2], János Kálmán[2]; [1]MTA-SZTE RGAI, Hungary; [2]University of Szeged, Hungary

Fri-O-1-4-4, Time: 12:00

Mild Cognitive Impairment (MCI), sometimes regarded as a prodromal stage of Alzheimer's disease, is a mental disorder that is difficult to diagnose. Recent studies reported that MCI causes slight changes in the speech of the patient. Our previous studies showed that MCI can be efficiently classified by machine learning methods such as Support-Vector Machines and Random Forest, using features describing the amount of pause in the spontaneous speech of the subject. Furthermore, as hesitation is the most important indicator of MCI, we took special care when handling filled pauses, which usually correspond to hesitation. In contrast to our previous studies which employed manually constructed feature sets, we now employ (automatic) correlation-based feature selection methods to find the relevant feature subset for MCI classification. By analyzing the selected feature subsets we also show that features related to filled pauses are useful for MCI detection from speech samples.

Towards an Automated Screening Tool for Developmental Speech and Language Impairments

Jen J. Gong[1], Maryann Gong[1], Dina Levy-Lambert[1], Jordan R. Green[2], Tiffany P. Hogan[2], John V. Guttag[1]; [1]MIT, USA; [2]MGH Institute of Health Professions, USA

Fri-O-1-4-5, Time: 12:20

Approximately 60% of children with speech and language impairments do not receive the intervention they need because their impairment was missed by parents and professionals who lack specialized training. Diagnoses of these disorders require a time-intensive battery of assessments, and these are often only administered after parents, doctors, or teachers show concern.

An automated test could enable more widespread screening for speech and language impairments. To build classification models to distinguish children with speech or language impairments from typically developing children, we use acoustic features describing

Notes

speech and pause events in story retell tasks. We developed and evaluated our method using two datasets. The smaller dataset contains many children with severe speech or language impairments and few typically developing children. The larger dataset contains primarily typically developing children. In three out of five classification tasks, even after accounting for age, gender, and dataset differences, our models achieve good discrimination performance (AUC > 0.70).

Spectral Enhancement of Cleft Lip and Palate Speech

Vikram C.M., Nagaraj Adiga, S.R. Mahadeva Prasanna; IIT Guwahati, India

Fri-O-1-4-6, Time: 12:40

The quality of cleft lip and palate (CLP) speech is affected due to hyper-nasality and mis-articulation. Surgery and speech therapy are required to correct the structural and functional defects of CLP, which will result in an enhanced speech signal. The quality of the enhanced speech is perceptually evaluated by speech-language pathologists and results are highly biased. In this work, a signal processing based two stage speech enhancement method is proposed to get the perceptual benchmark to compare the signal after the surgery / therapy. In the first stage, CLP speech is enhanced by suppressing the nasal formant and in the second stage, spectral peak-valley enhancement is carried out to reduce the hyper-nasality associated with the CLP speech. The evaluation results show that the perceptual quality of CLP speech signal is improved after enhancement in both stages. Further, the improvement in the quality of the enhanced signal is compared with the speech signal after palatal prosthesis / surgery. The perceptual evaluation results show that the enhanced speech signals are better than the speech after prosthesis / surgery

Fri-O-1-5 : Speech Coding and Audio Processing for Noise Reduction

Seacliff BCD, 11:00–13:00, Friday, 9 Sept. 2016
Chairs: Tom Bäckström, Dayana Ribas

Assessing Level-Dependent Segmental Contribution to the Intelligibility of Speech Processed by Single-Channel Noise-Suppression Algorithms

Tian Guan[1], Guangxing Chu[1], Fei Chen[2], Feng Yang[3]; [1]Tsinghua University, China; [2]SUSTC, China; [3]Shenzhen Children's Hospital, China

Fri-O-1-5-1, Time: 11:00

Most existing single-channel noise-suppression algorithms cannot improve speech intelligibility for normal-hearing listeners; however, the underlying reason for this performance deficit is still unclear. Given that various speech segments contain different perceptual contributions, the present work assesses whether the intelligibility of noisy speech can be improved when selectively suppressing its noise at high-level (vowel-dominated) or middle-level (containing vowel-consonant transitions) segments by existing single-channel noise-suppression algorithms. The speech signal was corrupted by speech-spectrum shaped noise and two-talker babble masker, and its noisy high- or middle-level segments were replaced by their noise-suppressed versions processed by four types of existing single-channel noise-suppression algorithms. Experimental results showed that performing segmental noise-suppression at high- or middle-level led to decreased intelligibility relative to noisy speech. This suggests that the lack of intelligibility improvement by existing noise-suppression algorithms is also present at segmental level, which may account for the deficit traditionally observed at full-sentence level.

Effectiveness of Near-End Speech Enhancement Under Equal-Loudness and Equal-Level Constraints

Tudor-Cătălin Zorilă[1], Sheila Flanagan[2], Brian C.J. Moore[2], Yannis Stylianou[1]; [1]Toshiba Research Europe, UK; [2]University of Cambridge, UK

Fri-O-1-5-2, Time: 11:20

Most recently proposed near-end speech enhancement methods have been evaluated with the overall power (RMS) of the speech held constant. While significant intelligibility gains have been reported in various noisy conditions, an equal-RMS constraint may lead to enhancement solutions that increase the loudness of the original speech. Comparable effects might be produced simply by increasing the power of the original speech, which also leads to an increase in loudness. Here we suggest modifying the equal-RMS constraint to one of equal loudness between the original and the modified signals, based on a loudness model for time-varying sounds. Four state-of-the-art speech-in-noise intelligibility enhancement systems were evaluated under the equal-loudness constraint, using intelligibility tests with normal-hearing listeners. Results were compared with those obtained under the equal-RMS constraint. The methods based on spectral shaping and dynamic range compression yielded significant intelligibility gains regardless of the constraint, while for the method without dynamic range compression the intelligibility gain was lower under the equal-loudness than under the equal-RMS constraint.

Speech Synthesis in Noisy Environment by Enhancing Strength of Excitation and Formant Prominence

Bidisha Sharma, S.R. Mahadeva Prasanna; IIT Guwahati, India

Fri-O-1-5-3, Time: 11:40

Text-to-speech (TTS) synthesis systems have grown popularity due to their diverse practical usability. While most of the technologies developed aims to meet requirements in laboratory environment, the practical appliance is not limited to a specific environment. This work aims towards improving intelligibility of synthesized speech to make it deployable in realism. Based on the comparison of Lombard speech and speech produced in quiet, strength of excitation is found to play a crucial role in making speech intelligible in noisy situation. A novel method for enhancement of strength of excitation is proposed which makes the synthesized speech more intelligible in practical scenario. Linear-prediction analysis based formant enhancement method is also employed to further improve the intelligibility. The proposed enhancement framework is applied in synthesized speech and evaluated in presence of different types and levels of noise. Subjective evaluation results show that, the proposed method makes the synthesized speech applicable in practical noisy environment.

NOTES

Relative Contributions of Amplitude and Phase to the Intelligibility Advantage of Ideal Binary Masked Sentences

Lei Wang, Shufeng Zhu, Diliang Chen, Yong Feng, Fei Chen; SUSTC, China

Fri-O-1-5-4, Time: 12:00

Many studies have shown the advantage of using ideal binary masking (IdBM) to improve the intelligibility of speech corrupted by interfering maskers. Given the fact that amplitude and phase are two important acoustic cues for speech perception, the present work further investigated the relative contributions of these two cues to the intelligibility advantage of IdBM-processed sentences. Three types of Mandarin IdBM-processed stimuli (i.e., amplitude-only, phase-only, and amplitude-and-phase) were generated, and played to normal-hearing listeners to recognize. Experiment results showed that amplitude- or phase-only cue could lead to significantly improved intelligibility of IdBM-processed sentences in relative to noise-masked sentences. A masker-dependent amplitude over phase advantage was observed when accounting for their relative contributions to the intelligibility advantage of IdBM-processed sentences. Under steady-state speech-spectrum shaped noise, both amplitude- and phase-only IdBM-processed sentences contained intelligibility information close to that contained in amplitude-and-phase IdBM-processed sentences. In contrast, under competing babble masker, amplitude-only IdBM-processed sentences were more intelligible than phase-only IdBM-processed sentences, and neither could account for the intelligibility advantage of amplitude-and-phase IdBM-processed sentences.

Predicting Binaural Speech Intelligibility from Signals Estimated by a Blind Source Separation Algorithm

Qingju Liu[1], Yan Tang[2], Philip J.B. Jackson[1], Wenwu Wang[1]; [1]University of Surrey, UK; [2]University of Salford, UK

Fri-O-1-5-5, Time: 12:20

State-of-the-art binaural objective intelligibility measures (OIMs) require individual source signals for making intelligibility predictions, limiting their usability in real-time online operations. This limitation may be addressed by a blind source separation (BSS) process, which is able to extract the underlying sources from a mixture. In this study, a speech source is presented with either a stationary noise masker or a fluctuating noise masker whose azimuth varies in a horizontal plane, at two speech-to-noise ratios (SNRs). Three binaural OIMs are used to predict speech intelligibility from the signals separated by a BSS algorithm. The model predictions are compared with listeners' word identification rate in a perceptual listening experiment. The results suggest that with SNR compensation to the BSS-separated speech signal, the OIMs can maintain their predictive power for individual maskers compared to their performance measured from the direct signals. It also reveals that the errors in SNR between the estimated signals are not the only factors that decrease the predictive accuracy of the OIMs with the separated signals. Artefacts or distortions on the estimated signals caused by the BSS algorithm may also be concerns.

Automated Pause Insertion for Improved Intelligibility Under Reverberation

Petko N. Petkov, Norbert Braunschweiler, Yannis Stylianou; Toshiba Research Europe, UK

Fri-O-1-5-6, Time: 12:40

Speech intelligibility in reverberant environments is reduced because of overlap-masking. Signal modification prior to presentation in such listening environments, e.g., with a public announcement system, can be employed to alleviate this problem. Time-scale modifications are particularly effective in reducing the effect of overlap-masking. A method for introducing linguistically-motivated pauses is proposed in this paper. Given the transcription of a sentence, pause strengths are predicted at word boundaries. Pause duration is obtained by combining the pause strength and the time it takes late reverberation to decay to a level where a target signal-to-late-reverberation ratio criterion is satisfied. Considering a moderate reverberation condition and both binary and continuous pause strengths, a formal listening test was performed. The results show that the proposed methodology offers a significant intelligibility improvement over unmodified speech while continuous pause strengths offer an advantage over binary pause strengths.

Fri-O-1-6 : Speech Analysis

Seacliff A, 11:00–13:00, Friday, 9 Sept. 2016

Chair: Hakan Erdogan

Automatic Classification of Phonation Modes in Singing Voice: Towards Singing Style Characterisation and Application to Ethnomusicological Recordings

Jean-Luc Rouas, Leonidas Ioannidis; LaBRI (UMR 5800), France

Fri-O-1-6-1, Time: 11:00

This paper describes our work on automatic classification of phonation modes on singing voice. In the first part of the paper, we will briefly review the main characteristics of the different phonation modes. Then, we will describe the isolated vowels databases we used, with emphasis on a new database we recorded specifically for the purpose of this work. The next section will be dedicated to the description of the proposed set of parameters (acoustic and glottal) and the classification framework. The results obtained with only acoustic parameters are close to 80% of correct recognition, which seems sufficient for experimenting with continuous singing. Therefore, we set up two other experiments in order to see if the system may be of any practical use for singing voice characterisation. The first experiment aims at assessing if automatic detection of phonation modes may help classify singing into different styles. This experiment is carried out using a database of one singer singing the same song in 8 styles. The second experiment is carried out on field recordings from ethnomusicologists and concerns the distinction between "normal" singing and "laments" from a variety of countries.

NOTES

Novel Nonlinear Prediction Based Features for Spoofed Speech Detection

Himanshu N. Bhavsar, Tanvina B. Patel, Hemant A. Patil; DA-IICT, India

Fri-O-1-6-2, Time: 11:20

Several speech synthesis and voice conversion techniques can easily generate or manipulate speech to deceive the speaker verification (SV) systems. Hence, there is a need to develop spoofing countermeasures to detect the human speech from spoofed speech. System-based features have been known to contribute significantly to this task. In this paper, we extend a recent study of Linear Prediction (LP) and Long-Term Prediction (LTP)-based features to LP and Nonlinear Prediction (NLP)-based features. To evaluate the effectiveness of the proposed countermeasure, we use the corpora provided at the ASVspoof 2015 challenge. A Gaussian Mixture Model (GMM)-based classifier is used and the % Equal Error Rate (EER) is used as a performance measure. On the development set, it is found that LP-LTP and LP-NLP features gave an average EER of 4.78% and 9.18%, respectively. Score-level fusion of LP-LTP (and LP-NLP) with Mel Frequency Cepstral Coefficients (MFCC) gave an EER of 0.8% (and 1.37%), respectively. After score-level fusion of LP-LTP, LP-NLP and MFCC features, the EER is significantly reduced to 0.57%. The LP-LTP and LP-NLP features have found to work well even for Blizzard Challenge 2012 speech database.

Robust Vowel Landmark Detection Using Epoch-Based Features

Sri Harsha Dumpala, Bhanu Teja Nellore, Raghu Ram Nevali, Suryakanth V. Gangashetty, B. Yegnanarayana; IIIT Hyderabad, India

Fri-O-1-6-3, Time: 11:40

Automatic detection of vowel landmarks is useful in many applications such as automatic speech recognition (ASR), audio search, syllabification of speech and expressive speech processing. In this paper, acoustic features extracted around epochs are proposed for detection of vowel landmarks in continuous speech. These features are based on zero frequency filtering (ZFF) and single frequency filtering (SFF) analyses of speech. Excitation source based features are extracted using ZFF method and vocal tract system based features are extracted using SFF method. Based on these features, a rule-based algorithm is developed for vowel landmark detection (VLD). Performance of the proposed VLD algorithm is studied on three different databases namely, TIMIT (read), NTIMIT (channel degraded) and Switchboard corpus (conversational speech). Results show that the proposed algorithm performs equally well compared to state-of-the-art techniques on TIMIT and better on NTIMIT and Switchboard corpora. Proposed algorithm also displays consistent performance on TIMIT and NTIMIT datasets for different levels of noise degradations.

Sensitivity of Quantitative RT-MRI Metrics of Vocal Tract Dynamics to Image Reconstruction Settings

Johannes Töger, Yongwan Lim, Sajan Goud Lingala, Shrikanth S. Narayanan, Krishna S. Nayak; University of Southern California, USA

Fri-O-1-6-4, Time: 12:00

Real-time Magnetic Resonance Imaging (RT-MRI) is a powerful method for quantitative analysis of speech. Current state-of-the-art methods use constrained reconstruction to achieve high frame rates and spatial resolution. The reconstruction involves two free parameters that can be retrospectively selected: 1) the temporal resolution and 2) the regularization parameter λ, which balances temporal regularization and fidelity to the collected MRI data. In this work, we study the sensitivity of derived quantitative measures of vocal tract function to these two parameters. Specifically, the cross-distance between the tongue tip and the alveolar ridge was investigated for different temporal resolutions (21, 42, 56 and 83 frames per second) and values of the regularization parameter. Data from one subject is included. The phrase ‘one two three four five’ was repeated 8 times at a normal pace. The results show that 1) a high regularization factor leads to lower cross-distance values 2) using a low value for the regularization parameter gives poor reproducibility and 3) a temporal resolution of at least 42 frames per second is desirable to achieve good reproducibility for all utterances in this speech task. The process employed here can be generalized to quantitative imaging of the vocal tract and other body parts.

Sound Pattern Matching for Automatic Prosodic Event Detection

Milos Cernak, Afsaneh Asaei, Pierre-Edouard Honnet, Philip N. Garner, Hervé Bourlard; Idiap Research Institute, Switzerland

Fri-O-1-6-5, Time: 12:20

Prosody in speech is manifested by variations of loudness, exaggeration of pitch, and specific phonetic variations of prosodic segments. For example, in the stressed and unstressed syllables, there are differences in place or manner of articulation, vowels in unstressed syllables may have a more central articulation, and vowel reduction may occur when a vowel changes from a stressed to an unstressed position.

In this paper, we characterize the sound patterns using phonological posteriors to capture the phonetic variations in a concise manner. The phonological posteriors quantify the posterior probabilities of the phonological classes given the input speech acoustics, and they are obtained using the deep neural network (DNN) computational method. Built on the assumption that there are unique sound patterns in different prosodic segments, we devise a sound pattern matching (SPM) method based on 1-nearest neighbour classifier. In this work, we focus on automatic detection of prosodic stress placed on words, called also emphasized words. We evaluate the SPM method on English and French data with emphasized words. The word emphasis detection works very well also on cross-lingual tests, that is using a French classifier on English data, and vice versa.

Automatic Classification of Lexical Stress in English and Arabic Languages Using Deep Learning

Mostafa Shahin[1], Julien Epps[2], Beena Ahmed[1]; [1]Texas A&M University, Qatar; [2]University of New South Wales, Australia

Fri-O-1-6-6, Time: 12:40

Prosodic features are important for the intelligibility and proficiency of stress-timed languages such as English and Arabic. Producing the appropriate lexical stress is challenging for second language (L2) learners, in particular, those whose first language (L1) is a syllable-timed language such as Spanish, French, etc. In this paper we introduce a method for automatic classification of lexical stress

NOTES

to be integrated into computer-aided pronunciation learning (CAPL) tools for L2 learning. We trained two different deep learning architectures, the deep feedforward neural network (DNN) and the deep convolutional neural network (CNN) using a set of temporal and spectral features related to the intensity, duration, pitch and energies in different frequency bands. The system was applied on both English (kids and adult) and Arabic (adult) speech corpora collected from native speakers. Our method results in error rates of 9%, 7% and 18% when tested on the English children corpus, English adult corpus and Arabic adult corpus respectively.

Fri-P-1-1 : First and Second Language Acquisition

Pacific Concourse – Poster A, 11:00–13:00, Friday, 9 Sept. 2016
Chair: Maria Luisa Garcia Lecumberri

Development of Mandarin Onset-Rime Detection in Relation to Age and Pinyin Instruction

Fei Chen, Nan Yan, Xunan Huang, Hao Zhang, Lan Wang, Gang Peng; Chinese Academy of Sciences, China

Fri-P-1-1-1, Time: 11:00

Development of explicit phonological awareness (PA) is thought to be dependent on formal instruction in reading or spelling. However, the development of implicit PA emerges before literacy instruction and interacts with how the phonological representations are constructed within a certain language. The present study systematically investigated the development of implicit PA of Mandarin onset-rime detection in relation to age and Pinyin instruction, involving 70 four- to seven-year-old kindergarten and first-grade children. Results indicated that the overall rate of correct responses in the rime detection task was much higher than that in the onset detection one, with better discrimination ability of larger units. Moreover, the underlying factors facilitating the development of Mandarin onset and rime detection were different, although both correlated positively with Pinyin instruction. On one hand, with age, development of rime detection appeared to develop naturally through spoken language experience before schooling, and was further optimized to the best after Pinyin instruction. On the other hand, the accuracy of onset detection exhibited a drastic improvement, boosting from 66% among preschoolers to 93% among first graders, establishing the primacy of Pinyin instruction responsible for the development of implicit onset awareness in Mandarin.

Joint Effect of Dialect and Mandarin on English Vowel Production: A Case Study in Changsha EFL Learners

Xinyi Wen, Yuan Jia; Chinese Academy of Social Sciences, China

Fri-P-1-1-2, Time: 11:00

Phonetic acquisition of English as a Foreign Language (EFL) for learners in dialectal areas has been increasingly regarded as an important research area in second language acquisition. However, most existing research has been focused on finding out the transfer effect of dialect on English production from a second language acquisition point of view, but ignores the impact of Mandarin. The present research aims to investigate the joint effect of dialect and Mandarin on Changsha EFL learners' vowel production through acoustic analysis, from both spectral and temporal perspectives. We will further explain the results with the Speech Learning Model (SLM). Three corner vowels, i.e., /a/ /i/ /u/, are studied, and the results show that: English vowels /i/ and /a/ produced by Changsha learners are significantly different from those of American speakers; specifically, /i/ is more affected by Mandarin, and /a/ is more affected by Changsha dialect, which can be explained by SLM. While /u/ produced by Changsha learners is similar to that of American speakers. Besides, Changsha learners produce shorter vowels in duration, due to dialect and Mandarin's transfer effect, but can still make tense-lax contrasts in /i-ɪ/ and /u-ʊ/ pairs.

Effects of L1 Phonotactic Constraints on L2 Word Segmentation Strategies

Tamami Katayama; Prefectural University of Hiroshima, Japan

Fri-P-1-1-3, Time: 11:00

In the present study, it was examined whether phonotactic constraints of the first language affect speech processing by Japanese learners of English and whether L2 proficiency influences it. Seventeen native English speakers (ES), 18 Japanese speakers with high proficiency of English (JH), and 20 Japanese speakers with relatively low English proficiency (JL) took part in a monitoring task. Two types of target words (CVC/CV, e.g., *team/tea*) were embedded in bisyllabic non-words (e.g., *teamfesh*) and given to the participants with other non-words in the lists. The three groups were instructed to respond as soon as they spot targets, and response times and error rates were analyzed. The results showed that all of the groups segmented the CVC target words significantly faster and more accurately than the CV targets. L1 phonotactic constraints did not hinder L2 speech processing, and a word segmentation strategy was not language-specific in the case of Japanese English learners.

Putting German [ʃ] and [ç] in Two Different Boxes: Native German vs L2 German of French Learners

Jane Wottawa[1], Martine Adda-Decker[1], Frédéric Isel[2]; [1]LPP (UMR 7018), France; [2]MoDyCo, France

Fri-P-1-1-4, Time: 11:00

French L2 Learners of German (FG) often replace the palatal fricative /ç/ absent in French with the post alveolar fricative /ʃ/. In our study we investigate which cues can be used to distinguish whether FG speakers produce [ʃ] or [ç] in words with the final syllables /ɪʃ/ or /ɪç/. In literature of German as an L2, to our knowledge, this contrast has not yet been studied. In this perspective, we first compared native German (GG) productions of [/ʃ/] and [ç] to the FG speaker productions. Comparisons concerned the F2 of the preceding vowel, the F2 transition between the preceding vowel and the fricative, the center of gravity and intensity of the fricatives in high and low frequencies. To decide which cues are effectively choices to separate [ʃ] and [ç], the Weka interface in R (RWeka) was used. Results show that for German native speech, the F2 of the preceding vowel and the F2 transition are valid cues to distinguish between [ʃ] and [ç]. For FG speakers these cues are not valid. To distinguish between [ʃ] and [ç] in FG speakers, the intensity of high and low frequencies as well as the center of gravity of the fricatives help to decide whether [ʃ] and [ç] was produced. In German native speech, cues furnished only by the fricative itself can as well be used to distinguish between [ʃ] and [ç].

NOTES

Naturalness Judgement of L2 English Through Dubbing Practice

Dean Luo [1], Ruxin Luo [2], Lixin Wang [3]; [1] Shenzhen Institute of Information Technology, China; [2] Shenzhen Polytechnic, China; [3] Shenzhen Seaskyland Technologies, China

Fri-P-1-1-5, Time: 11:00

This Study investigates how different prosodic features affect native speakers' perception of L2 English spoken by Chinese students through dubbing, or re-voicing practice on video clips. Learning oral foreign language through dubbing on movie or animation clips has become very popular in China. In this practice, learners try to reproduce utterances as closely as possible to the original speech by closely matching lip movements on the clips. The L2 utterances before and after substantial dubbing practices were recorded and categorized according to different prosodic error patterns. Objective acoustic features were extracted and analyzed with naturalness scores based on perceptual experiment. Experimental results show that stress and timing play key roles in native speakers' perception of naturalness. With the practice of dubbing, prosodic features, especially timing, can be considerably improved and thus the naturalness of the reproduced utterances increases.

Audiovisual Training Effects for Japanese Children Learning English /r/-/l/

Yasuaki Shinohara; Waseda University, Japan

Fri-P-1-1-6, Time: 11:00

In this study, the effects of audiovisual training were examined for Japanese children learning the English /r/-/l/ contrast. After 10 audiovisual training sessions, participants' improvement in English /r/-/l/ identification in audiovisual, visual-only and audio-only conditions was assessed. The results demonstrated that Japanese children significantly improved in their English /r/-/l/ identification accuracy in all three conditions. Although there was no significant modality effect on identification accuracy at pre test, the participants improved their identification accuracy in the audiovisual condition significantly more than in the audio-only condition. The improvement in the audiovisual condition was not significantly different from that in the visual-only condition. These results suggest that Japanese children can improve their identification accuracy of the English /r/-/l/ contrasts using each of visual and auditory modalities, and they appear to improve their lip-reading skills as much as audiovisual identification. Nonetheless, due to the ceiling effect in their improvement, it is unclear whether Japanese children improved their integrated processing of visual and auditory information.

L2 Acquisition and Production of the English Rhotic Pharyngeal Gesture

Sarah Harper, Louis Goldstein, Shrikanth S. Narayanan; University of Southern California, USA

Fri-P-1-1-7, Time: 11:00

This study is an investigation of L2 speakers' production of the pharyngeal gesture in the English /ɹ/. Real-time MRI recordings from one L1 French/L2 English and one L1 Greek/L2 English speaker were analyzed and compared with recordings from a native English speaker to examine whether the gestural composition of the rhotic consonant(s) in a speaker's L1, particularly the presence and location of a pharyngeal gesture, influences their production of English /ɹ/. While the L1 French speaker produced the expected high pharyngeal constriction in their production of the French rhotic, he did not appear to consistently produce an English-like low pharyngeal constriction in his production of English /ɹ/. Similarly, the native Greek speaker did not consistently produce a pharyngeal constriction of any kind in either his L1 rhotic (as expected) or in English /ɹ/. These results suggest that the acquisition and production of the pharyngeal gesture in the English rhotic approximant is particularly difficult for learners whose L1 rhotics lack an identical constriction, potentially due to a general difficulty of acquiring pharyngeal gestures that are not in the L1, the similarity of the acoustic consequences of the different components of a rhotic, or L1 transfer into the L2.

Fri-P-1-2 : Speech and Hearing Disorders & Perception

Pacific Concourse – Poster B, 11:00–13:00, Friday, 9 Sept. 2016
Chair: Frank Rudzicz

Auditory-Visual Perception of VCVs Produced by People with Down Syndrome: Preliminary Results

Alexandre Hennequin, Amélie Rochet-Capellan, Marion Dohen; GIPSA, France

Fri-P-1-2-1, Time: 11:00

Down Syndrome (DS) is a genetic disease involving a number of anatomical, physiological and cognitive impairments. More particularly it affects speech production abilities. This results in reduced intelligibility which has however only been evaluated auditorily. Yet, many studies have demonstrated that adding vision to audition helps perception of speech produced by people without impairments especially when it is degraded as is the case in noise. The present study aims at examining whether the visual information improves intelligibility of people with DS. 24 participants without DS were presented with VCV sequences (vowel-consonant-vowel) produced by four adults (2 with DS and 2 without DS). These stimuli were presented in noise in three modalities: auditory, auditory-visual and visual. The results confirm a reduced auditory intelligibility of speakers with DS. They also show that, for the speakers involved in this study, visual intelligibility is equivalent to that of speakers without DS and compensates for the auditory intelligibility loss. An analysis of the perceptual errors shows that most of them involve confusions between consonants. These results put forward the crucial role of multimodality in the improvement of the intelligibility of people with DS.

Combining Non-Pathological Data of Different Language Varieties to Improve DNN-HMM Performance on Pathological Speech

Emre Yılmaz, Mario Ganzeboom, Catia Cucchiarini, Helmer Strik; Radboud Universiteit Nijmegen, The Netherlands

Fri-P-1-2-2, Time: 11:00

Research on automatic speech recognition (ASR) of pathological speech is particularly hindered by scarce in-domain data resources. Collecting representative pathological speech data is difficult due to the large variability caused by the nature and severity of the disor-

Notes

ders, and the rigorous ethical and medical permission requirements. This task becomes even more challenging for languages which have fewer resources, fewer speakers and fewer patients than English, such as the mid-sized language Dutch. In this paper, we investigate the impact of combining speech data from different varieties of the Dutch language for training deep neural network (DNN)-based acoustic models. Flemish is chosen as the target variety for testing the acoustic models, since a Flemish database of pathological speech, the COPAS database, is available. We use non-pathological speech data from the northern Dutch and Flemish varieties and perform speaker-independent recognition using the DNN-HMM system trained on the combined data. The results show that this system provides improved recognition of pathological Flemish speech compared to a baseline system trained only on Flemish data. These findings open up new opportunities for developing useful ASR-based pathological speech applications for languages that are smaller in size and less resourced than English.

Evaluation of a Phone-Based Anomaly Detection Approach for Dysarthric Speech

Imed Laaridh[1], Corinne Fredouille[1], Christine Meunier[2]; [1]LIA, France; [2]LPL, France

Fri-P-1-2-3, Time: 11:00

Perceptual evaluation is still the most common method in clinical practice for the diagnosing and the following of the condition progression of people with speech disorders. Many automatic approaches were proposed to provide objective tools to deal with speech disorders and help professionals in the severity evaluation of speech impairments. This paper investigates an automatic phone-based anomaly detection approach implying an automatic text-constrained phone alignment. Here, anomalies are related to speech segments, for which an unexpected acoustic pattern is observed, compared with a normal speech production. This objective tool is applied to French dysarthric speech recordings produced by patients suffering from four different pathologies. The behavior of the anomaly detection approach is studied according to the precision of the automatic phone alignment. Faced with the difficulties of having a gold standard reference, especially for the phone-based anomaly annotation, this behavior is observed on both annotated and non-annotated corpora. As expected, alignment errors (large shifts compared with a manual segmentation) lead to a large amount of anomalies automatically detected. However, about 50% of correctly detected anomalies are not related to alignment errors. This behavior shows that the automatic approach is able to catch irregular acoustic patterns of phones.

Recognition of Dysarthric Speech Using Voice Parameters for Speaker Adaptation and Multi-Taper Spectral Estimation

Chitralekha Bhat, Bhavik Vachhani, Sunil Kopparapu; Tata Consultancy Services, India

Fri-P-1-2-4, Time: 11:00

Dysarthria is a motor speech disorder resulting from impairment in muscles responsible for speech production, often characterized by slurred or slow speech resulting in low intelligibility. With speech based applications such as voice biometrics and personal assistants gaining popularity, automatic recognition of dysarthric speech becomes imperative as a step towards including people with dysarthria into mainstream. In this paper we examine the applicability of voice parameters that are traditionally used for pathological voice classification such as jitter, shimmer, F0 and Noise Harmonic Ratio (NHR) contour in addition to Mel Frequency Cepstral Coefficients (MFCC) for dysarthric speech recognition. Additionally, we show that multi-taper spectral estimation for computing MFCC improves the unseen dysarthric speech recognition. A Deep neural network (DNN) - hidden Markov model (HMM) recognition system fared better than a Gaussian Mixture Model (GMM) - HMM based system for dysarthric speech recognition. We propose a method to optimally use incremental dysarthric data to improve dysarthric speech recognition for an ASR with DNN-HMM. All evaluations were done on Universal Access Speech Corpus.

Impaired Categorical Perception of Mandarin Tones and its Relationship to Language Ability in Autism Spectrum Disorders

Fei Chen[1], Nan Yan[1], Xiaojie Pan[2], Feng Yang[3], Zhuanzhuan Ji[1], Lan Wang[1], Gang Peng[1]; [1]Chinese Academy of Sciences, China; [2]Shenzhen Love Wisdom Special Children Rehabilitation Center, China; [3]Shenzhen Children's Hospital, China

Fri-P-1-2-5, Time: 11:00

While enhanced pitch processing appears to be characteristic of many individuals with autism spectrum disorders (ASD), it remains unclear whether enhancement in pitch perception applies to those who speak a tone language. Using a classic paradigm of categorical perception (CP), the present study investigated the perception of Mandarin tones in six- to eight-year-old children with ASD, and compared it with age-matched typically developing children. In stark contrast to controls, the child participants with ASD exhibited a much wider boundary width (i.e., more gentle slope), and showed no improved discrimination for pairs straddling the boundary, indicating impaired CP of Mandarin tones. Moreover, identification skills of different tone categories were positively correlated with language ability among children with ASD. These findings revealed aberrant tone processing in Mandarin-speaking individuals with ASD, especially in those with significant language impairment. Our results are in support of the notion of impaired change detection for the linguistic elements of speech in children with ASD.

Perceived Naturalness of Electrolaryngeal Speech Produced Using sEMG-Controlled vs. Manual Pitch Modulation

K.F. Nagle[1], J.T. Heaton[2]; [1]Seton Hall University, USA; [2]Massachusetts General Hospital, USA

Fri-P-1-2-6, Time: 11:00

Producing speech with natural prosodic patterns is an ongoing challenge for users of electrolaryngeal (EL) speech. This study describes speech produced using a method currently in development, wherein a prosodic pattern is derived from skin surface electromyographical (sEMG) signals recorded from under the chin (submental surface).

Eight laryngectomees who currently use a TruTone EL as their primary or backup mode of speech provided samples of EL speech in two modes: conventional thumb-pressure pitch-modulated control (represented by the TruTone EL; Griffin Laboratories, CA, U.S.A.) and sEMG-based pitch-modulated control (EMG-EL). Ratings of perceived

NOTES

naturalness were obtained from ten listeners unfamiliar with EL speech.

Listener ratings indicated that five speakers produced equally natural speech using both devices, and three produced significantly more natural speech using the EMG-EL than the TruTone EL. Mean fundamental frequency (f0) was similar within speakers for both modes; however, mean f0 range and standard deviation were significantly larger for the EMG-EL than for the TruTone EL, despite both devices having similar potential f0 range. This study showed that the EMG-EL provides an intuitive means of controlling f0-based prosodic patterns that are more natural-sounding than push-button control for some EL users.

Identifying Hearing Loss from Learned Speech Kernels

Shamima Najnin, Bonny Banerjee, Lisa Lucks Mendel, Masoumeh Heidari Kapourchali, Jayanta Kumar Dutta, Sungmin Lee, Chhayakanta Patro, Monique Pousson; University of Memphis, USA

Fri-P-1-2-7, Time: 11:00

Does a hearing-impaired individual's speech reflect his hearing loss? To investigate this question, we recorded at least four hours of speech data from each of 29 adult individuals, both male and female, belonging to four classes: 3 normal, and 26 severely-to-profoundly hearing impaired with high, medium or low speech intelligibility. Acoustic kernels were learned for each individual by capturing the distribution of his speech data points represented as 20 ms duration windows. These kernels were evaluated using a set of neurophysiological metrics, namely, distribution of characteristic frequencies, equal loudness contour, bandwidth and Q_{10} value of tuning curve. It turns out that, for our cohort, a feature vector can be constructed out of four properties of these metrics that would accurately classify hearing-impaired individuals with low intelligible speech from normal ones using a linear classifier. However, the overlap in the feature space between normal and hearing-impaired individuals increases as the speech becomes more intelligible. We conclude that a hearing-impaired individual's speech does reflect his hearing loss provided his loss of hearing has considerably affected the intelligibility of his speech.

Differential Effects of Velopharyngeal Dysfunction on Speech Intelligibility During Early and Late Stages of Amyotrophic Lateral Sclerosis

Panying Rong[1], Yana Yunusova[2], Jordan R. Green[1]; [1]MGH Institute of Health Professions, USA; [2]University of Toronto, Canada

Fri-P-1-2-8, Time: 11:00

The detrimental effects of velopharyngeal dysfunction (VPD) on speech intelligibility in persons with progressive motor speech disorders are poorly understood. In this study, we longitudinally investigated the velopharyngeal and articulatory performance of 142 individuals with varying severities of amyotrophic lateral sclerosis (ALS). Our goal was to determine the mechanisms that underlie the effects of VPD on speech intelligibility during early and late stages of ALS progression. We found that during the early stages of the disease, the effect of VPD on intelligibility was partially mitigated by an increase in articulatory (e.g., lower lip and jaw) movement speed. This apparent articulatory compensation eventually became unavailable during the late stages of disease progression, which led to rapid declines of speech intelligibility. The transition across the early and late stages was characterized by the slowing of the composite movement of lower lip and jaw below 138 mm/s, which indicated the onset of precipitous speech decline and thus, may provide important timing information for helping clinicians to plan interventions.

The Production of Intervocalic Glides in Non Dysarthric Parkinsonian Speech

V. Delvaux, V. Roland, K. Huet, M. Piccaluga, M.C. Haelewyck, B. Harmegnies; Université de Mons, Belgium

Fri-P-1-2-9, Time: 11:00

In the context of a research project aiming at investigating the relationships between speech disorders, quality of life and social participation in Parkinson's Disease (PD), we report here on an acoustic study of glides and steady vowels by non dysarthric parkinsonian and control speakers. Our specific aim is to explore the dynamics of supra-laryngeal articulators in PD. Results suggest that non dysarthric Parkinsonian speakers maintain an accurate production of glides in VC[glide]V pseudo-words at the expense of articulatory undershoot in the surrounding vowels, and some asymmetry between the V1-to-glide and glide-to-V2 articulatory movements. We discuss how these results both support and challenge the accuracy-tempo trade-off hypothesis (Ackermann and Ziegler, 1991).

Auditory Processing Impairments Under Background Noise in Children with Non-Syndromic Cleft Lip and/or Palate

Yang Feng, Zhang Lu; Shenzhen Children's Hospital, China

Fri-P-1-2-10, Time: 11:00

Cleft lip and/or palate (CL/P) disorders are commonly occurring congenital malformations and hearing impairment is a very common co-morbidity. Most previous research has only focused on middle ear disorders and related auditory consequences in this group. Studies of higher level auditory status and central auditory processing abilities of this group have been unsystematic. The present study was conducted in order to objectively investigate the central auditory abilities in children with non-syndromic cleft lip and/or palate (NSCLP). A structured behavioral central auditory test battery was conducted in a group of children with NSCLP and their age/sex matched normal peers. The following behavioral central auditory tasks were undertaken, including hearing in noise test (HINT), dichotic digits test (DDT), and gaps in noise test (GIN). Results showed that there were no significant group differences in DDT test, indicating that the binaural separation and integration abilities could be normal in children with NSCLP. However, the cleft group performed significantly poorer than their normal peers for each ear in HINT test under noise condition and GIN test, suggesting that the children with NSCLP could have impaired monaural low redundancy auditory processing ability, and at risk of temporal resolution disability.

NOTES

Modulation Spectral Features for Predicting Vocal Emotion Recognition by Simulated Cochlear Implants

Zhi Zhu[1]*, Ryota Miyauchi*[1]*, Yukiko Araki*[2]*, Masashi Unoki*[1]*;* [1]*JAIST, Japan;* [2]*Kanazawa University, Japan*

Fri-P-1-2-11, Time: 11:00

It has been reported that vocal emotion recognition is challenging for cochlear implant (CI) listeners due to the limited spectral cues with CI devices. As the mechanism of CI, modulation information is provided as a primarily cue. Previous studies have revealed that the modulation components of speech are important for speech intelligibility. However, it is unclear whether modulation information can contribute to vocal emotion recognition. We investigated the relationship between human perception of vocal emotion and the modulation spectral features of emotional speech. For human perception, we carried out a vocal-emotion recognition experiment using noise-vocoder simulations with normal-hearing listeners to predict the response from CI listeners. For modulation spectral features, we used auditory-inspired processing (auditory filterbank, temporal envelope extraction, modulation filterbank) to obtain the modulation spectrogram of emotional speech signals. Ten types of modulation spectral feature were then extracted from the modulation spectrogram. As a result, modulation spectral centroid, modulation spectral kurtosis, and modulation spectral tilt exhibited similar trends with the results of human perception. This suggests that these modulation spectral features may be important cues for voice emotion recognition with noise-vocoded speech.

Automatic Discrimination of Soft Voice Onset Using Acoustic Features of Breathy Voicing

Keiko Ochi[1]*, Koichi Mori*[2]*, Naomi Sakai*[2]*, Nobutaka Ono*[1]*;* [1]*NII, Japan;* [2]*National Rehabilitation Center for Persons with Disabilities, Japan*

Fri-P-1-2-12, Time: 11:00

Soft onset vocalization is used in certain speech therapies. However, it is not easy to practice it at home because the acoustical evaluation itself needs training. It would be helpful for speech patients to get objective feedback during training. In this paper, new parameters for identifying soft onset with high accuracy are described. One of the parameters measures an aspect of the soft voice onset, in which the vocal folds start to oscillate periodically before coming in contact with each other at the beginning of vocalization. Combined with an onset time exceeding a threshold, the proposed parameters gave about 99% accuracy in identifying soft onset vocalization.

Effect of Noise on Lexical Tone Perception in Cantonese-Speaking Amusics

Jing Shao, Caicai Zhang, Gang Peng, Yike Yang, William S.-Y. Wang; Hong Kong Polytechnic University, China

Fri-P-1-2-13, Time: 11:00

Congenital amusia is a neurogenetic disorder affecting musical pitch processing. It also affects lexical tone perception. It is well documented that noisy conditions impact speech perception in second language learners and cochlear implant users. However, it is yet unclear whether and how noise affects lexical tone perception in the amusics. This paper examined the effect of multi-talker babble noise [1] on lexical tone identification and discrimination in 14 Cantonese-speaking amusics and 14 controls at three levels of signal-to-noise ratio (SNR). Results reveal that the amusics were less accurate in the *identification* of tones compared to controls in all SNR conditions. They also showed degraded performance in the *discrimination*, but less severe than in the identification. These results confirmed that amusia influences lexical tone processing. But the amusics were not influenced more by noise than the controls in either identification or discrimination. This indicates that the deficits of amusia may not be due to the lack of native-like language processing mechanisms or are mechanical in nature, as in the case of second language learners and cochlear implant users. Instead, the amusics may be impaired in the linguistic processing of native tones, showing impaired tone perception already under the clear condition.

Audio-Visual Speech Recognition Using Bimodal-Trained Bottleneck Features for a Person with Severe Hearing Loss

Yuki Takashima[1]*, Ryo Aihara*[1]*, Tetsuya Takiguchi*[1]*, Yasuo Ariki*[1]*, Nobuyuki Mitani*[2]*, Kiyohiro Omori*[2]*, Kaoru Nakazono*[2]*;* [1]*Kobe University, Japan;* [2]*Hyogo Institute of Assistive Technology, Japan*

Fri-P-1-2-14, Time: 11:00

In this paper, we propose an audio-visual speech recognition system for a person with an articulation disorder resulting from severe hearing loss. In the case of a person with this type of articulation disorder, the speech style is quite different from those of people without hearing loss that a speaker-independent acoustic model for unimpaired persons is hardly useful for recognizing it. The audio-visual speech recognition system we present in this paper is for a person with severe hearing loss in noisy environments. Although feature integration is an important factor in multimodal speech recognition, it is difficult to integrate efficiently because those features are different intrinsically. We propose a novel visual feature extraction approach that connects the lip image to audio features efficiently, and the use of convolutive bottleneck networks (CBNs) increases robustness with respect to speech fluctuations caused by hearing loss. The effectiveness of this approach was confirmed through word-recognition experiments in noisy environments, where the CBN-based feature extraction method outperformed the conventional methods.

Perception of Tone in Whispered Mandarin Sentences: The Case for Singapore Mandarin

*Yuling Gu, Boon Pang Lim, Nancy F. Chen; A*STAR, Singapore*

Fri-P-1-2-15, Time: 11:00

Whispering is commonly used when one needs to speak softly (for instance, in a library). Whispered speech mainly differs from neutral speech in that voicing, and thus its acoustic correlate F0, is absent. It is well known that in tonal languages such as Mandarin, tone identity is primarily conveyed by the F0 contour. Previous works also suggest that secondary correlates are both consistent and sufficient to convey Mandarin tone in whisper. However, these results are focused on Standard Mandarin spoken in Mainland China and have only been obtained via small-scale experiments using citation-form speech. To investigate whether these results will carry over to continuous sentences in other variations of Mandarin, we

NOTES

present a study that is the first of its nature to explore native Singapore Mandarin. Unlike related works, our large-scale perceptual experiment thoroughly investigates lexical tones in whispered and neutral Mandarin by involving more diverse speech data, greater number of listeners and use syllables excised from continuous speech to better simulate natural speech conditions. Our findings differ significantly from earlier works in terms of the recognition patterns observed. We present further in-depth analysis on how various phonetic characteristics (vowel contexts, place and manner of articulation) affect whispered tone perception.

Fri-P-1-3 : Speech Synthesis Poster

Pacific Concourse – Poster C, 11:00–13:00, Friday, 9 Sept. 2016
Chair: Alan Black

A KL Divergence and DNN-Based Approach to Voice Conversion without Parallel Training Sentences

Feng-Long Xie[1], Frank K. Soong[2], Haifeng Li[1]; [1]Harbin Institute of Technology, China; [2]Microsoft, China

Fri-P-1-3-1, Time: 11:00

We extend our recently proposed approach to cross-lingual TTS training to voice conversion, without using parallel training sentences. It employs Speaker Independent, Deep Neural Net (SI-DNN) ASR to equalize the difference between source and target speakers and Kullback-Leibler Divergence (KLD) to convert spectral parameters probabilistically in the phonetic space via ASR senone posterior probabilities of the two speakers. With or without knowing the transcriptions of the target speaker's training speech, the approach can be either supervised or unsupervised. In a supervised mode, where adequate training data of the target speaker with transcriptions is used to train a GMM-HMM TTS of the target speaker, each frame of the source speakers input data is mapped to the closest senone in thus trained TTS. The mapping is done via the posterior probabilities computed by SI-DNN ASR and the minimum KLD matching. In a unsupervised mode, all training data of the target speaker is first grouped into phonetic clusters where KLD is used as the sole distortion measure. Once the phonetic clusters are trained, each frame of the source speakers input is then mapped to the mean of the closest phonetic cluster. The final converted speech is generated with the max probability trajectory generation algorithm. Both objective and subjective evaluations show the proposed approach can achieve higher speaker similarity and better spectral distortions, when comparing with the baseline system based upon our sequential error minimization trained DNN algorithm.

Parallel Dictionary Learning for Voice Conversion Using Discriminative Graph-Embedded Non-Negative Matrix Factorization

Ryo Aihara, Tetsuya Takiguchi, Yasuo Ariki; Kobe University, Japan

Fri-P-1-3-2, Time: 11:00

This paper proposes a discriminative learning method for Non-negative Matrix Factorization (NMF)-based Voice Conversion (VC). NMF-based VC has been researched because of the natural-sounding voice it produces compared with conventional Gaussian Mixture Model (GMM)-based VC. In conventional NMF-based VC, parallel exemplars are used as the dictionary; therefore, dictionary learning is not adopted. In order to enhance the conversion quality of NMF-based VC, we propose Discriminative Graph-embedded Non-negative Matrix Factorization (DGNMF). Parallel dictionaries of the source and target speakers are discriminatively estimated by using DGNMF based on the phoneme labels of the training data. Experimental results show that our proposed method can not only improve the conversion quality but also reduce the computational times.

Speech Bandwidth Extension Using Bottleneck Features and Deep Recurrent Neural Networks

Yu Gu, Zhen-Hua Ling, Li-Rong Dai; USTC, China

Fri-P-1-3-3, Time: 11:00

This paper presents a novel method for speech bandwidth extension (BWE) using deep structured neural networks. In order to utilize linguistic information during the prediction of high-frequency spectral components, the bottleneck (BN) features derived from a deep neural network (DNN)-based state classifier for narrowband speech are employed as auxiliary input. Furthermore, recurrent neural networks (RNNs) incorporating long short-term memory (LSTM) cells are adopted to model the complex mapping relationship between the feature sequences describing low-frequency and high-frequency spectra. Experimental results show that the BWE method proposed in this paper can achieve better performance than the conventional method based on Gaussian mixture models (GMMs) and the state-of-the-art approach based on DNNs in both objective and subjective tests.

Voice Conversion Based on Matrix Variate Gaussian Mixture Model Using Multiple Frame Features

Yi Yang, Hidetsugu Uchida, Daisuke Saito, Nobuaki Minematsu; University of Tokyo, Japan

Fri-P-1-3-4, Time: 11:00

This paper presents a novel voice conversion method based on matrix variate Gaussian mixture model (MV-GMM) using features of multiple frames. In voice conversion studies, approaches based on Gaussian mixture models (GMM) are still widely utilized because of their flexibility and easiness in handling. They treat the joint probability density function (PDF) of feature vectors from source and target speakers as that of joint vectors of the two vectors. Addition of dynamic features to the feature vectors in GMM-based approaches achieves certain performance improvements because the correlation between multiple frames is taken into account. Recently, a voice conversion framework based on MV-GMM, in which the joint PDF is modeled in a matrix variate space, has been proposed and it is able to precisely model both the characteristics of the feature spaces and the relation between the source and target speakers. In this paper, in order to additionally model the correlation between multiple frames in the framework more consistently, MV-GMM is constructed in a matrix variate space containing the features of neighboring frames. Experimental results show that an certain performance improvement in both objective and subjective evaluations is observed.

NOTES

Voice Conversion Based on Trajectory Model Training of Neural Networks Considering Global Variance

Naoki Hosaka, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, Keiichi Tokuda; Nagoya Institute of Technology, Japan

Fri-P-1-3-5, Time: 11:00

This paper proposes a new training method of deep neural networks (DNNs) for statistical voice conversion. DNNs are now being used as conversion models that represent mapping from source features to target features in statistical voice conversion. However, there are two major problems to be solved in conventional DNN-based voice conversion: 1) the inconsistency between the training and synthesis criteria, and 2) the over-smoothing of the generated parameter trajectories. In this paper, we introduce a parameter trajectory generation process considering the global variance (GV) into the training of DNNs for voice conversion. A consistent framework using the same criterion for both training and synthesis provides better conversion accuracy in the original static feature domain, and the over-smoothing can be avoided by optimizing the DNN parameters on the basis of the trajectory likelihood considering the GV. Experimental results show that the proposed method outperforms the DNN-based method in term of both speech quality and speaker similarity.

Comparing Articulatory and Acoustic Strategies for Reducing Non-Native Accents

Sandesh Aryal, Ricardo Gutierrez-Osuna; Texas A&M University, USA

Fri-P-1-3-6, Time: 11:00

This article presents an experimental comparison of two types of techniques, articulatory and acoustic, for transforming non-native speech to sound more native-like. Articulatory techniques use articulators from a native speaker to drive an articulatory synthesizer of the non-native speaker. These methods have a good theoretical justification, but articulatory measurements (e.g., via electromagnetic articulography) are difficult to obtain. In contrast, acoustic methods use techniques from the voice conversion literature to build a mapping between the two acoustic spaces, making them more attractive for practical applications (e.g., language learning). We compare two representative implementations of these approaches, both based on statistical parametric speech synthesis. Through a series of perceptual listening tests, we evaluate the two approaches in terms of accent reduction, speech intelligibility and speaker quality. Our results show that the acoustic method is more effective than the articulatory method in reducing perceptual ratings of non-native accents, and also produces synthesis of higher intelligibility while preserving voice quality.

Cross-Lingual Speaker Adaptation for Statistical Speech Synthesis Using Limited Data

Seyyed Saeed Sarfjoo, Cenk Demiroglu; Özyeğin Üniversitesi, Turkey

Fri-P-1-3-7, Time: 11:00

Cross-lingual speaker adaptation with limited adaptation data has many applications such as use in speech-to-speech translation systems. Here, we focus on cross-lingual adaptation for statistical speech synthesis (SSS) systems using limited adaptation data. To that end, we propose two techniques exploiting a bilingual Turkish-English speech database that we collected. In one approach, speaker-specific state-mapping is proposed for cross-lingual adaptation which performed significantly better than the baseline state-mapping algorithm in adapting the excitation parameter both in objective and subjective tests. In the second approach, eigenvoice adaptation is done in the input language which is then used to estimate the eigenvoice weights in the output language using weighted linear regression. The second approach performed significantly better than the baseline system in adapting the spectral envelope parameters both in objective and subjective tests.

Personalized, Cross-Lingual TTS Using Phonetic Posteriorgrams

Lifa Sun, Hao Wang, Shiyin Kang, Kun Li, Helen Meng; Chinese University of Hong Kong, China

Fri-P-1-3-8, Time: 11:00

We present a novel approach that enables a target speaker (e.g. monolingual Chinese speaker) to speak a new language (e.g. English) based on arbitrary textual input. Our system includes a trained English speaker-independent automatic speech recognition (SI-ASR) engine using TIMIT. Given the target speaker's speech in a non-target language, we generate Phonetic PosteriorGrams (PPGs) with the SI-ASR and then train a Deep Bidirectional Long Short-Term Memory based Recurrent Neural Networks (DBLSTM) to model the relationships between the PPGs and the acoustic signal. Synthesis involves input of arbitrary text to a general TTS engine (trained on any non-target speaker), the output of which is indexed by SI-ASR as PPGs. These are used by the DBLSTM to synthesize the target language in the target speaker's voice. A main advantage of this approach has very low training data requirement of the target speaker which can be in any language, as compared with a reference approach of training a special TTS engine using many recordings from the target speaker only in the target language. For a given target speaker, our proposed approach trained on 100 Mandarin (i.e. non-target language) utterances achieves comparable performance (in MOS and ABX test) of English synthetic speech as an HTS system trained on 1,000 English utterances.

Acoustic Analysis of Syllables Across Indian Languages

Anusha Prakash, Jeena J. Prakash, Hema A. Murthy; IIT Madras, India

Fri-P-1-3-9, Time: 11:00

Indian languages are broadly classified as Indo-Aryan or Dravidian. The basic set of phones is more or less the same, varying mostly in the phonotactics across languages. There has also been borrowing of sounds and words across languages over time due to intermixing of cultures. Since syllables are fundamental units of speech production and Indian languages are characterised by syllable-timed rhythm, acoustic analysis of syllables has been carried out.

In this paper, instances of common and most frequent syllables in continuous speech have been studied across six Indian languages, from both Indo-Aryan and Dravidian language groups. The distributions of acoustic features have been compared across these languages. This kind of analysis is useful for developing speech technologies in a multilingual scenario. Owing to similarities in the languages, text-to-speech (TTS) synthesisers have been developed by segmenting speech data at the phone level using hidden Markov models (HMM) from other languages as initial models. Degradation mean opinion scores and word error rates indicate that the quality of synthesised speech is comparable to that of TTSes developed by segmenting the data using language-specific HMMs.

Objective Evaluation Methods for Chinese Text-To-Speech Systems

Teng Zhang [1], Zhipeng Chen [1], Ji Wu [1], Sam Lai [2], Wenhui Lei [2], Carsten Isert [2]; [1] Tsinghua University, China; [2] BMW Group Technology Office China, China

Fri-P-1-3-10, Time: 11:00

To objectively evaluate the performance of text-to-speech (TTS) systems, many studies have been conducted in the straightforward way to compare synthesized speech and natural speech with the alignment. However, in most situations, there is no natural speech can be used. In this paper, we focus on machine learning approaches for the TTS evaluation. We exploit a subspace decomposition method to separate different components in speech, which generates distinctive acoustic features automatically. Furthermore, a pairwise based Support Vector Machine (SVM) model is used to evaluate TTS systems. With the original prosodic acoustic features and Support Vector Regression model, we obtain a ranking relevance of 0.7709. Meanwhile, with the proposed oblique matrix projection method and pairwise SVM model, we achieve a much better result of 0.9115.

Objective Evaluation Using Association Between Dimensions Within Spectral Features for Statistical Parametric Speech Synthesis

Yusuke Ijima [1], Taichi Asami [1], Hideyuki Mizuno [2]; [1] NTT, Japan; [2] Tokyo University of Science, Japan

Fri-P-1-3-11, Time: 11:00

This paper presents a novel objective evaluation technique for statistical parametric speech synthesis. One of its novel features is that it focuses on the association between dimensions within the spectral features. We first use a maximal information coefficient to analyze the relationship between subjective scores and associations of spectral features obtained from natural and various types of synthesized speech. The analysis results indicate that the scores improve as the association becomes weaker. We then describe the proposed objective evaluation technique, which uses a voice conversion method to detect the associations within spectral features. We perform subjective and objective experiments to investigate the relationship between subjective scores and objective scores. The proposed objective scores are compared to the mel-cepstral distortion. The results indicate that our objective scores achieve dramatically higher correlation to subjective scores than the mel-cepstral distortion.

A Hierarchical Predictor of Synthetic Speech Naturalness Using Neural Networks

Takenori Yoshimura [1], Gustav Eje Henter [2], Oliver Watts [2], Mirjam Wester [2], Junichi Yamagishi [2], Keiichi Tokuda [1]; [1] Nagoya Institute of Technology, Japan; [2] University of Edinburgh, UK

Fri-P-1-3-12, Time: 11:00

A problem when developing and tuning speech synthesis systems is that there is no well-established method of automatically rating the quality of the synthetic speech. This research attempts to obtain a new automated measure which is trained on the result of large-scale subjective evaluations employing many human listeners, *i.e.*, the Blizzard Challenge. To exploit the data, we experiment with linear regression, feed-forward and convolutional neural network models, and combinations of them to regress from synthetic speech to the perceptual scores obtained from listeners. The biggest improvements were seen when combining stimulus- and system-level predictions.

Text-to-Speech for Individuals with Vision Loss: A User Study

Monika Podsiadło [1], Shweta Chahar [2]; [1] Google, USA; [2] Google, UK

Fri-P-1-3-13, Time: 11:00

Individuals with vision loss use text-to-speech (TTS) for most of their interaction with devices, and rely on the quality of synthetic voices to a much larger extent than any other user group. A significant amount of local synthesis requests for Google TTS comes from TalkBack, the Android screenreader, making it our top client and making the visually-impaired users the heaviest consumers of the technology. Despite this, very little attention has been devoted to optimizing TTS voices for this user group and the feedback on TTS voices from the blind has been traditionally less-favourable. We present the findings from a TTS user experience study conducted by Google with visually-impaired screen reader users. The study comprised 14 focus groups and evaluated a total of 95 candidate voices with 90 participants across 3 countries. The study uncovered the distinctive usage patterns of this user group, which point to different TTS requirements and voice preferences from those of sighted users.

Speech Enhancement for a Noise-Robust Text-to-Speech Synthesis System Using Deep Recurrent Neural Networks

Cassia Valentini-Botinhao [1], Xin Wang [2], Shinji Takaki [2], Junichi Yamagishi [1]; [1] University of Edinburgh, UK; [2] NII, Japan

Fri-P-1-3-14, Time: 11:00

Quality of text-to-speech voices built from noisy recordings is diminished. In order to improve it we propose the use of a recurrent neural network to enhance acoustic parameters prior to training. We trained a deep recurrent neural network using a parallel database of noisy and clean acoustics parameters as input and output of the network. The database consisted of multiple speakers and diverse noise conditions. We investigated using text-derived features as an additional input of the network. We processed a noisy database of two other speakers using this network and used its output to train an HMM acoustic text-to-synthesis model for each voice. Listening experiment results showed that the voice built with enhanced parameters was ranked significantly higher than the ones trained with noisy speech and speech that has been enhanced using a conventional enhancement system. The text-derived features improved results only for the female voice, where it was ranked as highly as a voice trained with clean speech.

NOTES

Data Selection and Adaptation for Naturalness in HMM-Based Speech Synthesis

Erica Cooper, Alison Chang, Yocheved Levitan, Julia Hirschberg; Columbia University, USA

Fri-P-1-3-15, Time: 11:00

We describe experiments in building HMM text-to-speech voices on professional broadcast news data from multiple speakers. We build on earlier work comparing techniques for selecting utterances from the corpus and voice adaptation to produce the most natural-sounding voices. While our ultimate goal is to develop intelligible and natural-sounding synthetic voices in low-resource languages rapidly and without the expense of collecting and annotating data specifically for text-to-speech, we focus on English initially, in order to develop and evaluate our methods. We evaluate our approaches using crowdsourced listening tests for naturalness. We have found that removing utterances that are outliers with respect to hyper-articulation, as well as combining the selection of hypo-articulated utterances and low mean f0 utterances, produce the most natural-sounding voices.

Fri-P-1-4 : Topics in Speech Processing

Pacific Concourse – Poster D, 11:00–13:00, Friday, 9 Sept. 2016
Chair: Javier Hernando

A Portable Automatic *PA-TA-KA* Syllable Detection System to Derive Biomarkers for Neurological Disorders

Fei Tao[1], Louis Daudet[2], Christian Poellabauer[2], Sandra L. Schneider[3], Carlos Busso[1]; [1]University of Texas at Dallas, USA; [2]University of Notre Dame, USA; [3]Saint Mary's College, USA

Fri-P-1-4-1, Time: 11:00

Neurological disorders disrupt brain functions, affecting the life of many individuals. Conventional neurological disorder diagnosis methods require inconvenient and expensive devices. Several studies have identified speech biomarkers that are informative of neurological disorders, so speech-based interfaces can provide effective, convenient and affordable prescreening tools for diagnosis. We have investigated stand-alone automatic speech-based assessment tools for portable devices. Our current data collection protocol includes seven brief tests for which we have developed specialized *automatic speech recognition* (ASR) systems. The most challenging task from an ASR perspective is a popular diadochokinetic test consisting of fast repetitions of "PA-TA-KA", where subjects tend to alter, replace, insert or skip syllables. This paper presents our efforts to build a speech-based application specific for this task, where the computation is fast, efficient, and accurate on a portable device, not in the cloud. The tool recognizes the target syllables, providing phonetic alignment. This information is crucial to reliably estimate biomarkers such as the number of repetitions, insertions, mispronunciations, and temporal prosodic structure of the repetitions. We train and evaluate the application for two neurological disorders: *traumatic brain injuries* (TBIs) and Parkinson's disease. The results show low syllable error rates and high boundary detection, across populations.

Deep Neural Networks for i-Vector Language Identification of Short Utterances in Cars

Omid Ghahabi, Antonio Bonafonte, Javier Hernando, Asunción Moreno; Universitat Politècnica de Catalunya, Spain

Fri-P-1-4-2, Time: 11:00

This paper is focused on the application of the Language Identification (LID) technology for intelligent vehicles. We cope with short sentences or words spoken in moving cars in four languages: English, Spanish, German, and Finnish. As the response time of the LID system is crucial for user acceptance in this particular task, speech signals of different durations with total average of 3.8s are analyzed. In this paper, the authors propose the use of Deep Neural Networks (DNN) to model effectively the i-vector space of languages. Both raw i-vectors and session variability compensated i-vectors are evaluated as input vectors to DNNs. The performance of the proposed DNN architecture is compared with both conventional GMM-UBM and i-vector/LDA systems considering the effect of durations of signals. It is shown that the signals with durations between 2 and 3s meet the requirements of this application, i.e., high accuracy and fast decision, in which the proposed DNN architecture outperforms GMM-UBM and i-vector/LDA systems by 37% and 28%, respectively.

Improving i-Vector and PLDA Based Speaker Clustering with Long-Term Features

Abraham Woubie[1], Jordi Luque[2], Javier Hernando[1]; [1]Universitat Politècnica de Catalunya, Spain; [2]Telefónica I+D, Spain

Fri-P-1-4-3, Time: 11:00

i-vector modeling techniques have been successfully used for speaker clustering task recently. In this work, we propose the extraction of i-vectors from short- and long-term speech features, and the fusion of their PLDA scores within the frame of speaker diarization. Two sets of i-vectors are first extracted from short-term spectral and long-term voice-quality, prosodic and glottal to noise excitation ratio (GNE) features. Then, the PLDA scores of these two i-vectors are fused for speaker clustering task. Experiments have been carried out on single and multiple site scenario test sets of Augmented Multi-party Interaction (AMI) corpus. Experimental results show that i-vector based PLDA speaker clustering technique provides a significant diarization error rate (DER) improvement than GMM based BIC clustering technique.

Fri-S&T-1 : Show & Tell Session 1

Market Street Foyer, 11:00–13:00, Friday, 9 Sept. 2016
Chairs: Shiva Sundaram, Nicolas Scheffer

Open Language Interface for Voice Exploitation (OLIVE)

Aaron Lawson, Mitchell McLaren, Harry Bratt, Martin Graciarena, Horacio Franco, Christopher George, Allen Stauffer, Chris Bartels, Julien VanHout; SRI International, USA

Fri-S&T-1-1, Time: 11:00

We propose to demonstrate the Open Language Interface for Voice Exploitation (OLIVE) speech-processing system, which SRI Interna-

NOTES

tional developed under the DARPA Robust Automatic Transcription of Speech (RATS) program. The technology underlying OLIVE was designed to achieve robustness to high levels of noise and distortion for speech activity detection (SAD), speaker identification (SID), language and dialect identification (LID), and keyword spotting (KWS). Our demonstration will show OLIVE performing those four tasks. We will also demonstrate SRI's speaker recognition capability live on a mobile phone for visitors to interact with.

A Multimodal Dialogue System for Air Traffic Control Trainees Based on Discrete-Event Simulation

Luboš Šmídl, Adam Chýlek, Jan Švec; University of West Bohemia, Czech Republic

Fri-S&T-1-2, Time: 11:00

In this paper we present a multimodal dialogue system designed as a learning tool for air traffic control officer trainees (ATCO). It was developed using our discrete-event simulation dialogue management framework with cloud-based speech recognition and text-to-speech systems. Our system mimics pilots in an air traffic communication, allowing the ATCOs to practice a control of a virtual airspace using spoken commands from air traffic control English phraseology.

Lig-Aikuma: A Mobile App to Collect Parallel Speech for Under-Resourced Language Studies

Elodie Gauthier[1], David Blachon[1], Laurent Besacier[1], Guy-Noël Kouarata[2], Martine Adda-Decker[2], Annie Rialland[2], Gilles Adda[3], Grégoire Bachman[2]; [1]LIG (UMR 5217), France; [2]LPP (UMR 7018), France; [3]LIMSI, France

Fri-S&T-1-3, Time: 11:00

This paper reports on our ongoing efforts to collect speech data in under-resourced or endangered languages of Africa. Data collection is carried out using an improved version of the Android application (Aikuma) developed by Steven Bird and colleagues [1]. Features were added to the app in order to facilitate the collection of parallel speech data in line with the requirements of the French-German ANR/DFG BULB (Breaking the Unwritten Language Barrier) project. The resulting app, called Lig-Aikuma, runs on various mobile phones and tablets and proposes a range of different speech collection modes (recording, respeaking, translation and elicitation). It was used for field data collections in Congo-Brazzaville resulting in a total of over 80 hours of speech.

ARET — Automatic Reading of Educational Texts for Visually Impaired Students

Martin Grůber, Jindřich Matoušek, Zdeněk Hanzlíček, Zdeněk Krňoul, Zbyněk Zajíc; University of West Bohemia, Czech Republic

Fri-S&T-1-4, Time: 11:00

This paper deals with a presentation of an application which was developed to help in education of visually impaired pupils at a secondary school, i.e. at the pupils' age of 12 to 14 years. The web-based application integrates speech and language technologies to make the education easier in several areas, e.g. in mathematics, physics, chemistry or languages (Czech, English, German). TTS system is used for automatic reading of educational texts and it makes use of a special preprocessing of the texts, namely any formulas which may occur therein. The application is used by both teachers to create and manage the teaching material and pupils to view and listen to the prepared material. The application is currently being used by one special school for visually impaired pupils in daily lessons.

Fri-O-2-1 : New Trends in Neural Networks for Speech Recognition

Grand Ballroom A, 14:30–16:30, Friday, 9 Sept. 2016
Chairs: Arnab Ghoshal, Rohit Prabhavalkar

Segmental Recurrent Neural Networks for End-to-End Speech Recognition

Liang Lu[1], Lingpeng Kong[2], Chris Dyer[2], Noah A. Smith[3], Steve Renals[1]; [1]University of Edinburgh, UK; [2]Carnegie Mellon University, USA; [3]University of Washington, USA

Fri-O-2-1-1, Time: 14:30

We study the segmental recurrent neural network for end-to-end acoustic modelling. This model connects the segmental conditional random field (CRF) with a recurrent neural network (RNN) used for feature extraction. Compared to most previous CRF-based acoustic models, it does not rely on an external system to provide features or segmentation boundaries. Instead, this model marginalises out all the possible segmentations, and features are extracted from the RNN trained together with the segmental CRF. Essentially, this model is self-contained and can be trained end-to-end. In this paper, we discuss practical training and decoding issues as well as the method to speed up the training in the context of speech recognition. We performed experiments on the TIMIT dataset. We achieved 17.3% phone error rate (PER) from the first-pass decoding — the best reported result using CRFs, despite the fact that we only used a zeroth-order CRF and without using any language model.

Acoustic Modeling Using Bidirectional Gated Recurrent Convolutional Units

Markus Nussbaum-Thom, Jia Cui, Bhuvana Ramabhadran, Vaibhava Goel; IBM, USA

Fri-O-2-1-2, Time: 14:50

Convolutional and bidirectional recurrent neural networks have achieved considerable performance gains as acoustic models in automatic speech recognition in recent years. Latest architectures unify long short-term memory, gated recurrent unit and convolutional neural networks by stacking these different neural network types on each other, and providing short and long-term features to different depth levels of the network.

For the first time, we propose a unified layer for acoustic modeling which is simultaneously recurrent and convolutional, and which operates only on short-term features. Our unified model introduces a bidirectional gated recurrent unit that uses convolutional operations for the gating units. We analyze the performance behavior of the proposed layer, compare and combine it with bidirectional gated recurrent units, deep neural networks and frequency-domain convolutional neural networks on a 50 hour English broadcast news task. The analysis indicates that the proposed layer in combination with stacked bidirectional gated recurrent units outperforms other architectures.

Notes

Exploiting Depth and Highway Connections in Convolutional Recurrent Deep Neural Networks for Speech Recognition

Wei-Ning Hsu, Yu Zhang, Ann Lee, James Glass; MIT, USA

Fri-O-2-1-3, Time: 15:10

Deep neural network models have achieved considerable success in a wide range of fields. Several architectures have been proposed to alleviate the vanishing gradient problem, and hence enable training of very deep networks. In the speech recognition area, convolutional neural networks, recurrent neural networks, and fully connected deep neural networks have been shown to be complimentary in their modeling capabilities. Combining all three components, called CLDNN, yields the best performance to date. In this paper, we extend the CLDNN model by introducing a highway connection between LSTM layers, which enables direct information flow from cells of lower layers to cells of upper layers. With this design, we are able to better exploit the advantages of a deeper structure. Experiments on the GALE Chinese Broadcast Conversation/News Speech dataset indicate that our model outperforms all previous models and achieves a new benchmark, which is 22.41% character error rate on the dataset.

Stimulated Deep Neural Network for Speech Recognition

Chunyang Wu[1], Penny Karanasou[1], Mark J.F. Gales[1], Khe Chai Sim[2]; [1]University of Cambridge, UK; [2]NUS, Singapore

Fri-O-2-1-4, Time: 15:30

Deep neural networks (DNNs) and deep learning approaches yield state-of-the-art performance in a range of tasks, including speech recognition. However, the parameters of the network are hard to analyze, making network regularization and robust adaptation challenging. Stimulated training has recently been proposed to address this problem by encouraging the node activation outputs in regions of the network to be related. This kind of information aids visualization of the network, but also has the potential to improve regularization and adaptation. This paper investigates stimulated training of DNNs for both of these options. These schemes take advantage of the smoothness constraints that stimulated training offers. The approaches are evaluated on two large vocabulary speech recognition tasks: a U.S. English broadcast news (BN) task and a Javanese conversational telephone speech task from the IARPA Babel program. Stimulated DNN training acquires consistent performance gains on both tasks over unstimulated baselines. On the BN task, the proposed smoothing approach is also applied to rapid adaptation, again outperforming the standard adaptation scheme.

Phonetic Context Embeddings for DNN-HMM Phone Recognition

Leonardo Badino; Istituto Italiano di Tecnologia, Italy

Fri-O-2-1-5, Time: 15:50

This paper proposes an approach, named phonetic context embedding, to model phonetic context effects for deep neural network - hidden Markov model (DNN-HMM) phone recognition. Phonetic context embeddings can be regarded as continuous and distributed vector representations of context-dependent phonetic units (e.g., triphones). In this work they are computed using neural networks. First, all phone labels are mapped into vectors of binary distinctive features (DFs, e.g., nasal/not-nasal). Then for each speech frame the corresponding DF vector is concatenated with DF vectors of previous and next frames and fed into a neural network that is trained to estimate the acoustic coefficients (e.g., MFCCs) of that frame. The values of the first hidden layer represent the embedding of the input DF vectors. Finally, the resulting embeddings are used as secondary task targets in a multi-task learning (MTL) setting when training the DNN that computes phone state posteriors. The approach allows to easily encode a much larger context than alternative MTL-based approaches. Results on TIMIT with a fully connected DNN shows phone error rate (PER) reductions from 22.4% to 21.0% and from 21.3% to 19.8% on the test core and the validation set respectively and lower PER than an alternative strong MTL approach.

Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks

Ying Zhang, Mohammad Pezeshki, Philémon Brakel, Saizheng Zhang, César Laurent, Yoshua Bengio, Aaron Courville; Université de Montréal, Canada

Fri-O-2-1-6, Time: 16:10

Convolutional Neural Networks (CNNs) are effective models for reducing spectral variations and modeling spectral correlations in acoustic features for automatic speech recognition (ASR). Hybrid speech recognition systems incorporating CNNs with Hidden Markov Models/Gaussian Mixture Models (HMMs/GMMs) have achieved the state-of-the-art in various benchmarks. Meanwhile, Connectionist Temporal Classification (CTC) with Recurrent Neural Networks (RNNs), which is proposed for labeling unsegmented sequences, makes it feasible to train an 'end-to-end' speech recognition system instead of hybrid settings. However, RNNs are computationally expensive and sometimes difficult to train. In this paper, inspired by the advantages of both CNNs and the CTC approach, we propose an end-to-end speech framework for sequence labeling, by combining hierarchical CNNs with CTC directly without recurrent connections. By evaluating the approach on the TIMIT phoneme recognition task, we show that the proposed model is not only computationally efficient, but also competitive with the existing baseline systems. Moreover, we argue that CNNs have the capability to model temporal correlations with appropriate context information.

Fri-O-2-2 : Special Session: The RedDots Challenge: Towards Characterizing Speakers from Short Utterances

Grand Ballroom BC, 14:30–16:30, Friday, 9 Sept. 2016

Chairs: Kong Aik Lee, Anthony Larcher, Hagai Aronowitz, Guangsen Wang, Patrick Kenny

Joint Speaker and Lexical Modeling for Short-Term Characterization of Speaker

*Guangsen Wang, Kong Aik Lee, Trung Hieu Nguyen, Hanwu Sun, Bin Ma; A*STAR, Singapore*

Fri-O-2-2-1, Time: 14:30

For speech utterances of very short duration, speaker characterization has shown strong dependency on the lexical content. In this context, speaker verification is always performed by analyzing and matching speaker pronunciation of individual words, syllables,

NOTES

or phones. In this paper, we advocate the use of hidden Markov model (HMM) for joint modeling of speaker characteristic and lexical content. We then develop a scoring model that scores only the speaker part rather than the joint speaker-lexical component leading to a better speaker verification performance. Experiments were conducted on the text-prompted task of RSR2015 and the RedDots datasets. In the RSR2015, the prompted texts are limited to random sequences of digits. The RedDots dataset dictates an unconstrained scenario where the prompted texts are free-text sentences. Both RSR2015 and RedDots datasets are publicly available.

Tandem Features for Text-Dependent Speaker Verification on the RedDots Corpus

Md Jahangir Alam, Patrick Kenny, Vishwa Gupta; CRIM, Canada

Fri-O-2-2-2, Time: 14:45

We use tandem features and a fusion of four systems for text-dependent speaker verification on the RedDots corpus. In the tandem system, a senone-discriminant neural network provides a low-dimensional bottleneck feature at each frame which are concatenated with a standard Mel-frequency cepstral coefficients (MFCC) feature representation. The concatenated features are propagated to a conventional GMM/UBM speaker recognition framework. In order to capture complementary information to the MFCC, we also use linear frequency cepstral coefficients and wavelet-based cepstral coefficients features for score level fusion. We report results on the part 1 and part 4 (text-dependent) tasks of RedDots corpus. Both the tandem feature-based system and fused system provided significant improvements over the baseline GMM/UBM system in terms of equal error rates (EER) and detection cost functions (DCFs) as defined in the 2008 and 2010 NIST speaker recognition evaluations. On the part 1 task (impostor correct condition) the fused system reduced the EER from 2.63% to 2.28% for male trials and from 7.01% to 3.48% for female trials. On the part4 task (impostor correct condition) the fused system helped to reduce the EER from 2.49% to 1.96% and from 5.9% to 3.22% for male and female trials respectively.

Text Dependent Speaker Verification Using Un-Supervised HMM-UBM and Temporal GMM-UBM

Achintya Kr. Sarkar, Zheng-Hua Tan; Aalborg University, Denmark

Fri-O-2-2-3, Time: 15:00

In this paper, we investigate the Hidden Markov Model (HMM) and the temporal Gaussian Mixture Model (GMM) systems based on the Universal Background Model (UBM) concept to capture temporal information of speech for Text Dependent (TD) Speaker Verification (SV). In TD-SV, target speakers are constrained to use only predefined fixed sentence/s during both the enrollment and the test process. The temporal information is therefore important in the sense of utterance verification, i.e. whether the test utterance has the same sequence of textual content as the utterance used during the target enrollment. However, the temporal information is not considered in the classical GMM-UBM based TD-SV system. Moreover, no transcription knowledge of the speech is required in the HMM-UBM and temporal GMM-UBM based systems. We also study the fusion of the HMM-UBM, the temporal GMM-UBM and the classical GMM-UBM systems in SV. We show that the HMM-UBM system yields better performance than the other systems in most cases. Further, fusion of the systems improve the overall speaker verification performance. The results are shown in the different tasks of the RedDots challenge 2016 database.

Utterance Verification for Text-Dependent Speaker Recognition: A Comparative Assessment Using the RedDots Corpus

Tomi Kinnunen [1], Md. Sahidullah [1], Ivan Kukanov [1], Héctor Delgado [2], Massimiliano Todisco [2], Achintya Kr. Sarkar [3], Nicolai Bæk Thomsen [3], Ville Hautamäki [1], Nicholas Evans [2], Zheng-Hua Tan [3]; [1] University of Eastern Finland, Finland; [2] EURECOM, France; [3] Aalborg University, Denmark

Fri-O-2-2-4, Time: 15:15

Text-dependent automatic speaker verification naturally calls for the simultaneous verification of speaker identity and spoken content. These two tasks can be achieved with automatic speaker verification (ASV) and utterance verification (UV) technologies. While both have been addressed previously in the literature, a treatment of simultaneous speaker and utterance verification with a modern, standard database is so far lacking. This is despite the burgeoning demand for voice biometrics in a plethora of practical security applications. With the goal of improving overall verification performance, this paper reports different strategies for simultaneous ASV and UV in the context of short-duration, text-dependent speaker verification. Experiments performed on the recently released RedDots corpus are reported for three different ASV systems and four different UV systems. Results show that the combination of utterance verification with automatic speaker verification is (almost) universally beneficial with significant performance improvements being observed.

Parallel Speaker and Content Modelling for Text-Dependent Speaker Verification

Jianbo Ma, Saad Irtza, Kaavya Sriskandaraja, Vidhyasaharan Sethu, Eliathamby Ambikairajah; University of New South Wales, Australia

Fri-O-2-2-5, Time: 15:30

Text-dependent short duration speaker verification involves two challenges. The primary challenge of interest is the verification of the speaker's identity, and often a secondary challenge of interest is the verification of the lexical content of the pass-phrase. In this paper, we propose the use of two systems to handle these two tasks in parallel with one sub-system modelling speaker identity based on the assumption that lexical content is known and the other sub-system modelling lexical content in a speaker dependent manner. The text-dependent speaker verification sub-system is based on hidden Markov models and the lexical content verification system is based on models of speech segments that use a distinct Gaussian mixture model for each segment. Furthermore, a mixture selection method based on KL divergence was applied to refine the lexical content sub-system by making the models more discriminative. Experiments on part 1 of the RedDots database showed that the proposed combination of two sub-systems outperformed the baseline system by 39.8%, 51.1% and 37.3% in terms of the 'imposter_correct', 'target_wrong' and 'imposter_wrong' metrics respectively.

NOTES

i-Vector/HMM Based Text-Dependent Speaker Verification System for RedDots Challenge

Hossein Zeinali[1], Hossein Sameti[1], Lukáš Burget[2], Jan Černocký[2], Nooshin Maghsoodi[1], Pavel Matějka[2]; [1]Sharif University of Technology, Iran; [2]Brno University of Technology, Czech Republic

Fri-O-2-2-6, Time: 15:45

Recently, a new data collection was initiated within the RedDots project in order to evaluate text-dependent and text-prompted speaker recognition technology on data from a wider speaker population and with more realistic noise, channel and phonetic variability. This paper analyses our systems built for RedDots challenge — the effort to collect and compare the initial results on this new evaluation data set obtained at different sites. We use our recently introduced HMM based i-vector approach, where, instead of the traditional GMM, a set of phone specific HMMs is used to collect the sufficient statistics for i-vector extraction. Our systems are trained in a completely phrase-independent way on the data from RSR2015 and Libri speech databases. We compare systems making use of standard cepstral features and their combination with neural network based bottle-neck features. The best results are obtained with a score-level fusion of such systems.

Exploring Session Variability and Template Aging in Speaker Verification for Fixed Phrase Short Utterances

Rohan Kumar Das, Sarfaraz Jelil, S.R. Mahadeva Prasanna; IIT Guwahati, India

Fri-O-2-2-7, Time: 16:00

This work highlights the impact of session variability and template aging on speaker verification (SV) using fixed phrase short utterances from the RedDots database. These have been collected over a period of one year and contain a large number of sessions per speaker. Session variation has been found to have a direct influence on SV performance and its significance is even greater for the case of fixed phrase short utterances as a very small amount of speech data is involved for speaker modeling as well as testing. Similarly for a practical deployable SV system when there is large session variation involved over a period of time, the template aging of the speakers may effect the SV performance. This work attempts to address some issues related to session variability and template aging of speakers which are found for data having large session variability, that if considered can be utilized for improving the performance of an SV system.

Fri-O-2-3 : Articulatory Measurements and Analysis

Bayview A, 14:30–16:30, Friday, 9 Sept. 2016
Chairs: Jianwu Dang, Adam Lammert

Prediction of the Articulatory Movements of Unseen Phonemes of a Speaker Using the Speech Structure of Another Speaker

Hidetsugu Uchida, Daisuke Saito, Nobuaki Minematsu; University of Tokyo, Japan

Fri-O-2-3-1, Time: 14:30

In this paper, we propose a method to predict the articulatory movements of phonemes that are difficult for a speaker to pronounce correctly because those phonemes are not seen in the native language of that speaker. When one wants to predict the articulatory movements of those unseen phonemes, since he/she has difficulty to generate those sounds, the conventional acoustic-to-articulatory mapping cannot be applied as it is. Here, we propose a solution by using the speech structure of another reference speaker who can pronounce the unseen phonemes. Speech structure is a kind of speech feature that represents only the linguistic information by suppressing the non-linguistic information, e.g. speaker identity, of an input utterance. In the proposed method, by using the speech structure of those unseen phonemes and other phonemes as constraint, the articulatory movements of the unseen phonemes are searched for in the articulatory space of the original speaker. Experiments using English short vowels show that the averaged prediction error was 1.02 mm.

Vocal Tract Length Normalization for Speaker Independent Acoustic-to-Articulatory Speech Inversion

Ganesh Sivaraman[1], Vikramjit Mitra[2], Hosung Nam[3], Mark Tiede[4], Carol Espy-Wilson[1]; [1]University of Maryland, USA; [2]SRI International, USA; [3]Korea University, Korea; [4]Haskins Laboratories, USA

Fri-O-2-3-2, Time: 14:50

Speech inversion is a well-known ill-posed problem and addition of speaker differences typically makes it even harder. This paper investigates a vocal tract length normalization (VTLN) technique to transform the acoustic space of different speakers to a target speaker space such that speaker specific details are minimized. The speaker normalized features are then used to train a feed-forward neural network based acoustic-to-articulatory speech inversion system. The acoustic features are parameterized as time-contextualized mel-frequency cepstral coefficients and the articulatory features are represented by six tract-variable (TV) trajectories. Experiments are performed with ten speakers from the U. Wisc. X-ray microbeam database. Speaker dependent speech inversion systems are trained for each speaker as baselines to compare the performance of the speaker independent approach. For each target speaker, data from the remaining nine speakers are transformed using the proposed approach and the transformed features are used to train a speech inversion system. The performances of the individual systems are compared using the correlation between the estimated and the actual TVs on the target speaker's test set. Results show that the

NOTES

proposed speaker normalization approach provides a 7% absolute improvement in correlation as compared to the system where speaker normalization was not performed.

Investigation of Speed-Accuracy Tradeoffs in Speech Production Using Real-Time Magnetic Resonance Imaging

Adam C. Lammert[1], Christine H. Shadle[2], Shrikanth S. Narayanan[3], Thomas F. Quatieri[1]; [1]MIT Lincoln Laboratory, USA; [2]Haskins Laboratories, USA; [3]University of Southern California, USA

Fri-O-2-3-3, Time: 15:10

Motor actions in speech production are both rapid and highly dexterous, even though speed and accuracy are often thought to conflict. Fitts' law has served as a rigorous formulation of the fundamental speed-accuracy tradeoff in other domains of human motor action, but has not been directly examined with respect to speech production. This paper examines Fitts' law in speech articulation kinematics by analyzing USC-TIMIT, a large database of real-time magnetic resonance imaging data of speech production. This paper also addresses methodological challenges in applying Fitts-style analysis, including the definition and operational measurement of key variables in real-time MRI data. Results suggest high variability in the task demands associated with targeted articulatory kinematics, as well as a clear tradeoff between speed and accuracy for certain types of speech production actions. Consonant targets, and particularly those following vowels, show the strongest evidence of this tradeoff, with correlations as high as 0.71 between movement time and difficulty. Other speech actions seem to challenge Fitts' law. Results are discussed with respect to limitations of Fitts' law in the context of speech production, as well as future improvements and applications.

Characterizing Vocal Tract Dynamics Across Speakers Using Real-Time MRI

Tanner Sorensen, Asterios Toutios, Louis Goldstein, Shrikanth S. Narayanan; University of Southern California, USA

Fri-O-2-3-4, Time: 15:30

Real-time magnetic resonance imaging (rtMRI) provides information about the dynamic shaping of the vocal tract during speech production and valuable data for creating and testing models of speech production. In this paper, we use rtMRI videos to develop a dynamical system in the framework of Task Dynamics which controls vocal tract constrictions and induces deformation of the air-tissue boundary. This is the first task dynamical system explicitly derived from speech kinematic data. Simulation identifies differences in articulatory strategy across speakers ($n = 18$), specifically in the relative contribution of articulators to vocal tract constrictions.

Tracking Contours of Orofacial Articulators from Real-Time MRI of Speech

Mathieu Labrunie[1], Pierre Badin[1], Dirk Voit[2], Arun A. Joseph[2], Laurent Lamalle[3], Coriandre Vilain[1], Louis-Jean Boë[1], Jens Frahm[2]; [1]GIPSA, France; [2]BiomedNMR, Germany; [3]IRMaGe, France

Fri-O-2-3-5, Time: 15:50

We introduce a method for predicting midsagittal contours of orofacial articulators from real-time MRI data. A corpus of about 26 minutes of speech has been recorded of a French speaker at a rate of 55 images / s using highly undersampled radial gradient-echo MRI with image reconstruction by nonlinear inversion. The contours of each articulator have been manually traced for a set of about 60 images selected — by hierarchical clustering — to optimally represent the diversity of the speaker articulations. The data serve to build articulator-specific Principal Component Analysis (PCA) models of contours and associated image intensities, as well as multilinear regression (MLR) models that predict contour parameters from image parameters. The contours obtained by MLR are then refined, using the local information about pixel intensity profiles along the contours' normals, by means of modified Active Shape Models (ASM) trained on the same data. The method reaches RMS of predicted points to reference contour distances between 0.54 and 0.93 mm, depending on articulators. The processing of the corpus demonstrated the efficiency of the procedure, despite the possibility of further improvements. This work opens new perspectives for studying articulatory motion in speech.

State-of-the-Art MRI Protocol for Comprehensive Assessment of Vocal Tract Structure and Function

Sajan Goud Lingala[1], Asterios Toutios[1], Johannes Töger[1], Yongwan Lim[1], Yinghua Zhu[1], Yoon-Chul Kim[2], Colin Vaz[1], Shrikanth S. Narayanan[1], Krishna S. Nayak[1]; [1]University of Southern California, USA; [2]Samsung Medical Center, Korea

Fri-O-2-3-6, Time: 16:10

Magnetic Resonance Imaging (MRI) provides a safe and flexible means to study the vocal tract, and is increasingly used in speech production research. This work details a state-of-the-art MRI protocol for comprehensive assessment of vocal tract structure and function, and presents results from representative speakers. The system incorporates (a) custom upper airway coils that are maximally sensitive to vocal tract tissues, (b) graphical user interface for 2D real-time MRI that provides on-the-fly reconstruction for interactive localization, and correction of imaging artifacts, (c) off-line constrained reconstruction for generating high spatio-temporal resolution dynamic images at (83 frames per sec, 2.4 mm^2), (d) 3D static imaging of sounds sustained for 7 sec with full vocal tract coverage and isotropic resolution (resolution: 1.25 mm^3), (e) T2-weighted high-resolution, high-contrast depiction of soft-tissue boundaries of the full vocal tract (axial, coronal, sagittal sweeps with resolution: $0.58 \times 0.58 \times 3$ mm^3), and (f) simultaneous audio recording with off-line noise cancellation and temporal alignment of audio with 2D real-time MRI. A stimuli set was designed to capture efficiently salient, static and dynamic, articulatory and morphological aspects of speech production in 90-minute data acquisition sessions.

NOTES

Fri-O-2-4 : Automatic Assessment of Emotions

Bayview B, 14:30–16:30, Friday, 9 Sept. 2016
Chairs: Carlos Busso, Elizabeth Shriberg

DBN-ivector Framework for Acoustic Emotion Recognition

Rui Xia, Yang Liu; University of Texas at Dallas, USA

Fri-O-2-4-1, Time: 14:30

Deep learning and i-vectors have been successfully used in speech and speaker recognition recently. In this work we propose a framework based on deep belief network (DBN) and i-vector space modeling for acoustic emotion recognition. We use two types of labels for frame level DBN training. The first one is the vector of posterior probabilities calculated from the GMM universal background model (UBM). The second one is the predicted label based on the GMMs. The DBN is trained to minimize errors for both types. After DBN training, we use the vector of posterior probabilities estimated by DBN to replace the UBM for i-vector extraction. Finally the extracted i-vectors are used in backend classifiers for emotion recognition. Our experiments on the USC IEMOCAP data show the effectiveness of our proposed DBN-ivector framework. In particular, with decision level combination, our proposed system yields significant improvement on both unweighted and weighted accuracy.

An Investigation of Emotional Speech in Depression Classification

Brian Stasak[1], Julien Epps[1], Nicholas Cummins[1], Roland Goecke[2]; [1]University of New South Wales, Australia; [2]University of Canberra, Australia

Fri-O-2-4-2, Time: 14:50

Assessing depression via speech characteristics is a growing area of interest in quantitative mental health research with a view to a clinical mental health assessment tool. As a mood disorder, depression induces changes in response to emotional stimuli, which motivates this investigation into the relationship between emotion and depression affected speech. This paper investigates how emotional information expressed in speech (i.e. arousal, valence, dominance) contributes to the classification of minimally depressed and moderately-severely depressed individuals. Experiments based on a subset of the AVEC 2014 database show that manual emotion ratings alone are discriminative of depression and combining rating-based emotion features with acoustic features improves classification between mild and severe depression. Emotion-based data selection is also shown to provide improvements in depression classification and a range of threshold methods are explored. Finally, the experiments presented demonstrate that automatically predicted emotion ratings can be incorporated into a fully automatic depression classification to produce a 5% accuracy improvement over an acoustic-only baseline system.

Retrieving Categorical Emotions Using a Probabilistic Framework to Define Preference Learning Samples

Reza Lotfian, Carlos Busso; University of Texas at Dallas, USA

Fri-O-2-4-3, Time: 15:10

Preference learning is an appealing approach for affective recognition. Instead of predicting the underlying emotional class of a sample, this framework relies on pairwise comparisons to rank-order the testing data according to an emotional dimension. This framework is relevant not only for continuous attributes such as arousal or valence, but also for categorical classes (e.g., is this sample happier than the other?). A preference learning system for categorical classes can have applications in several domains including retrieving emotional behaviors conveying a target emotion, and defining the emotional intensity associated with a given class. One important challenge to build such a system is to define relative labels defining the preference between training samples. Instead of building these labels from scratch, we propose a probabilistic framework that creates relative labels from existing categorical annotations. The approach considers individual assessments instead of consensus labels, creating a metrics that is sensitive to the underlying ambiguity of emotional classes. The proposed metric quantifies the likelihood that a sample belong to a target emotion. We build *happy, angry* and *sad* rank-classifiers using this metric. We evaluate the approach over cross-corpus experiments, showing improved performance over binary classifiers and rank-based classifiers trained with consensus labels.

At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions in Speech

Maximilian Schmitt, Fabien Ringeval, Björn Schuller; Universität Passau, Germany

Fri-O-2-4-4, Time: 15:30

Recognition of natural emotion in speech is a challenging task. Different methods have been proposed to tackle this complex task, such as acoustic feature brute-forcing or even end-to-end learning. Recently, bag-of-audio-words (BoAW) representations of acoustic low-level descriptors (LLDs) have been employed successfully in the domain of acoustic event classification and other audio recognition tasks. In this approach, feature vectors of acoustic LLDs are quantised according to a learnt codebook of audio words. Then, a histogram of the occurring 'words' is built. Despite their massive potential, BoAW have not been thoroughly studied in emotion recognition. Here, we propose a method using BoAW created only of mel-frequency cepstral coefficients (MFCCs). Support vector regression is then used to predict emotion continuously in time and value, such as in the dimensions arousal and valence. We compare this approach with the computation of functionals based on the MFCCs and perform extensive evaluations on the RECOLA database, which features spontaneous and natural emotions. Results show that, BoAW representation of MFCCs does not only perform significantly better than functionals, but also outperforms by far most of recently published deep learning approaches, including convolutional and recurrent networks.

Speech Emotion Recognition Using Affective Saliency

Arodami Chorianopoulou[1], Polychronis Koutsakis[2], Alexandros Potamianos[3]; [1]Technical University of Crete, Greece; [2]Murdoch University, Australia; [3]NTUA, Greece

Fri-O-2-4-5, Time: 15:50

We investigate an affective saliency approach for speech emotion recognition of spoken dialogue utterances that estimates the amount of emotional information over time. The proposed saliency approach uses a regression model that combines features extracted from the acoustic signal and the posteriors of a segment-level classifier to obtain frame or segment-level ratings. The affective saliency model is trained using a minimum classification error (MCE) criterion that learns the weights by optimizing an objective loss function related to the classification error rate of the emotion recognition system. Affective saliency scores are then used to weight the contribution of frame-level posteriors and/or features to the speech emotion classification decision. The algorithm is evaluated for the task of anger detection on four call-center datasets for two languages, Greek and English, with good results.

Laughter Valence Prediction in Motivational Interviewing Based on Lexical and Acoustic Cues

Rahul Gupta [1], Nishant Nath [1], Taruna Agrawal [1], Panayiotis Georgiou [1], David C. Atkins [2], Shrikanth S. Narayanan [1]; [1] University of Southern California, USA; [2] University of Washington, USA

Fri-O-2-4-6, Time: 16:10

Motivational Interviewing (MI) is a goal oriented psychotherapy counseling that aims to instill positive change in a client through discussion. Since the discourse is in the form of semi-structured natural conversation, it often involves a variety of non-verbal social and affective behaviors such as laughter. Laughter carries information related to affect, mood and personality and can offer a window into the mental state of a person. In this work, we conduct an analytical study on predicting the valence of laughters (positive, neutral or negative) based on lexical and acoustic cues, within the context of MI. We hypothesize that the valence of laughter can be predicted using a window of past and future context around the laughter and, design models to incorporate context, from both text and audio. Through these experiments we validate the relation of the two modalities to perceived laughter valence. Based on the outputs of the prediction experiment, we perform a follow up analysis of the results including: (i) identification of the optimal past and future context in the audio and lexical channels, (ii) investigation of the differences in the prediction patterns for the counselor and the client and, (iii) analysis of feature patterns across the two modalities.

Fri-O-2-5 : Acoustic and Articulatory Phonetics

Seacliff BCD, 14:30–16:30, Friday, 9 Sept. 2016
Chairs: Ailbhe Ni Chasaide, Mattias Heldner

Respiratory Belts and Whistles: A Preliminary Study of Breathing Acoustics for Turn-Taking

Marcin Włodarczak, Mattias Heldner; Stockholm University, Sweden

Fri-O-2-5-1, Time: 14:30

This paper presents first results on using acoustic intensity of inhalations as a cue to speech initiation in spontaneous multiparty conversations. We demonstrate that inhalation intensity significantly differentiates between cycles coinciding with no speech activity, shorter (< 1 s) and longer stretches of speech. While the model fit is relatively weak, it is comparable to the fit of a model using kinematic features collected with Respiratory Inductance Plethysmography. We also show that incorporating both kinematic and acoustic features further improves the model. Given the ease of capturing breath acoustics, we consider the results to be a promising first step towards studying communicative functions of respiratory sounds. We discuss possible extensions to the data collection procedure with a view to improving predictive power of the model.

/r/ as Language Marker in Bilingual Speech Production and Perception

Constantijn Kaland, Vincenzo Galatà, Lorenzo Spreafico, Alessandro Vietti; Libera Università di Bolzano, Italy

Fri-O-2-5-2, Time: 14:50

Across languages of the world /r/ is known for its variability. Recent literature incorporates sociolinguistic factors, such as bilingualism, in order to explain /r/ variation. The current study investigates to what extent /r/ is a marker of a bilingual's dominant language. Specifically, the effects of several sociolinguistic and phonotactic factors on the production and perception of /r/ are investigated, such as the bilingual speaker's linguistic background, the language spoken as well as syllable position and place of articulation. To this end a reading task is carried out with bilingual speakers from South Tyrol (Italy). The major languages spoken in this region are Tyrolean (German dialect) and Italian. The recorded reading data is subsequently used in a perception experiment to investigate whether South Tyrolean listeners can identify the dominant language of the speaker on the basis of the presence of /r/ and the /r/ variant. Results show that listeners can identify the dominant language of the bilingual speakers on the basis of /r/. Specifically, the more Italian dominant the sociolinguistic background of the speaker, the more /r/ is produced frontally and the more that speaker is perceived as Italian dominant.

Evaluation of Phonatory Behavior of German and French Speakers in Native and Non-Native Speech

Manfred Pützer [1], Frank Zimmerer [1], Wolfgang Wokurek [2], Jeanin Jügler [1]; [1] Universität des Saarlandes, Germany; [2] Universität Stuttgart, Germany

Fri-O-2-5-3, Time: 15:10

Phonatory behavior of German speakers (GS) and French speakers (FS) in native (L1) and non-native (L2) speech was instrumentally examined. Vowel productions of the two groups were analyzed using a parametrization of phonatory behavior and phonatory quality properties in the acoustic signal. The behavior of GS is characterized by more strained adduction of the vocal folds whereas FS show more incomplete glottal closure. Furthermore, GS change their phonatory behavior in the foreign language (=French) by adapting phonatory strategies of FS, whereas FS do not show this tendency. In addition, German beginners (BEG) and partly German advanced learners (ADV) are already orientated on production characteristics of the L2. French BEG however retain their phonatory behavior in L2 (=German) by showing less vocal fold adduction in comparison to their L1. French ADV show the opposite behavior. Finally, ADV of the two speaker groups generally show more strained behavior in L2 productions than BEG. The results provide evidence that GS and FS apply different laryngeal phonatory settings and that they altered their settings in L2 differently. Perceptual evaluation of voice quality of the speech material and a correlation analysis between acoustic and perceptual results are suggested for future research.

Notes

Today's Most Frequently Used F_0 Estimation Methods, and Their Accuracy in Estimating Male and Female Pitch in Clean Speech

Sofia Strömbergsson; Karolinska Institute, Sweden

Fri-O-2-5-4, Time: 15:30

Variation in fundamental frequency (F_0) constitutes a valuable source of information for researches across many disciplines, with a shared interest in speech. Different methods for estimating F_0 vary in estimation accuracy and accessibility, and there is yet no gold standard. Through a bibliometric survey, this study examines what methods were the most frequently used in the speech scientific community during the years 2010–2016. Secondly, the most used methods are evaluated against a ground truth reference, with a specific focus on their accuracy in estimating F_0 in male and female speakers, respectively.

The results show that Praat is the dominant method by far, followed by STRAIGHT, RAPT and YIN. This pattern holds across a range of different research areas, although within Acoustics and Engineering, Praat's dominance is less pronounced. In the evaluation including Praat, RAPT and YIN — with their default and gender-adapted settings — Praat also proved to be the most accurate. The finding that adapting Praat's pitch range settings by gender leads to further improvements should encourage researchers to do this routinely.

A Praat-Based Algorithm to Extract the Amplitude Envelope and Temporal Fine Structure Using the Hilbert Transform

Lei He, Volker Dellwo; Universität Zürich, Switzerland

Fri-O-2-5-5, Time: 15:50

A speech signal can be viewed as a high frequency carrier signal containing the temporal fine structure (TFS) that is modulated by a low frequency envelope (ENV). A widely used method to decompose a speech signal into the TFS and ENV is the Hilbert transform. Although this method has been available for about one century and is widely applied in various kinds of speech processing tasks (e.g. speech chimeras), there are only very few speech processing packages that contain readily available functions for the Hilbert transform, and there is very little textbook type literature tailored for speech scientists to explain the processes behind the transform. With this paper we provide the code for carrying out the Hilbert operation to obtain the TFS and ENV in the widely used speech processing software Praat, and explain the basics of the procedure. To verify our code, we compare the Hilbert transform in Praat with a widely applied function for the same purpose in MATLAB ("hilbert(...)"). We can confirm that both methods arrive at identical outputs.

Likelihood Ratio Calculation in Acoustic-Phonetic Forensic Voice Comparison: Comparison of Three Statistical Modelling Approaches

Ewald Enzinger; University of New South Wales, Australia

Fri-O-2-5-6, Time: 16:10

This study compares three statistical models used to calculate likelihood ratios in acoustic-phonetic forensic-voice-comparison systems: Multivariate kernel density, principal component analysis kernel density, and a multivariate normal model. The data were coefficient values obtained from discrete cosine transforms fitted to human-supervised formant-trajectory measurements of tokens of /iau/ from a database of recordings of 60 female speakers of Chinese. Tests were conducted using high-quality recordings as nominal suspect samples and mobile-to-landline transmitted recordings as nominal offender samples. Performance was assessed before and after fusion with a baseline automatic mel frequency cepstral coefficient Gaussian mixture model universal background model system. In addition, Monte Carlo simulations were used to compare the output of the statistical models to true likelihood-ratio values calculated on the basis of the distribution specified for a simulated population.

Fri-O-2-6 : Source Separation and Spatial Audio

Seacliff A, 14:30–16:30, Friday, 9 Sept. 2016
Chair: Emmanuel Vincent

A Sparse Spherical Harmonic-Based Model in Subbands for Head-Related Transfer Functions

Xiaoke Qi, Jianhua Tao; Chinese Academy of Sciences, China

Fri-O-2-6-1, Time: 14:30

Several functional models for head-related transfer function (HRTF) have been proposed based on spherical harmonic (SH) orthogonal functions, which yield an encouraging performance level in terms of log-spectral distortion (LSD). However, since the properties of subbands are quite different and highly subject-dependent, the degree of SH expansion should be adapted to the subband and the subject, which is quite challenging. In this paper, a sparse spherical harmonic-based model termed SSHM is proposed in order to achieve an intelligent frequency truncation. Different from SH-based model (SHM) which assigns the degree for each subband, SSHM constrains the number of SH coefficients by using an l_1 penalty, and automatically preserves the significant coefficients in each subband. As a result, SSHM requires less coefficients at the same SD level than other truncation methods to reconstruct HRTFs . Furthermore, when used for interpolation, SSHM gives a better fitting precision since it naturally reduces the influence of the fluctuation caused by the movement of the subject and the processing error. The experiments show that even using about 40% less coefficients, SSHM has a slightly lower LSD than SHM. Therefore, SSHM can achieve a better tradeoff between efficiency and accuracy.

Single-Channel Multi-Speaker Separation Using Deep Clustering

Yusuf Isik, Jonathan Le Roux, Zhuo Chen, Shinji Watanabe, John R. Hershey; MERL, USA

Fri-O-2-6-2, Time: 14:50

Deep clustering is a recently introduced deep learning architecture that uses discriminatively trained embeddings as the basis for clustering. It was recently applied to spectrogram segmentation, resulting in impressive results on speaker-independent multi-speaker separation. In this paper we extend the baseline system with an end-to-end signal approximation objective that greatly improves performance on a challenging speech separation. We first significantly improve upon the baseline system performance by

NOTES

incorporating better regularization, larger temporal context, and a deeper architecture, culminating in an overall improvement in signal to distortion ratio (SDR) of 10.3 dB compared to the baseline of 6.0 dB for two-speaker separation, as well as a 7.1 dB SDR improvement for three-speaker separation. We then extend the model to incorporate an enhancement layer to refine the signal estimates, and perform end-to-end training through both the clustering and enhancement stages to maximize signal fidelity. We evaluate the results using automatic speech recognition. The new signal approximation objective, combined with end-to-end training, produces unprecedented performance, reducing the word error rate (WER) from 89.1% down to 30.8%. This represents a major advancement towards solving the cocktail party problem.

Jointly Optimizing Activation Coefficients of Convolutive NMF Using DNN for Speech Separation

Hao Li[1], Shuai Nie[2], Xueliang Zhang[1], Hui Zhang[1]; [1]Inner Mongolia University, China; [2]Chinese Academy of Sciences, China

Fri-O-2-6-3, Time: 15:10

Convolutive non-negative matrix factorization (CNMF) and deep neural networks (DNN) are two efficient methods for monaural speech separation. Conventional DNN focuses on building the non-linear relationship between mixture and target speech. However, it ignores the prominent structure of the target speech. Conventional CNMF model concentrates on capturing prominent harmonic structures and temporal continuities of speech but it ignores the non-linear relationship between the mixture and target. Taking these two aspects into consideration at the same time may result in better performance. In this paper, we propose a joint optimization of DNN models with an extra CNMF layer for speech separation task. We also utilize an extra masking layer on the proposed model to constrain the speech reconstruction. Moreover, a discriminative training criterion is proposed to further enhance the performance of the separation. Experimental results show that the proposed model has significant improvement in PESQ, SAR, SIR and SDR compared with conventional methods.

A Feature Study for Masking-Based Reverberant Speech Separation

Masood Delfarah, DeLiang Wang; Ohio State University, USA

Fri-O-2-6-4, Time: 15:30

Monaural speech separation in reverberant conditions is very challenging. In masking-based separation, features extracted from speech mixtures are employed to predict a time-frequency mask. Robust feature extraction is crucial for the performance of supervised speech separation in adverse acoustic environments. Using objective speech intelligibility as the metric, we investigate a wide variety of monaural features in low signal-to-noise ratios and moderate to high reverberation. Deep neural networks are employed as the learning machine in our feature investigation. We find considerable performance gain using a contextual window in reverberant speech processing, likely due to temporal structure of reverberation. In addition, we systematically evaluate feature combinations. In unmatched noise and reverberation conditions, the resulting feature set from this study substantially outperforms previously employed sets for speech separation in anechoic conditions.

Discriminative Layered Nonnegative Matrix Factorization for Speech Separation

Chung-Chien Hsu, Tai-Shih Chi, Jen-Tzung Chien; National Chiao Tung University, Taiwan

Fri-O-2-6-5, Time: 15:50

This paper proposes a discriminative layered nonnegative matrix factorization (DL-NMF) for monaural speech separation. The standard NMF conducts the parts-based representation using a single-layer of bases which was recently upgraded to the layered NMF (L-NMF) where a tree of bases was estimated for multi-level or multi-aspect decomposition of a complex mixed signal. In this study, we develop the DL-NMF by extending the generative bases in L-NMF to the discriminative bases which are estimated according to a discriminative criterion. The discriminative criterion is conducted by optimizing the recovery of the mixed spectra from the separated spectra and minimizing the reconstruction errors between separated spectra and original source spectra. The experiments on single-channel speech separation show the superiority of DL-NMF to NMF and L-NMF in terms of the SDR, SIR and SAR measures.

On Discriminative Framework for Single Channel Audio Source Separation

Arpita Gang, Pravesh Biyani; IIIT Delhi, India

Fri-O-2-6-6, Time: 16:10

Single channel source separation (SCSS) algorithms that utilise discriminative source models perform better in comparison to those that are trained independently. However, all the aspects of training discriminative models have not been addressed in the literature. For instance, the choice of dimensions of source models (number of columns of NMF, Dictionary etc) not only influences the fidelity of a given source but also impacts the interference introduced in it. Therefore choosing a right dimension parameter for every source model is crucial for an effective separation. In fact, the similarity between the constituent sources can be different for different mixtures and thus, dimensions should also be chosen specific to the sources in the concerned mixture. Further, separation of a given constituent from a mixture, assuming remaining to be interferers, offers more freedom for the particular constituent and hence provide better separation. In this paper, we propose a generic discriminative learning framework where we separate one source at a time and embed our dimension search algorithm in the training of discriminative source models. We apply our framework on the NMF based SCSS algorithms and demonstrate a performance improvement in separation for both speech-speech and speech-music mixture.

NOTES

Fri-P-2-1 : Special Session: Auditory-Visual Expressive Speech and Gesture in Humans and Machines

Pacific Concourse – Poster A, 14:30–16:30, Friday, 9 Sept. 2016
Chairs: Jeesun Kim, Gérard Bailly

Generating Natural Video Descriptions via Multimodal Processing

Qin Jin[1], Junwei Liang[2], Xiaozhu Lin[1]; [1]Renmin University of China, China; [2]Carnegie Mellon University, USA

Fri-P-2-1-1, Time: 14:30

Generating natural language descriptions of visual content is an intriguing task which has wide applications such as assisting blind people. The recent advances in image captioning stimulate further study of this task in more depth including generating natural descriptions for videos. Most works of video description generation focus on visual information in the video. However, audio provides rich information for describing video contents as well. In this paper, we propose to generate video descriptions in natural sentences via multimodal processing, which refers to using both audio and visual cues via unified deep neural networks with both convolutional and recurrent structure. Experimental results on the Microsoft Research Video Description (MSVD) corpus prove that fusing audio information greatly improves the video description performance. We also investigate the impact of image amount vs caption amount on the image caption performance and see the trend that when limited amount of training is available, number of various captions is more important than number of various images. This will guide us to investigate in the future how to improve the video description system via increasing amount of training data.

Feature-Level Decision Fusion for Audio-Visual Word Prominence Detection

Martin Heckmann; Honda Research Institute Europe, Germany

Fri-P-2-1-2, Time: 14:30

Common fusion techniques in audio-visual speech processing operate on the modality level. I.e. they either combine the features extracted from the two modalities directly or derive a decision for each modality separately and then combine the modalities on the decision level. We investigate the audio-visual processing of linguistic prosody, more precisely the extraction of word prominence. In this context the different features for each modality can be assumed to be only partially dependent. Hence we propose to train a classifier for each of these features, acoustic and visual modality, and then combine them on a decision level. We compare this approach with conventional fusion methods, i.e. feature fusion and decision fusion on the modality level. Our results show that the feature-level decision fusion clearly outperforms the other approaches, in particular when we also additionally integrate the features resulting from the feature fusion. Compared to a detection based only on the full audio stream we obtain relative improvements from the audio-visual detection of 19% for clean audio and up to 50% for noisy audio.

Acoustic and Visual Analysis of Expressive Speech: A Case Study of French Acted Speech

Slim Ouni[1], Vincent Colotte[1], Sara Dahmani[1], Soumaya Azzi[2]; [1]LORIA, France; [2]Polytech Clermont-Ferrand, France

Fri-P-2-1-3, Time: 14:30

Within the framework of developing an expressive audiovisual speech synthesis, an acoustic and visual analysis of expressive acted speech is proposed in this paper. Our purpose is to identify the main characteristics of audiovisual expressions that need to be integrated during synthesis to provide believable emotions to the virtual 3D talking head. We conducted a case study of a semi-professional actor who uttered a set of sentences for 6 different emotions in addition to neutral speech. We have recorded concurrently audio and motion capture data. The acoustic and the visual data have been analyzed. The main finding is that although some expressions are not well identified, some expressions were well characterized and tied in both acoustic and visual space.

Characterization of Audiovisual Dramatic Attitudes

Adela Barbulescu[1], Rémi Ronfard[1], Gérard Bailly[2]; [1]Inria, France; [2]GIPSA, France

Fri-P-2-1-4, Time: 14:30

In this work we explore the capability of audiovisual parameters (such as fundamental frequency, rhythm, head motion or facial expressions) to discriminate among different dramatic attitudes. We extract the audiovisual parameters from an acted corpus of attitudes and structure them as frame, syllable, and sentence-level features. Using Linear Discriminant Analysis classifiers, we show that sentence-level features present a higher discriminating rate among the attitudes. We also compare the classification results with the perceptual evaluation tests, showing that F0 is correlated to the perceptual results for all attitudes, while other features, such as head motion, contribute differently, depending both on the attitude and the speaker.

Conversational Engagement Recognition Using Auditory and Visual Cues

Yuyun Huang, Emer Gilmartin, Nick Campbell; Trinity College Dublin, Ireland

Fri-P-2-1-5, Time: 14:30

Automatic prediction of engagement in human-human and human-machine dyadic and multiparty interaction scenarios could greatly aid in evaluation of the success of communication. A corpus of eight face-to-face dyadic casual conversations was recorded and used as the basis for an engagement study, which examined the effectiveness of several methods of engagement level recognition. A convolutional neural network based analysis was seen to be the most effective.

An Acoustic Analysis of Child-Child and Child-Robot Interactions for Understanding Engagement during Speech-Controlled Computer Games

Theodora Chaspari, Jill Fain Lehman; Disney Research, USA

Fri-P-2-1-6, Time: 14:30

Engagement is an essential factor towards successful game design and effective human-computer interaction. We analyze the

Notes

prosodic patterns of child-child and child-robot pairs playing a language-based computer game. Acoustic features include speech loudness and fundamental frequency. We use a linear mixed-effects model to capture the coordination of acoustic patterns between interactors as well as its relation to annotated engagement levels. Our results indicate that the considered acoustic features are related to engagement levels for both the child-child and child-robot interaction. They further suggest significant association of the prosodic patterns during the child-child scenario, which is moderated by the co-occurring engagement. This acoustic coordination is not present in the child-robot interaction, since the robot's behavior was not automatically adjusted to the child. These findings are discussed in relation to automatic robot adaptation and provide a foundation for promoting engagement and enhancing rapport during the considered game-based interactions.

Auditory-Visual Lexical Tone Perception in Thai Elderly Listeners with and without Hearing Impairment

Benjawan Kasisopa[1], Chutamanee Onsuwan[2], Charturong Tantibundhit[2], Nittayapa Klangpornkun[2], Suparak Techacharoenrungrueang[3], Sudaporn Luksaneeyanawin[3], Denis Burnham[1]; [1]Western Sydney University, Australia; [2]Thammasat University, Thailand; [3]Chulalongkorn University, Thailand

Fri-P-2-1-7, Time: 14:30

Lexical tone perception was investigated in elderly Thais with Normal Hearing (NH), or Hearing Impairment (HI), the latter with and without Hearing Aids. Auditory-visual (AV), auditory-only (AO), and visual-only (VO) discrimination of Thai tones was investigated. Both groups performed poorly in VO. In AV and AO, the NH performed better than the HI group, and Hearing Aids facilitated tone discrimination. There was slightly more visual augmentation (AV>AO) for the HI group, but not the NH group. The Falling-Rising (FR) pair of tones was easiest to discriminate for both groups and there was a similar ranking of relative discriminability of all 10 tone contrasts for the HI group with and without hearing aids, but this differed from the ranking in the NH group. These results show that the Hearing Impaired elderly with and without hearing aids can, and do, use visual speech information to augment tone perception, but do so in a similar, *not a significantly more* enhanced manner than the Normal Hearing elderly. Thus hearing loss in the Thai elderly does not result in greater use of visual information for discrimination of lexical tone; rather, *all* Thai elderly use visual information to augment their auditory perception of tone.

Use of Agreement/Disagreement Classification in Dyadic Interactions for Continuous Emotion Recognition

Hossein Khaki, Engin Erzin; Koç Üniversitesi, Turkey

Fri-P-2-1-8, Time: 14:30

Natural and affective handshakes of two participants define the course of dyadic interaction. Affective states of the participants are expected to be correlated with the nature or type of the dyadic interaction. In this study, we investigate relationship between affective attributes and nature of dyadic interaction. In this investigation we use the JESTKOD database, which consists of speech and full-body motion capture data recordings for dyadic interactions under agreement and disagreement scenarios. The dataset also has affective annotations in activation, valence and dominance (AVD) attributes. We pose the continuous affect recognition problem under agreement and disagreement scenarios of dyadic interactions. We define a statistical mapping using the support vector regression (SVR) from speech and motion modalities to affective attributes with and without the dyadic interaction type (DIT) information. We observe an improvement in estimation of the valence attribute when the DIT is available. Furthermore this improvement sustains even we estimate the DIT from the speech and motion modalities of the dyadic interaction.

Fri-P-2-2 : Special Session: Intelligibility Under the Microscope

Pacific Concourse – Poster B, 14:30–16:30, Friday, 9 Sept. 2016
Chairs: Ricard Marxer, Martin Cooke, Jon Barker

Microscopic Multilingual Matrix Test Predictions Using an ASR-Based Speech Recognition Model

Marc René Schädler, David Hülsmeier, Anna Warzybok, Sabine Hochmuth, Birger Kollmeier; Carl von Ossietzky Universität Oldenburg, Germany

Fri-P-2-2-1, Time: 14:30

In an attempt to predict the outcomes of matrix sentence tests in different languages and various noise conditions for native listeners, the simulation framework for auditory discrimination experiments (FADE) and the extended Speech Intelligibility Index (eSII) is employed. FADE uses an automatic speech recognition system to simulate recognition experiments and reports the highest achievable performance as the outcome, which showed good predictions for the German matrix test in noise. The eSII is based on the short-time analysis of weighted signal-to-noise ratios in different frequency bands. In contrast to many other approaches, including the eSII, FADE uses no empirical reference. In this work, the FADE approach is evaluated for predictions of the German, Polish, Russian, and Spanish matrix test in stationary and fluctuating noise conditions. The FADE-based predictions yield a high correlation (Pearsons R^2 = 0.94) with the empirical data and a root-mean-square (RMS) prediction error of 1.9 dB outperforming the eSII-based predictions (R^2 = 0.78, RMS = 4.2 dB). FADE can also predict the data of subgroups with only stationary or only fluctuating noises, while the eSII cannot. The FADE-based predictions seem to generalize over different languages and noise conditions.

DNN-Based Automatic Speech Recognition as a Model for Human Phoneme Perception

Mats Exter[1], Bernd T. Meyer[2]; [1]Carl von Ossietzky Universität Oldenburg, Germany; [2]Johns Hopkins University, USA

Fri-P-2-2-2, Time: 14:30

In this paper, we test the applicability of state-of-the-art automatic speech recognition (ASR) to predict phoneme confusions in human listeners. Phoneme-specific response rates are obtained from ASR based on deep neural networks (DNNs) and from listening tests with six normal-hearing subjects. The measure for model quality is the correlation of phoneme recognition accuracies obtained in ASR and

Notes

in human speech recognition (HSR). Various feature representations are used as input to the DNNs to explore their relation to overall ASR performance and model prediction power. Standard filterbank output and perceptual linear prediction (PLP) features result in best predictions, with correlation coefficients reaching $r = 0.9$.

Undoing Misperceptions: A Microscopic Analysis of Consistent Confusions Through Signal Modifications

Attila Máté Tóth[1], Martin Cooke[2]; [1]Universidad del País Vasco, Spain; [2]Ikerbasque, Spain

Fri-P-2-2-3, Time: 14:30

Consistent confusions — word misperceptions reported in an open set task with a high agreement across listeners — can be especially valuable in understanding the detailed processes underlying speech perception. The current study investigates the origin of a set of consistent confusions collected in a variety of masking conditions, by applying signal-level modifications to the stimuli eliciting the confusion, and subsequently reevaluating listeners' percepts. Modifications were selected to provide release from either the energetic or the informational component of the maskers and involved manipulations of signal-to-noise ratio, fundamental frequency, and resynthesis of the noise-mixture in glimpsed regions of the target speech. Increasing signal-to-noise ratio and glimpse resynthesis showed the expected release from energetic and informational masking respectively. However, manipulations targeting informational masking release, including fundamental frequency modification, affected a surprisingly high number of confusions stemming from energetic maskers. The degree of fundamental frequency shift did not have a significant effect on the response patterns observed. Around 30% of confusions can be explained solely based on the information contained within the target glimpses surviving energetic masking, while for the rest of the cases additional factors, such as recruitment of information from the masker, appear to be involved.

Blind Non-Intrusive Speech Intelligibility Prediction Using Twin-HMMs

Mahdie Karbasi[1], Ahmed Hussen Abdelaziz[2], Hendrik Meutzner[1], Dorothea Kolossa[1]; [1]Ruhr-Universität Bochum, Germany; [2]ICSI, USA

Fri-P-2-2-4, Time: 14:30

Automatic prediction of speech intelligibility is highly desirable in the speech research community, since listening tests are time-consuming and can not be used online. Most of the available objective speech intelligibility measures are intrusive methods, as they require a clean reference signal in addition to the corresponding noisy/processed signal at hand. In order to overcome the problem of predicting the speech intelligibility in the absence of the clean reference signal, we have proposed in [1] to employ a recognition/synthesis framework called twin hidden Markov model (THMM) for synthesizing the clean features, required inside an intrusive intelligibility prediction method. The new framework can predict the speech intelligibility equally well as well-known intrusive methods like the short-time objective intelligibility (STOI). The original THMM, however, requires the correct transcription for synthesizing the clean reference features, which is not always available. In this paper, we go one step further and investigate the use of the recognized transcription instead of the oracle transcription for obtaining a more widely applicable speech intelligibility prediction. We show that the output of the newly-proposed blind approach is highly correlated with the human speech recognition results, collected via crowdsourcing in different noise conditions.

Misperceptions Arising from Speech-in-Babble Interactions

Attila Máté Tóth[1], Martin Cooke[2], Jon Barker[3]; [1]Universidad del País Vasco, Spain; [2]Ikerbasque, Spain; [3]University of Sheffield, UK

Fri-P-2-2-5, Time: 14:30

The deterioration of speech intelligibility in the presence of other sound sources has been explained in terms of both energetic masking, which renders parts of the speech signal inaudible, and informational masking, in which audible components of the masker interfere with speech identification. The current study focuses on the role of a specific form of informational masking in which audible glimpses of both target and masker combine to produce an incorrect listener percept. We examine a corpus of word misperceptions in Spanish which occur when target words are combined with a babble masker. Glimpses originating in both the target and the masker are force-aligned to the reported misperceived word in order to identify the most likely acoustic evidential basis for the confusion. In this way, the degree of involvement of both target and masker can be quantified. In nearly all cases, the best explanation for the misperception involves recruiting evidence from the babble masker (type I error), and in more than 80% of the tokens some of the audible target evidence is ignored (type II error). These findings suggest misallocation of acoustic-phonetic material plays a significant role in the generation of speech-in-babble confusions.

Introducing Temporal Rate Coding for Speech in Cochlear Implants: A Microscopic Evaluation in Humans and Models

Anja Eichenauer[1], Mathias Dietz[1], Bernd T. Meyer[2], Tim Jürgens[1]; [1]Carl von Ossietzky Universität Oldenburg, Germany; [2]Johns Hopkins University, USA

Fri-P-2-2-6, Time: 14:30

Standard cochlea implant (CI) speech coding strategies transmit formant information only via the place of the stimulated electrode. In acoustic hearing, however, formant frequencies are additionally coded via the temporal rate of auditory nerve firing. This study presents a novel CI coding strategy ("Formant Locking (FL)-strategy") that varies stimulation rates in relation to extracted fundamental and formant frequencies. Simulated auditory nerve activity resulting from stimulation with the FL-strategy shows that the FL-strategy triggers spike rates that are related to the formant frequencies similar as in normal hearing, and greatly different than in a standard CI strategy. Vowel recognition in seven CI users via direct stimulation of their electrode array shows that the FL-strategy results in significantly increased scores of the vowels /u/ and /i/ compared to a standard CI strategy. However, at the same time, a decrease in scores for /o/ and /e/ occurred. A microscopic speech intelligibility model involving an automatic speech recognizer reveals good agreement between modeled and predicted confusion matrices for the FL-strategy. This suggests that microscopic models can be used to test CI strategies in the development phase, and gives indications which cues might be used by the listeners for speech recognition.

NOTES

Language Effects in Noise-Induced Word Misperceptions

Maria Luisa Garcia Lecumberri[1], Jon Barker[2], Ricard Marxer[2], Martin Cooke[3]; [1]Universidad del País Vasco, Spain; [2]University of Sheffield, UK; [3]Ikerbasque, Spain

Fri-P-2-2-7, Time: 14:30

Speech misperceptions provide a window into the processes underlying spoken language comprehension. One approach shown to catalyse robust misperceptions is to embed words in noise. However, the use of masking noise makes it difficult to measure the relative contributions of low-level auditory processing and higher-level factors which involve the deployment of linguistic experience. The current study addresses this confound by comparing noise-induced misperceptions in two languages, Spanish and English, which display marked phonological differences in properties such as consonant-vowel ratio, rhythm and syllable structure. An analysis of over 5000 word-level misperceptions generated using a common experimental framework in the two languages reveals some striking similarities: the proportion of confusions generated by three distinct types of masker are almost identical for the two languages, as are the proportions of phonemic and syllabic insertions, deletions and substitutions. The biggest difference is seen for babble noise, which tends to induce relatively complex confusions in English and simpler confusions in Spanish. We speculate that the inflectional morphology of Spanish lends itself to more easily recruit single elements from a babble masker into valid word hypotheses.

Speech Reductions Cause a De-Weighting of Secondary Acoustic Cues

Léo Varnet[1], Fanny Meunier[2], Michel Hoen[3]; [1]LSP (UMR 8248), France; [2]BCL (UMR 7320), France; [3]Oticon Medical, France

Fri-P-2-2-8, Time: 14:30

The ability of the auditory system to change the perceptual weighting of acoustic cues when faced with degraded speech has long been evidenced. However, the exact changes that occur remain mostly unknown. Here, we proposed to use the Auditory Classification Image (ACI) methodology to reveal the acoustic cues used in natural speech comprehension and in reduced (i.e. noise-vocoded or re-synthesized) speech comprehension. The results show that in the latter case the auditory system updates its listening strategy by de-weighting secondary acoustic cues. Indeed, these are often weaker and thus more easily erased in adverse listening conditions. Furthermore our data suggests that this de-weighting does not directly depend on the actual reliability of the cues, but rather on the expected change in informativeness.

Using Phonologically Weighted Levenshtein Distances for the Prediction of Microscopic Intelligibility

Lionel Fontan[1], Isabelle Ferrané[2], Jérôme Farinas[2], Julien Pinquier[2], Xavier Aumont[1]; [1]Archean Technologies, France; [2]IRIT, France

Fri-P-2-2-9, Time: 14:30

This article presents a new method for analyzing Automatic Speech Recognition (ASR) results at the phonological feature level. To this end the Levenshtein distance algorithm is refined in order to take into account the distinctive features opposing substituted phonemes. This method allows to survey features additions or deletions, providing microscopic qualitative information as a complement to word recognition scores. To explore the relevance of the qualitative data gathered by this method, a study is conducted on a speech corpus simulating presbycusis effects on speech perception at eight severity stages. Consonantic features additions and deletions in ASR outputs are analyzed and put in relation with intelligibility data collected in 30 human subjects. ASR results show monotonic trends in most consonantic features along the degradation conditions, which appear to be consistent with the misperceptions that could be observed in human subjects.

The Impact of Manner of Articulation on the Intelligibility of Voicing Contrast in Noise: Cross-Linguistic Implications

Mayuki Matsui; NINJAL, Japan

Fri-P-2-2-10, Time: 14:30

The current study addresses the impact of manner of articulation on the intelligibility of voicing contrast in noise from a cross-linguistic perspective. Previous noise-masking studies have suggested that the impact of manner of articulation on the intelligibility of voicing contrast in noise is apparently different in Russian and English. In order to further assess the source of such a cross-linguistic inconsistency, the current study examines how Russian voicing contrast is perceived by English listeners. Native listeners of English performed a forced-choice identification task with Russian voiced and voiceless stimuli in quiet and noisy conditions. The results showed that the voicing contrast in stops were more confused than that in fricatives for English listeners, showing a pattern similar to Russian listeners. The results suggest that the source of the cross-linguistic difference identified in previous studies comes from the difference in the acoustic properties of the stimuli, reflecting the difference in phonetic implementation of voicing contrasts in each language. The results in turn suggest that perceptual cue weighting strategies for perceiving voicing contrast in different manners of articulation is similar among Russian and English listeners.

Directly Comparing the Listening Strategies of Humans and Machines

Michael I. Mandel; CUNY Brooklyn College, USA

Fri-P-2-2-11, Time: 14:30

In a given noisy environment, human listeners can more accurately identify spoken words than automatic speech recognizers. It is not clear, however, what information the humans are able to utilize in doing so that the machines are not. This paper uses a recently introduced technique to directly characterize the information used by humans and machines on the same task. The task was a forced choice between eight sentences spoken by a single talker from the small-vocabulary GRID corpus that were selected to be maximally confusable with one another. These sentences were mixed with “bubble” noise, which is designed to reveal randomly selected time-frequency glimpses of the sentence. Responses to these noisy mixtures allowed the identification of time-frequency regions that were important for each listener to recognize each sentence, i.e., regions that were frequently audible when a sentence was correctly identified and inaudible when it was not. In comparing these regions across human and machine listeners, we found that dips in noise allowed the humans to recognize words based on informative speech

NOTES

cues. In contrast, the baseline CHiME-2-GRID recognizer correctly identified sentences only when the time-frequency profile of the noisy mixture matched that of the underlying speech.

Fri-P-2-3 : Spoken Documents, Spoken Understanding and Semantic Analysis

Pacific Concourse – Poster C, 14:30–16:30, Friday, 9 Sept. 2016
Chair: Helen Meng

LSTM-Based NeuroCRFs for Named Entity Recognition

Marc-Antoine Rondeau [1], Yi Su [2]; [1] McGill University, Canada; [2] Nuance Communications, Canada

Fri-P-2-3-1, Time: 14:30

Although NeuroCRF, an augmented Conditional Random Fields (CRF) model whose feature function is parameterized as a Feed-Forward Neural Network (FF NN) on word embeddings, has soundly outperformed traditional linear-chain CRF on many sequence labeling tasks, it is held back by the fact that FF NNs have a fixed input length and therefore cannot take advantage of the full input sentence. We propose to address this issue by replacing the FF NN with a Long Short-Term Memory (LSTM) NN, which can summarize an input of arbitrary length into a fixed dimension representation. The resulting model obtains F_1=89.28 on WikiNER dataset, a significant improvement over the NeuroCRF baseline's F_1=87.58, which is already a highly competitive result.

Exploring Word Mover's Distance and Semantic-Aware Embedding Techniques for Extractive Broadcast News Summarization

Shih-Hung Liu [1], Kuan-Yu Chen [1], Yu-Lun Hsieh [1], Berlin Chen [2], Hsin-Min Wang [1], Hsu-Chun Yen [3], Wen-Lian Hsu [1]; [1] Academia Sinica, Taiwan; [2] National Taiwan Normal University, Taiwan; [3] National Taiwan University, Taiwan

Fri-P-2-3-2, Time: 14:30

Extractive summarization is a process that manages to select the most salient sentences from a document (or a set of documents) and subsequently assemble them to form an informative summary, facilitating users to browse and assimilate the main theme of the document efficiently. Our work in this paper continues this general line of research and its main contributions are two-fold. First, we explore to leverage the recently proposed word mover's distance (WMD) metric, in conjunction with semantic-aware continuous space representations of words, to authentically capture finer-grained sentence-to-document and/or sentence-to-sentence semantic relatedness for effective use in the summarization process. Second, we investigate to combine our proposed approach with several state-of-the-art summarization methods, which originally adopted the conventional term-overlap or bag-of-words (BOW) approaches for similarity calculation. A series of experiments conducted on a typical broadcast news summarization task seem to suggest the performance merits of our proposed approach, in comparison to the mainstream methods.

Improved Neural Bag-of-Words Model to Retrieve Out-of-Vocabulary Words in Speech Recognition

Imran Sheikh [1], Irina Illina [1], Dominique Fohr [1], Georges Linarès [2]; [1] LORIA, France; [2] LIA, France

Fri-P-2-3-3, Time: 14:30

Many Proper Names (PNs) are Out-Of-Vocabulary (OOV) words for speech recognition systems used to process diachronic audio data. To enable recovery of the PNs missed by the system, relevant OOV PNs can be retrieved by exploiting the semantic context of the spoken content. In this paper, we explore the Neural Bag-of-Words (NBOW) model, proposed previously for text classification, to retrieve relevant OOV PNs. We propose a Neural Bag-of-Weighted-Words (NBOW2) model in which the input embedding layer is augmented with a context anchor layer. This layer learns to assign importance to input words and has the ability to capture (task specific) key-words in a NBOW model. With experiments on French broadcast news videos we show that the NBOW and NBOW2 models outperform earlier methods based on raw embeddings from LDA and Skip-gram. Combining NBOW with NBOW2 gives faster convergence during training.

Beyond Utterance Extraction: Summary Recombination for Speech Summarization

Jérémy Trione, Benoit Favre, Frederic Bechet; LIF (UMR 7279), France

Fri-P-2-3-4, Time: 14:30

This paper describes a template filling approach for creating conversation summaries. The templates are generated from generalized summary fragments from a training corpus. Necessary pieces of information for filling them are extracted automatically from the conversation transcripts given linguistic features, and drive the fragment selection process. The approach obtains ROUGE-2 scores of 0.08471 on the RATP-DECODA corpus, which represents a significant improvement over extractive baselines and hand-written templates.

Attention-Based Recurrent Neural Network Models for Joint Intent Detection and Slot Filling

Bing Liu, Ian Lane; Carnegie Mellon University, USA

Fri-P-2-3-5, Time: 14:30

Attention-based encoder-decoder neural network models have recently shown promising results in machine translation and speech recognition. In this work, we propose an attention-based neural network model for joint intent detection and slot filling, both of which are critical steps for many speech understanding and dialog systems. Unlike in machine translation and speech recognition, alignment is explicit in slot filling. We explore different strategies in incorporating this alignment information to the encoder-decoder framework. Learning from the attention mechanism in encoder-decoder model, we further propose introducing attention to the alignment-based RNN models. Such attentions provide additional information to the intent classification and slot label prediction. Our independent task models achieve state-of-the-art intent detection error rate and slot filling F1 score on the benchmark ATIS task. Our joint training model further obtains 0.56% absolute (23.8% relative) error reduction on intent detection and 0.23% absolute gain on slot filling over the independent task models.

Notes

Domain Adaptation of Recurrent Neural Networks for Natural Language Understanding

Aaron Jaech[1], Larry Heck[2], Mari Ostendorf[1]; [1]University of Washington, USA; [2]Google, USA

Fri-P-2-3-6, Time: 14:30

The goal of this paper is to use multi-task learning to efficiently scale slot filling models for natural language understanding to handle multiple target tasks or domains. The key to scalability is reducing the amount of training data needed to learn a model for a new task. The proposed multi-task model delivers better performance with less data by leveraging patterns that it learns from the other tasks. The approach supports an open vocabulary, which allows the models to generalize to unseen words, which is particularly important when very little training data is used. A newly collected crowd-sourced data set, covering four different domains, is used to demonstrate the effectiveness of the domain adaptation and open vocabulary techniques.

LatticeRnn: Recurrent Neural Networks Over Lattices

Faisal Ladhak, Ankur Gandhe, Markus Dreyer, Lambert Mathias, Ariya Rastrow, Björn Hoffmeister; Amazon.com, USA

Fri-P-2-3-7, Time: 14:30

We present a new model called LatticeRnn, which generalizes recurrent neural networks (RNNs) to process weighted lattices as input, instead of sequences. A LatticeRnn can encode the complete structure of a lattice into a dense representation, which makes it suitable to a variety of problems, including rescoring, classifying, parsing, or translating lattices using deep neural networks (DNNs). In this paper, we use LatticeRnns for a classification task: each lattice represents the output from an automatic speech recognition (ASR) component of a spoken language understanding (SLU) system, and we classify the intent of the spoken utterance based on the lattice embedding computed by a LatticeRnn. We show that making decisions based on the full ASR output lattice, as opposed to 1-best or n-best hypotheses, makes SLU systems more robust to ASR errors. Our experiments yield improvements of 13% over a baseline RNN system trained on transcriptions and 10% over an n-best list rescoring system for intent classification.

Learning Document Representations Using Subspace Multinomial Model

Santosh Kesiraju, Lukáš Burget, Igor Szőke, Jan Černocký; Brno University of Technology, Czech Republic

Fri-P-2-3-8, Time: 14:30

Subspace multinomial model (SMM) is a log-linear model and can be used for learning low dimensional continuous representation for discrete data. SMM and its variants have been used for speaker verification based on prosodic features and phonotactic language recognition. In this paper, we propose a new variant of SMM that introduces sparsity and call the resulting model as ℓ_1 SMM. We show that ℓ_1 SMM can be used for learning document representations that are helpful in topic identification or classification and clustering tasks. Our experiments in document classification show that SMM achieves comparable results to models such as latent Dirichlet allocation and sparse topical coding, while having a useful property that the resulting document vectors are Gaussian distributed.

Attention-Based Convolutional Neural Networks for Sentence Classification

Zhiwei Zhao, Youzheng Wu; Sony, China

Fri-P-2-3-9, Time: 14:30

Sentence classification is one of the foundational tasks in spoken language understanding (SLU) and natural language processing (NLP). In this paper we propose a novel convolutional neural network (CNN) with attention mechanism to improve the performance of sentence classification. In traditional CNN, it is not easy to encode long term contextual information and correlation between non-consecutive words effectively. In contrast, our attention-based CNN is able to capture these kinds of information for each word without any external features. We conducted experiments on various public and in-house datasets. The experimental results demonstrate that our proposed model significantly outperforms the traditional CNN model and achieves competitive performance with the ones that exploit rich syntactic features.

Spoken Language Understanding in a Latent Topic-Based Subspace

Mohamed Morchid, Mohamed Bouaziz, Waad Ben Kheder, Killian Janod, Pierre-Michel Bousquet, Richard Dufour, Georges Linarès; LIA, France

Fri-P-2-3-10, Time: 14:30

Performance of spoken language understanding applications declines when spoken documents are automatically transcribed in noisy conditions due to high Word Error Rates (WER). To improve the robustness to transcription errors, recent solutions propose to map these automatic transcriptions in a latent space. These studies have proposed to compare classical topic-based representations such as Latent Dirichlet Allocation (LDA), supervised LDA and author-topic (AT) models. An original compact representation, called c-vector, has recently been introduced to walk around the tricky choice of the number of latent topics in these topic-based representations. Moreover, c-vectors allow to increase the robustness of document classification with respect to transcription errors by compacting different LDA representations of a same speech document in a reduced space and then compensate most of the noise of the document representation. The main drawback of this method is the number of sub-tasks needed to build the c-vector space. This paper proposes to both improve this compact representation (c-vector) of spoken documents and to reduce the number of needed sub-tasks, using an original framework in a robust low dimensional space of features from a set of AT models called “Latent Topic-based Subspace” (LTS). In comparison to LDA, the AT model considers not only the dialogue content (words), but also the class related to the document. Experiments are conducted on the DECODA corpus containing speech conversations from the call-center of the RATP Paris transportation company. Results show that the original LTS representation outperforms the best previous compact representation (c-vector), with a substantial gain of more than 2.5% in terms of correctly labeled conversations.

Notes

Multi-Domain Joint Semantic Frame Parsing Using Bi-Directional RNN-LSTM

Dilek Hakkani-Tür[1], Gokhan Tur[1], Asli Celikyilmaz[1], Yun-Nung Chen[2], Jianfeng Gao[1], Li Deng[1], Ye-Yi Wang[1]; [1]Microsoft, USA; [2]National Taiwan University, Taiwan

Fri-P-2-3-11, Time: 14:30

Sequence-to-sequence deep learning has recently emerged as a new paradigm in supervised learning for spoken language understanding. However, most of the previous studies explored this framework for building single domain models for each task, such as slot filling or domain classification, comparing deep learning based approaches with conventional ones like conditional random fields. This paper proposes a holistic multi-domain, multi-task (i.e. slot filling, domain and intent detection) modeling approach to estimate complete semantic frames for all user utterances addressed to a conversational system, demonstrating the distinctive power of deep learning methods, namely bi-directional recurrent neural network (RNN) with long-short term memory (LSTM) cells (RNN-LSTM) to handle such complexity. The contributions of the presented work are three-fold: (i) we propose an RNN-LSTM architecture for joint modeling of slot filling, intent determination, and domain classification; (ii) we build a joint multi-domain model enabling multi-task deep learning where the data from each domain reinforces each other; (iii) we investigate alternative architectures for modeling lexical context in spoken language understanding. In addition to the simplicity of the single model framework, experimental results show the power of such an approach on Microsoft Cortana real user data over alternative methods based on single domain/task deep learning.

Deep Stacked Autoencoders for Spoken Language Understanding

Killian Janod, Mohamed Morchid, Richard Dufour, Georges Linarès, Renato De Mori; LIA, France

Fri-P-2-3-12, Time: 14:30

The automatic transcription process of spoken document results in several word errors, especially when very noisy conditions are encountered. Document representations based on neural embedding frameworks have recently shown significant improvements in different Spoken and Natural Language Understanding tasks such as denoising and filtering. Nonetheless, these methods mainly need clean representations, failing to properly remove noise contained in noisy representations. This paper proposes to study the impact of residual noise contained into automatic transcripts of spoken dialogues in highly abstract spaces from deep neural networks. The paper makes the assumption that the noise learned from “clean” manual transcripts of spoken documents moves down dramatically the performance of theme identification systems in noisy conditions. The proposed deep neural network takes, as input and output, highly imperfect transcripts from spoken dialogues to improve the robustness of the document representation in a noisy environment. Results obtained on the DECODA theme classification task of dialogues reach an accuracy of 82% with a significant gain of about 5%.

Labeled Data Generation with Encoder-Decoder LSTM for Semantic Slot Filling

Gakuto Kurata[1], Bing Xiang[2], Bowen Zhou[2]; [1]IBM, Japan; [2]IBM, USA

Fri-P-2-3-13, Time: 14:30

To train a model for semantic slot filling, manually labeled data in which each word is annotated with a semantic slot label is necessary while manually preparing such data is costly. Starting from a small amount of manually labeled data, we propose a method to generate the labeled data with using the encoder-decoder LSTM. We first train the encoder-decoder LSTM that accepts and generates the same manually labeled data. Then, to generate a wide variety of labeled data, we add perturbations to the vector that encodes the manually labeled data and generate labeled data with the decoder LSTM based on the perturbated encoded vector. We also try to enhance the encoder-decoder LSTM to generate the word sequences and their label sequences separately to obtain new pairs of words and their labels. Through the experiments with the standard ATIS slot filling task, by using the generated data, we obtained improvement in slot filling accuracy over the strong baseline with the NN-based slot filling model.

Exploring the Correlation of Pitch Accents and Semantic Slots for Spoken Language Understanding

Sabrina Stehwien, Ngoc Thang Vu; Universität Stuttgart, Germany

Fri-P-2-3-14, Time: 14:30

We investigate the correlation between pitch accents and semantic slots in human-machine speech. Using an automatic pitch accent detector on the ATIS corpus, we find that most words labelled with semantic slots also carry a pitch accent. Most of the pitch accented words that are not associated with a semantic label are still meaningful, pointing towards the speaker’s intention. Our findings show that prosody constitutes a relevant and useful resource for spoken language understanding, especially considering the fact that our pitch accent detector does not require any kind of manual transcriptions during testing time.

Analysis on Gated Recurrent Unit Based Question Detection Approach

Yaodong Tang, Zhiyong Wu, Helen Meng, Mingxing Xu, Lianhong Cai; Tsinghua University, China

Fri-P-2-3-15, Time: 14:30

Recent studies have shown various kinds of recurrent neural networks (RNNs) are becoming powerful sequence models in speech related applications. Our previous work in detecting questions of Mandarin speech presents that gated recurrent unit (GRU) based RNN can achieve significantly better results. In this paper, we try to open the black box to find the correlations between inner architecture of GRU and phonetic features of question sentences. We find that both update gate and reset gate in GRU blocks react when people begin to pronounce a word. According to the reactions, experiments are conducted to show the behavior of GRU based question detection approach on three important factors, including keywords or special structure of questions, final particles and interrogative intonation. We also observe that update gate and reset gate don’t collaborate well on our dataset. Based on the asynchronous acts of update gate and reset gate in GRU, we adapt the structure of GRU block to

NOTES

our dataset and get further performance improvement in question detection task.

Fri-P-2-4 : Spoken Term Detection

Pacific Concourse – Poster D, 14:30–16:30, Friday, 9 Sept. 2016
Chairs: Laura June Mayfield Tomokiyo

Combining State-Level Spotting and Posterior-Based Acoustic Match for Improved Query-by-Example Spoken Term Detection

Shuji Oishi, Tatsuya Matsuba, Mitsuaki Makino, Atsuhiko Kai; Shizuoka University, Japan

Fri-P-2-4-1, Time: 14:30

In spoken term detection (STD) systems, automatic speech recognition (ASR) frontend is often employed for its reasonable accuracy and efficiency. However, out-of-vocabulary (OOV) problem at ASR stage has a great impact on the STD performance for spoken query. In this paper, we propose combining feature-based acoustic match which is often employed in the STD systems for low resource languages, along with the other ASR-derived features. First, automatic transcripts for spoken document and spoken query are decomposed into corresponding acoustic model state sequences and used for spotting plausible speech segments. Second, DTW-based acoustic match between the query and candidate segment is performed using the posterior features derived from a monophone-state DNN. Finally, an integrated score is obtained by a logistic regression model, which is trained with a large spoken document and automatically generated spoken queries as development data. The experimental results on NTCIR-12 SpokenQuery&Doc-2 task showed that the proposed method significantly outperforms the baseline systems which use the subword-level or state-level spotting alone. Also, our universal scoring model trained with a separate set of development data could achieve the best STD performance, and showed the effectiveness of additional ASR-derived features regarding the confidence measure and query length.

A Novel Discriminative Score Calibration Method for Keyword Search

Zhiqiang Lv, Meng Cai, Wei-Qiang Zhang, Jia Liu; Tsinghua University, China

Fri-P-2-4-2, Time: 14:30

The performance of keyword search systems depends heavily on the quality of confidence scores. In this work, a novel discriminative score calibration method has been proposed. By training an MLP classifier employing the word posterior probability and several novel normalized scores, we can obtain a relative improvement of 4.67% for the actual term-weighted value (ATWV) metric on the OpenKWS15 development test dataset. In addition, a LSTM-CTC based keyword verification method has been proposed to supply extra acoustic information. After the information is added, a further improvement of 7.05% over the baseline can be observed.

Segmented Dynamic Time Warping for Spoken Query-by-Example Search

Jorge Proença, Fernando Perdigão; Instituto de Telecomunicações, Portugal

Fri-P-2-4-3, Time: 14:30

This paper describes a low-resource approach to a Query-by-Example task, where spoken queries must be matched in a large dataset of spoken documents sometimes in complex or non-exact ways. Our approach tackles these complex match cases by using Dynamic Time Warping to obtain alternative paths that account for reordering of words, small extra content and small lexical variations. We also report certain advances on calibration and fusion of sub-systems that improve overall results, such as manipulating the score distribution per query and using an average posteriorgram distance matrix as an extra sub-system. Results are evaluated on the MediaEval task of Query-by-Example Search on Speech (QUESST). For this task, the language of the audio being searched is almost irrelevant, approaching the use case scenario to a language of very low resources. For that, we use as features the posterior probabilities obtained from five phonetic recognizers trained with five different languages.

Generating Complementary Acoustic Model Spaces in DNN-Based Sequence-to-Frame DTW Scheme for Out-of-Vocabulary Spoken Term Detection

Shi-wook Lee [1], Kazuyo Tanaka [2], Yoshiaki Itoh [3]; [1] AIST, Japan; [2] University of Tsukuba, Japan; [3] Iwate Prefectural University, Japan

Fri-P-2-4-4, Time: 14:30

This paper proposes a sequence-to-frame dynamic time warping (DTW) combination approach to improve out-of-vocabulary (OOV) spoken term detection (STD) performance gain. The goal of this paper is twofold: first, we propose a method that directly adopts the posterior probability of deep neural network (DNN) and Gaussian mixture model (GMM) as the similarity distance for sequence-to-frame DTW. Second, we investigate combinations of diverse schemes in GMM and DNN, with different subword units and acoustic models, estimate the complementarity in terms of performance gap and correlation of the combined systems, and discuss the performance gain of the combined systems. The results of evaluations conducted of the combined systems on an out-of-vocabulary spoken term detection task show that the performance gain of DNN-based systems is better than that of GMM-based systems. However, the performance gain obtained by combining DNN- and GMM-based systems is insignificant, even though DNN and GMM are highly heterogeneous. This is because the performance gap between DNN-based systems and GMM-based systems is quite large. On the other hand, score fusion of two heterogeneous subword units, triphone and sub-phonetic segments, in DNN-based systems provides significantly improved performance.

Multi-Task Learning and Weighted Cross-Entropy for DNN-Based Keyword Spotting

Sankaran Panchapagesan, Ming Sun, Aparna Khare, Spyros Matsoukas, Arindam Mandal, Björn Hoffmeister, Shiv Vitaladevuni; Amazon.com, USA

Fri-P-2-4-5, Time: 14:30

We propose improved Deep Neural Network (DNN) training loss functions for more accurate single keyword spotting on resource-

NOTES

constrained embedded devices. The loss function modifications consist of a combination of multi-task training and weighted cross entropy. In the multi-task architecture, the keyword DNN acoustic model is trained with two tasks in parallel — the main task of predicting the keyword-specific phone states, and an auxiliary task of predicting LVCSR senones. We show that multi-task learning leads to comparable accuracy over a previously proposed transfer learning approach where the keyword DNN training is initialized by an LVCSR DNN of the same input and hidden layer sizes. The combination of LVCSR-initialization and Multi-task training gives improved keyword detection accuracy compared to either technique alone. We also propose modifying the loss function to give a higher weight on input frames corresponding to keyword phone targets, with a motivation to balance the keyword and background training data. We show that weighted cross-entropy results in additional accuracy improvements. Finally, we show that the combination of 3 techniques — LVCSR-initialization, multi-task training and weighted cross-entropy gives the best results, with significantly lower False Alarm Rate than the LVCSR-initialization technique alone, across a wide range of Miss Rates.

Audio Word2Vec: Unsupervised Learning of Audio Segment Representations Using Sequence-to-Sequence Autoencoder

Yu-An Chung, Chao-Chung Wu, Chia-Hao Shen, Hung-Yi Lee, Lin-Shan Lee; National Taiwan University, Taiwan

Fri-P-2-4-6, Time: 14:30

The vector representations of fixed dimensionality for words (in text) offered by Word2Vec have been shown to be very useful in many application scenarios, in particular due to the semantic information they carry. This paper proposes a parallel version, the Audio Word2Vec. It offers the vector representations of fixed dimensionality for variable-length audio segments. These vector representations are shown to describe the sequential phonetic structures of the audio segments to a good degree, with very attractive real world applications such as query-by-example Spoken Term Detection (STD). In this STD application, the proposed approach significantly outperformed the conventional Dynamic Time Warping (DTW) based approaches at significantly lower computation requirements. We propose unsupervised learning of Audio Word2Vec from audio data without human annotation using Sequence-to-sequence Autoencoder (SA). SA consists of two RNNs equipped with Long Short-Term Memory (LSTM) units: the first RNN (encoder) maps the input audio sequence into a vector representation of fixed dimensionality, and the second RNN (decoder) maps the representation back to the input audio sequence. The two RNNs are jointly trained by minimizing the reconstruction error. Denoising Sequence-to-sequence Autoencoder (DSA) is further proposed offering more robust learning.

Non-Uniform Boosted MCE Training of Deep Neural Networks for Keyword Spotting

Zhong Meng, Biing-Hwang Juang; Georgia Institute of Technology, USA

Fri-P-2-4-7, Time: 14:30

Keyword spotting can be formulated as a non-uniform error automatic speech recognition (ASR) problem. It has been demonstrated [1] that this new formulation with the non-uniform MCE training technique can lead to improved system performance in keyword spotting applications. In this paper, we demonstrate that deep neural networks (DNNs) can be successfully trained on the non-uniform minimum classification error (MCE) criterion which weighs the errors on keywords much more significantly than those on non-keywords in an ASR task. The integration with a DNN-HMM system enables modeling of multi-frame distributions, which conventional systems find difficult to accomplish. To further improve the performance, more confusable data is generated by boosting the likelihood of the sentences that have more errors. The keyword spotting system is implemented within a weighted finite state transducer (WFST) framework and the DNN is optimized using standard backpropagation and stochastic gradient decent. We evaluate the performance of the proposed framework on a large vocabulary spontaneous conversational telephone speech dataset (Switchboard-1 Release 2). The proposed approach achieves an absolute figure of merit improvement of 3.65% over the baseline system.

Language Model Data Augmentation for Keyword Spotting in Low-Resourced Training Conditions

Arseniy Gorin[1], Rasa Lileikytė[1], Guangpu Huang[1], Lori Lamel[1], Jean-Luc Gauvain[1], Antoine Laurent[2]; [1]LIMSI, France; [2]Vocapia Research, France

Fri-P-2-4-8, Time: 14:30

This research extends our earlier work on using machine translation (MT) and word-based recurrent neural networks to augment language model training data for keyword search in conversational Cantonese speech. MT-based data augmentation is applied to two language pairs: English-Lithuanian and English-Amharic. Using filtered N-best MT hypotheses for language modeling is found to perform better than just using the 1-best translation. Target language texts collected from the Web and filtered to select conversational-like data are used in several manners. In addition to using Web data for training the language model of the speech recognizer, we further investigate using this data to improve the language model and phrase table of the MT system to get better translations of the English data. Finally, generating text data with a character-based recurrent neural network is investigated. This approach allows new word forms to be produced, providing a way to reduce the out-of-vocabulary rate and thereby improve keyword spotting performance. We study how these different methods of language model data augmentation impact speech-to-text and keyword spotting performance for the Lithuanian and Amharic languages. The best results are obtained by combining all of the explored methods.

Fri-S&T-2 : Show & Tell Session 2

Market Street Foyer, 14:30–16:30, Friday, 9 Sept. 2016
Chairs: Nicolas Scheffer, Shiva Sundaram

STON: Efficient Subtitling in Dutch Using State-of-the-Art Tools

Lyan Verwimp[1], Brecht Desplanques[2], Kris Demuynck[2], Joris Pelemans[1], Marieke Lycke[3], Patrick Wambacq[1]; [1]Katholieke Universiteit Leuven, Belgium; [2]Ghent University, Belgium; [3]VRT, Belgium

Fri-S&T-2-1, Time: 14:30

We present a modular video subtitling platform that integrates speech/non-speech segmentation, speaker diarisation, language identification, Dutch speech recognition with state-of-the-art acous-

NOTES

tic models and language models optimised for efficient subtitling, appropriate pre- and postprocessing of the data and alignment of the final result with the video fragment. Moreover, the system is able to learn from subtitles that are newly created. The platform is developed for the Flemish national broadcaster VRT in the context of the project STON, and enables the easy upload of a new fragment and inspection of both the timings and results of each step in the subtitling process.

An Automatic Training Tool for Air Traffic Control Training

Petr Stanislav, Luboš Šmídl, Jan Švec; University of West Bohemia, Czech Republic

Fri-S&T-2-2, Time: 14:30

In this paper we presents an automatic training tool (ATT) for air traffic control officer (ATCO) trainees. It was developed using our cloud-based speech recognition and text-to-speech systems and allows dynamically generate the content. Our system significantly expands the available training materials, allowing ATCOs practice the basics of communication and phraseology. Furthermore, the automatic training tool is designed generally to be used for teaching in various areas, from specialized skills to a simple general knowledge.

Digitala: An Augmented Test and Review Process Prototype for High-Stakes Spoken Foreign Language Examination

Reima Karhila[1], Aku Rouhe[1], Peter Smit[1], André Mansikkaniemi[1], Heini Kallio[2], Erik Lindroos[2], Raili Hildén[2], Martti Vainio[2], Mikko Kurimo[1]; [1]Aalto University, Finland; [2]University of Helsinki, Finland

Fri-S&T-2-3, Time: 14:30

This paper introduces the first prototype for a computerised examination procedure of spoken foreign languages in Finland, intended for national scale upper secondary school final examinations. Speech technology and profiling of reviewers are used to minimise the otherwise massive reviewing effort.

Exploring Collections of Multimedia Archives Through Innovative Interfaces in the Context of Digital Humanities

Géraldine Damnati, Delphine Charlet, Marc Denjean; Orange Labs, France

Fri-S&T-2-4, Time: 14:30

STIK is a platform that gathers Speech, Texts and Images of Knowledge. It allows browsing and navigating through collections of multimedia, facilitating access to archives in the domain of Knowledge resources. STIK includes a back-end with a specific automatic metadata extraction pipeline, a front-end with innovative interfaces for navigating within a document and a specific implementation of a search engine with dedicated key-word search functionality. It gathers multimedia contents from Canal-U, a French institution that exploits audiovisual archives produced by Higher Education and Research, with various formats and various academic disciplines. STIK is a contribution to the emerging domain of Digital Humanities.

Fri-O-3-1 : Feature Extraction and Acoustic Modeling Using Neural Networks for ASR

Grand Ballroom A, 17:00–19:00, Friday, 9 Sept. 2016
Chairs: Hui Jiang, Kate Knill

Learning Neural Network Representations Using Cross-Lingual Bottleneck Features with Word-Pair Information

*Yougen Yuan[1], Cheung-Chi Leung[2], Lei Xie[1], Bin Ma[2], Haizhou Li[2]; [1]Northwestern Polytechnical University, China; [2]A*STAR, Singapore*

Fri-O-3-1-1, Time: 17:00

We assume that only word pairs identified by human are available in a low-resource target language. The word pairs are parameterized by a bottleneck feature (BNF) extractor that is trained using transcribed data in a high-resource language. The cross-lingual BNFs of the word pairs are used for training another neural network to generate a new feature representation in the target language. Pairwise learning of frame-level and word-level feature representations are investigated. Our proposed feature representations were evaluated in a word discrimination task on the Switchboard telephone speech corpus. Our learned features could bring 27.5% relative improvement over the previously best reported result on the task.

Novel Front-End Features Based on Neural Graph Embeddings for DNN-HMM and LSTM-CTC Acoustic Modeling

Yuzong Liu[1], Katrin Kirchhoff[2]; [1]Amazon.com, USA; [2]University of Washington, USA

Fri-O-3-1-2, Time: 17:20

In this paper we investigate neural graph embeddings as front-end features for various deep neural network (DNN) architectures for speech recognition. Neural graph embedding features are produced by an autoencoder that maps graph structures defined over speech samples to a continuous vector space. The resulting feature representation is then used to augment the standard acoustic features at the input level of a DNN classifier. We compare two different neural graph embedding methods, one based on a local neighborhood graph encoding, and another based on a global similarity graph encoding. They are evaluated in DNN-HMM-based and LSTM-CTC-based ASR systems on a 110-hour Switchboard conversational speech recognition task. Significant improvements in word error rates are achieved by both methods in the DNN-HMM system, and by global graph embeddings in the LSTM-CTC system.

Articulatory Feature Extraction Using CTC to Build Articulatory Classifiers Without Forced Frame Alignments for Speech Recognition

Basil Abraham, S. Umesh, Neethu Mariam Joy; IIT Madras, India

Fri-O-3-1-3, Time: 17:40

Articulatory features provide robustness to speaker and environment variability by incorporating speech production knowledge. Pseudo articulatory features are a way of extracting articulatory

NOTES

features using articulatory classifiers trained from speech data. One of the major problems faced in building articulatory classifiers is the requirement of speech data aligned in terms of articulatory feature values at frame level. Manually aligning data at frame level is a tedious task and alignments obtained from the phone alignments using phone-to-articulatory feature mapping are prone to errors. In this paper, a technique using connectionist temporal classification (CTC) criterion to train an articulatory classifier using bidirectional long short-term memory (BLSTM) recurrent neural network (RNN) is proposed. The CTC criterion eliminates the need for forced frame level alignments. Articulatory classifiers were also built using different neural network architectures like deep neural networks (DNN), convolutional neural network (CNN) and BLSTM with frame level alignments and were compared to the proposed approach of using CTC. Among the different architectures, articulatory features extracted using articulatory classifiers built with BLSTM gave better recognition performance. Further, the proposed approach of BLSTM with CTC gave the best overall performance on both SVitchboard (6 hours) and Switchboard 33 hours data set.

On the Role of Nonlinear Transformations in Deep Neural Network Acoustic Models

Tasha Nagamine [1], Michael L. Seltzer [2], Nima Mesgarani [1]; [1] Columbia University, USA; [2] Microsoft, USA

Fri-O-3-1-4, Time: 18:00

Deep neural networks (DNNs) are widely utilized for acoustic modeling in speech recognition systems. Through training, DNNs used for phoneme recognition nonlinearly transform the time-frequency representation of a speech signal into a sequence of invariant phonemic categories. However, little is known about how this nonlinear mapping is performed and what its implications are for the classification of individual phones and phonemic categories. In this paper, we analyze a sigmoid DNN trained for a phoneme recognition task and characterized several aspects of the nonlinear transformations that occur in hidden layers. We show that the function learned by deeper hidden layers becomes increasingly nonlinear, and that network selectively warps the feature space so as to increase the discriminability of acoustically similar phones, aiding in their classification. We also demonstrate that the nonlinear transformation of the feature space in deeper layers is more dedicated to the phone instances that are more difficult to discriminate, while the more separable phones are dealt with in the superficial layers of the network. This study describes how successive nonlinear transformations are applied to the feature space non-uniformly when a deep neural network model learns categorical boundaries, which may partly explain their superior performance in pattern classification applications.

Complex Linear Projection (CLP): A Discriminative Approach to Joint Feature Extraction and Acoustic Modeling

Ehsan Variani, Tara N. Sainath, Izhak Shafran, Michiel Bacchiani; Google, USA

Fri-O-3-1-5, Time: 18:20

State-of-the-art automatic speech recognition (ASR) systems typically rely on pre-processed features. This paper studies the time-frequency duality in ASR feature extraction methods and proposes extending the standard acoustic model with a complex-valued linear projection layer to learn and optimize features that minimize standard cost functions such as cross-entropy. The proposed Complex Linear Projection (CLP) features achieve superior performance compared to pre-processed Log Mel features.

Modeling Time-Frequency Patterns with LSTM vs. Convolutional Architectures for LVCSR Tasks

Tara N. Sainath, Bo Li; Google, USA

Fri-O-3-1-6, Time: 18:40

Various neural network architectures have been proposed in the literature to model 2D correlations in the input signal, including convolutional layers, frequency LSTMs and 2D LSTMs such as time-frequency LSTMs, grid LSTMs and ReNet LSTMs. It has been argued that frequency LSTMs can model translational variations similar to CNNs, and 2D LSTMs can model even more variations [1], but no proper comparison has been done for speech tasks. While convolutional layers have been a popular technique in speech tasks, this paper compares convolutional and LSTM architectures to model time-frequency patterns as the first layer in an LDNN [2] architecture. This comparison is particularly interesting when the convolutional layer degrades performance, such as in noisy conditions or when the learned filterbank is not constant-Q [3]. We find that grid-LDNNs offer the best performance of all techniques, and provide between a 1–4% relative improvement over an LDNN and CLDNN on 3 different large vocabulary Voice Search tasks.

Fri-O-3-2 : Special Session: The Speakers in the Wild (SITW) Speaker Recognition Challenge

Grand Ballroom BC, 17:00–19:00, Friday, 9 Sept. 2016

Chairs: Mitchell McLaren, Aaron Lawson, Luciana Ferrer, Diego Castan

The Speakers in the Wild (SITW) Speaker Recognition Database

Mitchell McLaren [1], Luciana Ferrer [2], Diego Castan [1], Aaron Lawson [1]; [1] SRI International, USA; [2] Universidad de Buenos Aires, Argentina

Fri-O-3-2-1, Time: 17:00

The Speakers in the Wild (SITW) speaker recognition database contains hand-annotated speech samples from open-source media for the purpose of benchmarking text-independent speaker recognition technology on single and multi-speaker audio acquired across unconstrained or “wild” conditions. The database consists of recordings of 299 speakers, with an average of eight different sessions per person. Unlike existing databases for speaker recognition, this data was not collected under controlled conditions and thus contains real noise, reverberation, intra-speaker variability and compression artifacts. These factors are often convolved in the real world, as the SITW data shows, and they make SITW a challenging database for single- and multi-speaker recognition.

NOTES

The 2016 Speakers in the Wild Speaker Recognition Evaluation

Mitchell McLaren[1], Luciana Ferrer[2], Diego Castan[1], Aaron Lawson[1]; [1]SRI International, USA; [2]Universidad de Buenos Aires, Argentina

Fri-O-3-2-2, Time: 17:15

The newly collected Speakers in the Wild (SITW) database was central to a text-independent speaker recognition challenge held as part of a special session at Interspeech 2016. The SITW database is composed of audio recordings from 299 speakers collected from open source media, with an average of 8 sessions per speaker. The recordings contain unconstrained or "wild" acoustic conditions, rarely found in large speaker recognition datasets, and multi-speaker recordings for both speaker enrollment and verification. This article provides details of the SITW speaker recognition challenge and analysis of evaluation results. There were 25 international teams involved in the challenge of which 11 teams participated in an evaluation track. Teams were tasked with applying existing and novel speaker recognition algorithms to the challenges associated with the real world conditions of SITW. We provide an analysis of some of the top performing systems submitted during the evaluation and provide future research directions.

Analysis of Speaker Recognition Systems in Realistic Scenarios of the SITW 2016 Challenge

Ondřej Novotný, Pavel Matějka, Oldřich Plchot, Ondřej Glembek, Lukáš Burget, Jan Černocký; Brno University of Technology, Czech Republic

Fri-O-3-2-3, Time: 17:30

In this paper, we summarize our efforts for the Speakers In The Wild (SITW) challenge, and we present our findings with this new dataset for speaker recognition. Apart from the standard comparison of different SRE systems, we analyze the use of diarization for dealing with audio segments containing multiple speakers, as in part of the newly introduced enrollment and test protocols, diarization is a necessary system component. Our state-of-the-art systems used in this work utilize both cepstral and DNN-based bottleneck features and are based on i-vectors followed by Probabilistic Linear Discriminant Analysis (PLDA) classifier and logistic regression calibration/fusion. We present both narrow-band (8 kHz) and wide-band (16 kHz) systems together with their fusions.

A Speaker Recognition System for the SITW Challenge

Oleg Kudashev[1], Sergey Novoselov[1], Konstantin Simonchik[1], Alexandr Kozlov[2]; [1]ITMO University, Russia; [2]Speech Technology Center, Russia

Fri-O-3-2-4, Time: 17:45

This paper presents an ITMO university system submitted to the Speakers in the Wild (SITW) Speaker Recognition Challenge. During evaluation track of the SITW challenge we explored conventional universal background model (UBM) Gaussian mixture model (GMM) i-vector systems and recently developed DNN-posteriors based i-vector systems. The systems were investigated under the real-world media channel conditions represented in the challenge. This paper discusses practical issues of the robust i-vector systems training and performs investigation of denoising autoencoder (DAE) based back-end when applied to "in the wild" conditions. Our speak-er diarization approach for "multi-speaker in the file" conditions is also briefly presented in the paper. Experiments per-formed on the evaluation dataset demonstrate that DNN- based i-vector systems are superior to the UBM-GMM based sys-tems and applying DAE-based back-end helps to improve system performance.

Speakers In The Wild (SITW): The QUT Speaker Recognition System

H. Ghaemmaghami, M.H. Rahman, Ivan Himawan, David Dean, Ahilan Kanagasundaram, Sridha Sridharan, Clinton Fookes; Queensland University of Technology, Australia

Fri-O-3-2-5, Time: 18:00

This paper presents the QUT speaker recognition system, as a competing system in the *Speakers In The Wild* (SITW) speaker recognition challenge. Our proposed system achieved an overall ranking of second place, in the main *core-core* condition evaluations of the SITW challenge. This system uses an i-vector/PLDA approach, with domain adaptation and a deep neural network (DNN) trained to provide feature statistics. The statistics are accumulated by using class posteriors from the DNN, in place of GMM component posteriors in a typical GMM-UBM i-vector/PLDA system. Once the statistics have been collected, the i-vector computation is carried out as in a GMM-UBM based system. We apply domain adaptation to the extracted i-vectors to ensure robustness against dataset variability, PLDA modelling is used to capture speaker and session variability in the i-vector space, and the processed i-vectors are compared using the batch likelihood ratio. The final scores are calibrated to obtain the calibrated likelihood scores, which are then used to carry out speaker recognition and evaluate the performance of the system. Finally, we explore the practical application of our system to the *core-multi* condition recordings of the SITW data and propose a technique for speaker recognition in recordings with multiple speakers.

AUT System for SITW Speaker Recognition Challenge

Abbas Khosravani, Mohammad Mehdi Homayounpour; Amirkabir University of Technology, Iran

Fri-O-3-2-6, Time: 18:15

This document intends to present AUT speaker recognition system submitted to SITW (Speakers in the Wild) speaker recognition challenge. This challenge aims to provide real world data across a wide range of acoustic and environmental conditions in the context of automatic speaker recognition so as to facilitate the development of new algorithms. The presented system is based on the state-of-the-art *i*-vector/PLDA and source normalization techniques. The system has been developed on publically available databases and evaluated on the data provided by SITW challenge. Taking advantage of the challenge development data, our experiments indicate that source normalization can help speaker recognition system to better adapt to the evaluation condition. Post evaluation analysis is conducted on the conditions of SITW database.

LIA System for the SITW Speaker Recognition Challenge

Waad Ben Kheder, Moez Ajili, Pierre-Michel Bousquet, Driss Matrouf, Jean-François Bonastre; LIA, France

Fri-O-3-2-7, Time: 18:30

This paper presents the speaker verification systems developed in

NOTES

the LIA lab at the University of Avignon for the SITW (Speakers In The Wild) challenge. We present the algorithms used to deal with additive noise, short utterances and propose an improved scoring scheme using a discriminative classifier and integrating the homogeneity of the two compared recordings. Due to the heterogeneity of this database (presence of background noise, reverberation, Lombard effect, etc.), it is hard to analyze the contribution of individual techniques used to deal with each problem. For this reason, a subset of the trials will be studied for each algorithm in order to emphasize its contribution.

Investigating Various Diarization Algorithms for Speaker in the Wild (SITW) Speaker Recognition Challenge

Yi Liu, Yao Tian, Liang He, Jia Liu; Tsinghua University, China

Fri-O-3-2-8, Time: 18:45

Collecting training data for real-world text-independent speaker recognition is challenging. In practice, utterances for a specific speaker are often mixed with many other acoustic signals. To guarantee the recognition performance, the segments spoken by target speakers should be precisely picked out. An automatic detection could be developed to reduce the cost of expensive human handmade annotations. One way to achieve this goal is by using speaker diarization as a pre-processing step in the speaker enrollment phase. To this end, three speaker diarization algorithms based on Bayesian information criterion (BIC), agglomerative information bottleneck (aIB) and i-vector are investigated in this paper. The corresponding impacts on the results of speaker recognition system are also studied. Experiments conducted on Speaker in the Wild (SITW) Speaker Recognition Challenge (SRC) 2016 showed that the utilization of a proper speaker diarization improves the overall performance. Some more efforts are made to combine these methods together as well.

Fri-O-3-3 : Non-Native Speech Perception

Bayview A, 17:00–19:00, Friday, 9 Sept. 2016
Chairs: Outi Tuomainen, Florian Hintz

Does the Importance of Word-Initial and Word-Final Information Differ in Native versus Non-Native Spoken-Word Recognition?

Odette Scharenborg [1], Juul Coumans [1], Sofoklis Kakouros [2], Roeland van Hout [1]; [1] Radboud Universiteit Nijmegen, The Netherlands; [2] Aalto University, Finland

Fri-O-3-3-1, Time: 17:00

This paper investigates whether the importance and use of word-initial and word-final information in spoken-word recognition is dependent on whether one is listening in a native or a non-native language and on the presence of background noise. Native English and non-native Dutch and Finnish listeners participated in an English word recognition experiment, where either a word's onset or offset was masked by speech-shaped noise with different signal-to-noise ratios. The results showed that for all listener groups the masking of word onset information was more detrimental to spoken-word recognition than the masking of word offset information. The reliance on word-initial information was larger in harder listening conditions for the English but not so for the Dutch and Finnish listeners. Moreover, no significant differences in the use of word-initial and word-final information were found between the two non-native listener groups. Taken together, these results show that the reliance on word-initial information in deteriorating listening conditions seems to be dependent on whether one is listening in one's native or a non-native language rather than on the listener's native language.

The Effect of Sentence Accent on Non-Native Speech Perception in Noise

Odette Scharenborg [1], Elea Kolkman [1], Sofoklis Kakouros [2], Brechtje Post [3]; [1] Radboud Universiteit Nijmegen, The Netherlands; [2] Aalto University, Finland; [3] University of Cambridge, UK

Fri-O-3-3-2, Time: 17:20

This paper investigates the uptake and use of prosodic information signalling sentence accent during native and non-native speech perception in the presence of background noise. A phoneme monitoring experiment was carried out in which English, Dutch, and Finnish listeners were presented with target phonemes in semantically unpredictable yet meaningful English sentences. Sentences were presented in different levels of speech-shaped noise and, crucially, in two prosodic contexts in which the target-bearing word was either deaccented or accented. Results showed that overall performance was high for both the native and the non-native listeners; however, where native listeners seemed able to partially overcome the problems at the acoustic level in degraded listening conditions by using prosodic information signalling upcoming sentence accent, non-native listeners could not do so to the same extent. These results support the hypothesis that the performance difference between native and non-native listeners in the presence of background noise is, at least partially, caused by a reduced exploitation of contextual information during speech processing by non-native listeners.

The Effects of Modified Speech Styles on Intelligibility for Non-Native Listeners

Martin Cooke [1], Maria Luisa Garcia Lecumberri [2]; [1] Ikerbasque, Spain; [2] Universidad del País Vasco, Spain

Fri-O-3-3-3, Time: 17:40

Speech output, including modified and synthetic speech, is used increasingly in natural settings where message reception might be affected by noise. Recent evaluations have demonstrated the effect of different speech styles on intelligibility for native listeners, but their impact on listening in a second language is less well-understood. The current study measured the intelligibility of four speech styles in the presence of stationary and fluctuating maskers for a non-native listener cohort, and compared the results with those of native listeners on the same task. Both groups showed a similar pattern of effects, but the scale of intelligibility gains and losses with respect to plain speech was significantly compressed for the non-native group relative to native listeners. In addition, non-native listeners identified speech from the four styles in the absence of noise, revealing that styles shown to be beneficial in noise lost their benefits or were harmful in quiet conditions. This result suggests that while enhanced styles lead to gains by reducing the effect of masking noise, the same styles distort the acoustic-phonetic integrity of the speech signal. More work is needed to develop speech modification approaches that simultaneously preserve speech information and promote unmasking.

Notes

The Influence of Language Experience on the Categorical Perception of Vowels: Evidence from Mandarin and Korean

Hao Zhang[1], Fei Chen[1], Nan Yan[1], Lan Wang[1], Feng Shi[2], Manwa L. Ng[3]; [1]Chinese Academy of Sciences, China; [2]Nankai University, China; [3]University of Hong Kong, China

Fri-O-3-3-4, Time: 18:00

Previous research on categorical perception of speech sounds has demonstrated a strong influence of language experience on the categorical perception of consonants and lexical tones. In order to explore the influence of language experience on vowel perception, the present study investigated the perceptual performance for Mandarin and Korean listeners along a vowel continuum, which spanned three vowel categories /a/, /ɜ/, and /u/. The results showed that both language groups exhibited categorical features in vowel perception, with a sharper categorical boundary of /ɜ/-/u/ than that of /a/-/ɜ/. Moreover, the differences found between the two groups revealed that the Korean listeners' perception tended to be more categorical along the /a/-/ɜ/-/u/ vowel continuum than that of the Mandarin listeners. Furthermore, the Mandarin listeners tended to label stimuli more often as /a/ and less often as /u/ than the Korean counterparts. These perceptual differences between the Mandarin and Korean groups might be attributed to the different acoustic distribution in the F1×F2 vowel space of the two different native languages.

Multiple Influences on Vocabulary Acquisition: Parental Input Dominates

Dominic W. Massaro; University of California at Santa Cruz, USA

Fri-O-3-3-5, Time: 18:20

How spoken language is acquired has been an active area of inquiry in linguistic, psychological, and speech science. New advances in this controversial field are promising given the recent accumulation of large databases of children's speech understanding and production, as well as various properties of words. This paper explores the contribution of a variety of potential influences on vocabulary acquisition including difficulty of articulation, iconicity, log parental input frequency, lexical category, and imageability. The influence of difficulty of articulation, iconicity ratings, and imagery ratings decreased more or less linearly with increasing age. Lexical category effects were fairly small. Parental input in terms of child directed speech has by far the largest influence. Multiple regressions with these variables give a fairly complete account of spoken vocabulary acquisition. The increasing availability of large databases promises progress in this area of inquiry.

Can Intensive Exposure to Foreign Language Sounds Affect the Perception of Native Sounds?

Jian Gong[1], Maria Luisa Garcia Lecumberri[2], Martin Cooke[3]; [1]JUST, China; [2]Universidad del País Vasco, Spain; [3]Ikerbasque, Spain

Fri-O-3-3-6, Time: 18:40

A possible side-effect of exposure to non-native sounds is a change in the way we perceive native sounds. Previous studies have demonstrated that native speakers' speech production can change as a result of learning a new language, but little work has been carried out to measure the perceptual consequences of exposure. The current study examined how intensive exposure to Spanish intervocalic consonants affected Chinese learners with no prior experience of Spanish. Before, during and after a training period, listeners undertook both an adaptive noise task, which measured the noise level at which listeners could identify native language consonants, and an assimilation task, in which listeners assigned Spanish consonants to Chinese consonant categories. Listeners exhibited a significantly reduced noise tolerance for the Chinese consonants /l/ and /w/ following exposure to Spanish. These two consonants also showed the largest reductions in Spanish to Chinese category assimilations. Taken together, these findings suggest that Chinese listeners modified their native language categories boundaries as a result of exposure to Spanish sounds in order to accommodate them, and that as a consequence their identification performance in noise reduced. Some differences between the two sounds in the time-course of recovery from perceptual adaptation were observed.

Fri-O-3-4 : Behavioral Signal Processing and Speaker State and Traits Analytics

Bayview B, 17:00–19:00, Friday, 9 Sept. 2016

Chairs: Joseph Tepperman, Chi-Chun (Jeremy) Lee

Privacy-Preserving Speech Analytics for Automatic Assessment of Student Collaboration

Nikoletta Bassiou, Andreas Tsiartas, Jennifer Smith, Harry Bratt, Colleen Richey, Elizabeth Shriberg, Cynthia D'Angelo, Nonye Alozie; SRI International, USA

Fri-O-3-4-1, Time: 17:00

This work investigates whether nonlexical information from speech can automatically predict the quality of small-group collaborations. Audio was collected from students as they collaborated in groups of three to solve math problems. Experts in education annotated 30-second time windows by hand for collaboration quality. Speech activity features (computed at the group level) and spectral, temporal and prosodic features (extracted at the speaker level) were explored. After the latter were transformed from the speaker level to the group level, features were fused. Results using support vector machines and random forests show that feature fusion yields best classification performance. The corresponding unweighted average F_1 measure on a 4-class prediction task ranges between 40% and 50%, significantly higher than chance (12%). Speech activity features alone are strong predictors of collaboration quality, achieving an F_1 measure between 35% and 43%. Speaker-based acoustic features alone achieve lower classification performance, but offer value in fusion. These findings illustrate that the approach under study offers promise for future monitoring of group dynamics, and should be attractive for many collaboration activity settings in which privacy is desired.

NOTES

Complexity in Prosody: A Nonlinear Dynamical Systems Approach for Dyadic Conversations; Behavior and Outcomes in Couples Therapy

Md. Nasir [1], Brian Baucom [2], Shrikanth S. Narayanan [1], Panayiotis Georgiou [1]; [1] University of Southern California, USA; [2] University of Utah, USA

Fri-O-3-4-2, Time: 17:20

In this paper, we model dyadic human conversational interactions from a nonlinear dynamical systems perspective. We focus on deriving measures of the underlying system complexity using the observed dyadic behavioral signals. Specifically, we analyze different measures of complexity in prosody of speech (pitch and energy) during dyadic conversations of couples with marital conflict. We evaluate the importance of these measures as features by correlating them with different behavioral attributes of the couple codified in terms of behavioral codes. Furthermore, we investigate the relation between the computed complexity and outcomes of couples therapy. The results show that the derived complexity measures are more correlated to session level behavioral codes, and to the marital therapy outcomes, compared to traditional speech prosody features. It shows that nonlinear dynamical analysis of speech acoustic features can be a useful tool for behavioral analysis.

Couples Behavior Modeling and Annotation Using Low-Resource LSTM Language Models

Shao-Yen Tseng [1], Sandeep Nallan Chakravarthula [1], Brian Baucom [2], Panayiotis Georgiou [1]; [1] University of Southern California, USA; [2] University of Utah, USA

Fri-O-3-4-3, Time: 17:40

Observational studies on couple interactions are often based on manual annotations of a set of behavior codes. Such annotations are expensive, time-consuming, and often suffer from low inter-annotator agreement. In previous studies it has been shown that the lexical channels contain sufficient information for capturing behavior and predicting the interaction labels, and various automated processes using language models have been proposed. However, current methods are restricted to a small context window due to the difficulty of training language models with limited data as well as the lack of frame-level labels. In this paper we investigate the application of recurrent neural networks for capturing behavior trajectories through larger context windows. We solve the issue of data sparsity and improve robustness by introducing out-of-domain knowledge through pretrained word representations. Finally, we show that our system can accurately estimate true rating values of couples interactions using a fusion of the frame-level behavior trajectories. The ratings predicted by our proposed system achieve inter-annotator agreements comparable to those of trained human annotators.

Importantly, our system promises robust handling of out of domain data, exploitation of longer context, on-line feedback with continuous labels and easy fusion with other modalities.

Speech Likability and Personality-Based Social Relations: A Round-Robin Analysis over Communication Channels

Laura Fernández Gallardo, Benjamin Weiss; T-Labs, Germany

Fri-O-3-4-4, Time: 18:00

The Social Relations Model is well-known for analyses of interpersonal attraction. As a novelty in this paper, the model is applied to assess different effects on likability ratings from speech only. A group of 30 unacquainted participants is considered in our experiment. Their voices were recorded and transmitted through communication channels, and ratings of speech likability and speaker personality were then collected from the same individuals following a round-robin approach. This setup enabled us to detect the influence of participants' personality and of narrowband and wideband speech on the sources of variance according to the Social Relations Model. An analysis of acoustic correlates of speech likability has also been conducted, which shows differences in the relevance of speech features and in the description of likability ratings depending on the speech bandwidth.

Behavioral Coding of Therapist Language in Addiction Counseling Using Recurrent Neural Networks

Bo Xiao [1], Doğan Can [1], James Gibson [1], Zac E. Imel [2], David C. Atkins [3], Panayiotis Georgiou [1], Shrikanth S. Narayanan [1]; [1] University of Southern California, USA; [2] University of Utah, USA; [3] University of Washington, USA

Fri-O-3-4-5, Time: 18:20

Manual annotation of human behaviors with domain specific codes is a primary method of research and treatment fidelity evaluation in psychotherapy. However, manual annotation has a prohibitively high cost and does not scale to coding large amounts of psychotherapy session data. In this paper, we present a case study of modeling therapist language in addiction counseling, and propose an automatic coding approach. The task objective is to code therapist utterances with domain specific codes. We employ Recurrent Neural Networks (RNNs) to predict these behavioral codes based on session transcripts. Experiments show that RNNs outperform the baseline method using Maximum Entropy models. The model with bi-directional Gated Recurrent Units and domain specific word embeddings achieved the highest overall accuracy. We also briefly discuss about client code prediction and comparison to previous work.

Factor Analysis Based Speaker Normalisation for Continuous Emotion Prediction

Ting Dang, Vidhyasaharan Sethu, Eliathamby Ambikairajah; University of New South Wales, Australia

Fri-O-3-4-6, Time: 18:40

Speaker variability has been shown to be a significant confounding factor in speech based emotion classification systems and a number of speaker normalisation techniques have been proposed. However, speaker normalisation in systems that predict continuous multidimensional descriptions of emotion such as arousal and valence has not been explored. This paper investigates the effect

Notes

of speaker variability in such speech based continuous emotion prediction systems and proposes a factor analysis based speaker normalisation technique. The proposed technique operates directly on the feature space and decomposes it into speaker and emotion specific sub-spaces. The proposed technique is validated on both the USC CreativeIT database and the SEMAINE database and leads to improvements of 8.2% and 11.0% (in terms of correlation coefficient) on the two databases respectively when predicting arousal.

Fri-O-3-5 : Spoken Term Detection

Seacliff BCD, 17:00–19:00, Friday, 9 Sept. 2016
Chairs: Isabel Trancoso, Roland Kuhn

Subspace Detection of DNN Posterior Probabilities via Sparse Representation for Query by Example Spoken Term Detection

Dhananjay Ram, Afsaneh Asaei, Hervé Bourlard; Idiap Research Institute, Switzerland

Fri-O-3-5-1, Time: 17:00

We cast the query by example spoken term detection (QbE-STD) problem as subspace detection where query and background subspaces are modeled as union of low-dimensional subspaces. The speech exemplars used for subspace modeling are class-conditional posterior probabilities estimated using deep neural network (DNN). The query and background training exemplars are exploited to model the underlying low-dimensional subspaces through dictionary learning for sparse representation. Given the dictionaries characterizing the query and background subspaces, QbE-STD is performed based on the ratio of the two corresponding sparse representation reconstruction errors. The proposed subspace detection method can be formulated as the generalized likelihood ratio test for composite hypothesis testing. The experimental evaluation demonstrate that the proposed method is able to detect the query given a single example and performs significantly better than a highly competitive QbE-STD baseline system based on dynamic time warping (DTW) for exemplar matching.

Unsupervised Bottleneck Features for Low-Resource Query-by-Example Spoken Term Detection

*Hongjie Chen[1], Cheung-Chi Leung[2], Lei Xie[1], Bin Ma[2], Haizhou Li[2]; [1]Northwestern Polytechnical University, China; [2]A*STAR, Singapore*

Fri-O-3-5-2, Time: 17:20

We propose a framework which ports Dirichlet Gaussian mixture model (DPGMM) based labels to deep neural network (DNN). The DNN trained using the unsupervised labels is used to extract a low-dimensional unsupervised speech representation, named as unsupervised bottleneck features (uBNFs), which capture considerable information for sound cluster discrimination. We investigate the performance of uBNF in query-by-example spoken term detection (QbE-STD) on the TIMIT English speech corpus. Our uBNF performs comparably with the cross-lingual bottleneck features (BNFs) extracted from a DNN trained using 171 hours of transcribed telephone speech in another language (Mandarin Chinese). With the score fusion of uBNFs and cross-lingual BNFs, we gain about 10% relative improvement in terms of mean average precision (MAP) comparing with the cross-lingual BNFs. We also study the performance of the framework with different input features and different lengths of temporal context.

A Nonparametric Bayesian Approach for Spoken Term Detection by Example Query

Amir Hossein Harati Nejad Torbati, Joseph Picone; Temple University, USA

Fri-O-3-5-3, Time: 17:40

State of the art speech recognition systems use data-intensive context-dependent phonemes as acoustic units. However, these approaches do not translate well to low resourced languages where large amounts of training data is not available. For such languages, automatic discovery of acoustic units is critical. In this paper, we demonstrate the application of nonparametric Bayesian models to acoustic unit discovery. We show that the discovered units are correlated with phonemes and therefore are linguistically meaningful.

We also present a spoken term detection (STD) by example query algorithm based on these automatically learned units. We show that our proposed system produces a P@N of 61.2% and an EER of 13.95% on the TIMIT dataset. The improvement in the EER is 5% while P@N is only slightly lower than the best reported system in the literature.

Rescoring Hypothesized Detections of Out-of-Vocabulary Keywords Using Subword Samples

*Van Tung Pham[1], Haihua Xu[2], Xiong Xiao[2], Nancy F. Chen[3], Eng Siong Chng[1], Haizhou Li[1]; [1]NTU, Singapore; [2]TL@NTU, Singapore; [3]A*STAR, Singapore*

Fri-O-3-5-4, Time: 18:00

Rescoring hypothesized detections, using keyword's audio samples extracted from training data, is an effective way to improve the performance of a Keyword Search (KWS) system. Unfortunately such rescoring framework cannot be applied directly to Out-of-Vocabulary (OOV) keywords since there is no sample in the training data. To address this limitation, we propose two techniques for OOV keywords in this work. The first technique generates samples for an OOV keyword by concatenating samples of its constituent subwords. The second technique splits hypothesized detections into segments, then estimates the acoustic similarities between detections and subword's samples according to the similarities between segments and these samples. The similarity scores from these two techniques are used to rescore and re-rank the list of detections returned by the automatic speech recognition (ASR) systems. The experiments show that incorporating the proposed similarity scores results in a better separation between the correct and false alarm detections than using the ASR scores alone. Furthermore, experimental results on the NIST OpenKWS15 Evaluation show that rescoring with the proposed similarity scores significantly outperforms the raw ASR scores, and other methods that do not use the similarity scores, in both Maximum Term Weighted Value (MTWV) and Mean Average Precision (MAP) metrics.

Unrestricted Vocabulary Keyword Spotting Using LSTM-CTC

Yimeng Zhuang, Xuankai Chang, Yanmin Qian, Kai Yu; Shanghai Jiao Tong University, China

Fri-O-3-5-5, Time: 18:20

Keyword spotting (KWS) aims to detect predefined keywords in continuous speech. Recently, direct deep learning approaches have been used for KWS and achieved great success. However, these approaches mostly assume fixed keyword vocabulary and require significant retraining efforts if new keywords are to be detected. For

NOTES

unrestricted vocabulary, HMM based keyword-filler framework is still the mainstream technique. In this paper, a novel deep learning approach is proposed for unrestricted vocabulary KWS based on Connectionist Temporal Classification (CTC) with Long Short-Term Memory (LSTM). Here, an LSTM is trained to discriminant phones with the CTC criterion. During KWS, an arbitrary keyword can be specified and it is represented by one or more phone sequences. Due to the property of peaky phone posteriors of CTC, the LSTM can produce a phone lattice. Then, a fast substring matching algorithm based on minimum edit distance is used to search the keyword phone sequence on the phone lattice. The approach is highly efficient and vocabulary independent. Experiments showed that the proposed approach can achieve significantly better results compared to a DNN-HMM based keyword-filler decoding system. In addition, the proposed approach is also more efficient than the DNN-HMM KWS baseline.

Interactive Spoken Content Retrieval by Deep Reinforcement Learning

Yen-Chen Wu, Tzu-Hsiang Lin, Yang-De Chen, Hung-Yi Lee, Lin-Shan Lee; National Taiwan University, Taiwan

Fri-O-3-5-6, Time: 18:40

User-machine interaction is important for spoken content retrieval. For text content retrieval, the user can easily scan through and select on a list of retrieved item. This is impossible for spoken content retrieval, because the retrieved items are difficult to show on screen. Besides, due to the high degree of uncertainty for speech recognition, the retrieval results can be very noisy. One way to counter such difficulties is through user-machine interaction. The machine can take different actions to interact with the user to obtain better retrieval results before showing to the user. The suitable actions depend on the retrieval status, for example requesting for extra information from the user, returning a list of topics for user to select, etc. In our previous work, some hand-crafted states estimated from the present retrieval results are used to determine the proper actions. In this paper, we propose to use Deep-Q-Learning techniques instead to determine the machine actions for interactive spoken content retrieval. Deep-Q-Learning bypasses the need for estimation of the hand-crafted states, and directly determine the best action base on the present retrieval status even without any human knowledge. It is shown to achieve significantly better performance compared with the previous hand-crafted states.

Fri-O-3-6 : Co-Inference of Production and Acoustics

Seacliff A, 17:00–19:00, Friday, 9 Sept. 2016
Chair: Carol Epsy-Wilson

Relating Estimated Cyclic Spectral Peak Frequency to Measured Epilarynx Length Using Magnetic Resonance Imaging

Elizabeth Godoy, Andrew Dumas, Jennifer Melot, Nicolas Malyska, Thomas F. Quatieri; MIT Lincoln Laboratory, USA

Fri-O-3-6-1, Time: 17:00

The epilarynx plays an important role in speech production, carrying information about the individual speaker and manner of articulation. However, precise acoustic behavior of this lower vocal tract structure is difficult to establish. Focusing on acoustics observable in natural speech, recent spectral processing techniques isolate a unique resonance with characteristics of the epilarynx previously shown via simulation, specifically cyclicity (i.e. energy differences between the closed and open phases of the glottal cycle) in a 3–5kHz region observed across vowels. Using Magnetic Resonance Imaging (MRI), the present work relates this estimated cyclic peak frequency to measured epilarynx length. Assuming a simple quarter wavelength relationship, the cavity length estimated from the cyclic peak frequency is shown to be directly proportional (linear fit slope =1.1) and highly correlated ($\rho = 0.85$, pval<10^{-4}) to the measured epilarynx length across speakers. Results are discussed, as are implications in speech science and application domains.

Acoustic-to-Articulatory Inversion Mapping Based on Latent Trajectory Gaussian Mixture Model

Patrick Lumban Tobing [1], Tomoki Toda [2], Hirokazu Kameoka [3], Satoshi Nakamura [1]; [1] NAIST, Japan; [2] Nagoya University, Japan; [3] NTT, Japan

Fri-O-3-6-2, Time: 17:20

A maximum likelihood parameter trajectory estimation based on a Gaussian mixture model (GMM) has been successfully implemented for acoustic-to-articulatory inversion mapping. In the conventional method, GMM parameters are optimized by maximizing a likelihood function for joint static and dynamic features of acoustic-articulatory data, and then, the articulatory parameter trajectories are estimated for given the acoustic data by maximizing a likelihood function for only the static features, imposing a constraint between static and dynamic features to consider the inter-frame correlation. Due to the inconsistency of the training and mapping criterion, the trained GMM is not optimum for the mapping process. This inconsistency problem is addressed within a trajectory training framework, but it becomes more difficult to optimize some parameters, e.g., covariance matrices and mixture component sequences. In this paper, we propose an inversion mapping method based on a latent trajectory GMM (LT-GMM) as yet another way to overcome the inconsistency issue. The proposed method makes it possible to use a well-formulated algorithm, such as EM algorithm, to optimize the LT-GMM parameters, which is not feasible in the traditional trajectory training. Experimental results demonstrate that the proposed method yields higher accuracy in the inversion mapping compared to the conventional GMM-based method.

Formant Estimation and Tracking Using Deep Learning

Yehoshua Dissen, Joseph Keshet; Bar-Ilan University, Israel

Fri-O-3-6-3, Time: 17:40

Formant frequency estimation and tracking are among the most fundamental problems in speech processing. In the former task the input is a stationary speech segment such as the middle part of a vowel and the goal is to estimate the formant frequencies, whereas in the latter task the input is a series of speech frames and the goal is to track the trajectory of the formant frequencies throughout the signal. Traditionally, formant estimation and tracking is done using ad-hoc signal processing methods. In this paper we propose using machine learning techniques trained on an annotated corpus of read speech for these tasks. Our feature set is composed of LPC-based cepstral coefficients with a range of model orders and

NOTES

pitch-synchronous cepstral coefficients. Two deep network architectures are used as learning algorithms: a deep feed-forward network for the estimation task and a recurrent neural network for the tracking task. The performance of our methods compares favorably with mainstream LPC-based implementations and state-of-the-art tracking algorithms.

Convex Hull Convolutive Non-Negative Matrix Factorization for Uncovering Temporal Patterns in Multivariate Time-Series Data

Colin Vaz, Asterios Toutios, Shrikanth S. Narayanan; University of Southern California, USA

Fri-O-3-6-4, Time: 18:00

We propose the Convex Hull Convolutive Non-negative Matrix Factorization (CH-CNMF) algorithm to learn temporal patterns in multivariate time-series data. The algorithm factors a data matrix into a basis tensor that contains temporal patterns and an activation matrix that indicates the time instants when the temporal patterns occurred in the data. Importantly, the temporal patterns correspond closely to the observed data and represent a wide range of dynamics. Experiments with synthetic data show that the temporal patterns found by CH-CNMF match the data better and provide more meaningful information than the temporal patterns found by Convolutive Non-negative Matrix Factorization with sparsity constraints (CNMF-SC). Additionally, CH-CNMF applied on vocal tract constriction data yields a wider range of articulatory gestures compared to CNMF-SC. Moreover, we find that the gestures comprising the CH-CNMF basis generalize better to unseen data and capture vocal tract structure and dynamics significantly better than those comprising the CNMF-SC basis.

Majorisation-Minimisation Based Optimisation of the Composite Autoregressive System with Application to Glottal Inverse Filtering

Lauri Juvela [1], Hirokazu Kameoka [2], Manu Airaksinen [1], Junichi Yamagishi [3], Paavo Alku [1]; [1] Aalto University, Finland; [2] University of Tokyo, Japan; [3] NII, Japan

Fri-O-3-6-5, Time: 18:20

The composite autoregressive system can be used to estimate a speech source-filter decomposition in a rigorous manner, thus having potential use in glottal inverse filtering. By introducing a suitable prior, spectral tilt can be introduced into the source component estimation to better correspond to human voice production. However, the current expectation-maximisation based composite autoregressive model optimisation leaves room for improvement in terms of speed. Inspired by majorisation-minimisation techniques used for nonnegative matrix factorisation, this work derives new update rules for the model, resulting in faster convergence compared to the original approach. Additionally, we present a new glottal inverse filtering method based on the composite autoregressive system and compare it with inverse filtering methods currently used in glottal excitation modelling for parametric speech synthesis. These initial results show that the proposed method performs comparatively well, sometimes outperforming the reference methods.

F_0 Contour Analysis Based on Empirical Mode Decomposition for DNN Acoustic Modeling in Mandarin Speech Recognition

Xiaoyun Wang [1], Xugang Lu [1], Hisashi Kawai [1], Seiichi Yamamoto [2]; [1] NICT, Japan; [2] Doshisha University, Japan

Fri-O-3-6-6, Time: 18:40

Tone information provides a strong distinction for many ambiguous characters in Mandarin Chinese. The use of tonal acoustic units and F_0 related tonal features have been shown to be effective at improving the accuracy of Mandarin automatic speech recognition (ASR) systems, as F_0 contains the most prominent tonal information for distinguishing words that are phonemically identical. Both long-term temporal intonations and short-term quick variations coexist in F_0. Using untreated F_0 as an acoustic feature renders the F_0 contour patterns differently from their citation form and downplays the significance of tonal information in ASR. In this paper, we explore the empirical mode decomposition (EMD) on F_0 contours to reconstruct F_0 related tonal features with a view to removing the components that are irrelevant for Mandarin ASR.We investigate both GMM-HMM and DNN-HMM based acoustic modeling with the reconstructed tonal features. In comparison with the baseline systems using typical tonal features, our best system using reconstructed tonal features leads to a 4.5% relative word error rate reduction for the GMM-HMM system and a 3.5% relative word error rate reduction for the DNN-HMM system.

Fri-P-3-1 : Acoustic and Articulatory Phonetics

Pacific Concourse – Poster A, 17:00–19:00, Friday, 9 Sept. 2016
Chair: Marija Tabain

Vowels and Diphthongs in Cangnan Southern Min Chinese Dialect

Fang Hu, Chunyu Ge; Chinese Academy of Social Sciences, China

Fri-P-3-1-1, Time: 17:00

This paper gives an acoustic phonetic description of the vowels and diphthongs in Cangnan Southern Min Chinese dialect. Vowel formant data from 10 speakers (5 male and 5 female) show that the distribution of Cangnan monophthongs in the acoustic vowel space is of particular typological interest. Diphthong production is examined in terms of temporal organization, spectral property, and dynamic aspects. Results suggest that the production of falling diphthongs tends to be a single articulatory event with a dynamic spectral target, while the production of rising diphthongs and level diphthongs is a sequence of two spectral targets.

Diphthongization of Nuclear Vowels and the Emergence of a Tetraphthong in Hetang Cantonese

Wenqi Hu [1], Fang Hu [2], Jian Jin [1]; [1] Sun Yat-Sen University, China; [2] Chinese Academy of Social Sciences, China

Fri-P-3-1-2, Time: 17:00

This paper is an acoustic phonetic description of vowels in Hetang Cantonese, and focuses on the diphthongization of nuclear vowels. Different to the representative dialect such as Guangzhou or Hong

NOTES

Kong Cantonese, the Hetang dialect exhibits its unique characteristics regarding the phonetics and phonology of vowels. A noticeable phenomenon is the diphthongization of nuclear vowels. And, a tetraphthong [uɔ$^{\text{ɐ}}$i] emerges when the nuclear vowel is diphthongized in a triphthong.

PhonVoc: A Phonetic and Phonological Vocoding Toolkit

Milos Cernak, Philip N. Garner; Idiap Research Institute, Switzerland

Fri-P-3-1-3, Time: 17:00

We present the PhonVoc toolkit, a cascaded deep neural network (DNN) composed of speech analyser and synthesizer that use a shared phonetic and/or phonological speech representation. The free toolkit is distributed as open-source software under a BSD 3-Clause License, available at `https://github.com/idiap/phonvoc` with the pre-trained US English analysis and synthesis DNNs, and thus it is ready for immediate use.

In a broader context, the toolkit implements training and testing of the analysis by synthesis heuristic model. It is thus designed for the wider speech community working in acoustic phonetics, laboratory phonology, and parametric speech coding. The toolkit interprets the phonetic posterior probabilities as a sequential scheme, whereas the phonological posterior-class probabilities are considered as a parallel via *K* different phonological classes. A case study is presented on a LibriSpeech database and a LibriVox US English native female speaker. The phonetic and phonological vocoding yield comparable performance, improving speech quality by merging the phonetic and phonological speech representation.

Vowels and Diphthongs in the Taiyuan Jin Chinese Dialect

Liping Xia, Fang Hu; Chinese Academy of Social Sciences, China

Fri-P-3-1-4, Time: 17:00

On the basis of an acoustic phonetic analysis of monophthongs and diphthongs, this paper describes vowel phonology in the Taiyuan Jin dialect. The results show that Taiyuan has a comparable but different vowel inventory for C(G)V versus C(G)VN syllables. And the vowel contrast is dramatically reduced in checked syllables. The asymmetry between falling and rising diphthongs suggests a dynamic account of vowels, rather than a sequential taxonomy of vowels into monophthongs and diphthongs. Phonetically, monophthongs are composed of a static spectral target, falling diphthongs are composed of a dynamic spectral target, and rising diphthongs are sequences of two spectral targets. Phonologically, falling diphthongs are grouped with monophthongs, rather than rising diphthongs.

The Effects of Prosody on French V-to-V Coarticulation: A Corpus-Based Study

Giuseppina Turco, Cécile Fougeron, Nicolas Audibert; LPP (UMR 7018), France

Fri-P-3-1-5, Time: 17:00

This study examines whether the degree of vowel-to-vowel coarticulation in French (better known as "vowel height harmony", V2-to-V1 henceforth) varies as a function of position in prosodic domain (i.e. IP initial vs. word-medial) and duration of V1 (i.e. short vs. long). Following the literature on the phonetics-prosody interface, segments at stronger edges are more resistant to coarticulatory effects induced by their neighboring vowel. While previous studies have mainly looked at non-local V-to-V coarticulation across prosodic boundaries/domains (e.g.,V#(C)V), here we look at V2-to-V1 coarticulation within an Intonational Phrase according to whether target V1 is in absolute initial position (#V1C(C)V2, e.g., #essaie [esɛ]/[ɛsɛ] – 'try') or not (word-medial, e.g., épaissit [epɛsi]/[epesi] – 'thikened'). The analyses are based on 33k words presenting possible V1C(C)V2 harmonic contexts, which were extracted from a corpus of French running speech. V2-to-V1 coarticulation is measured as the lowering of the first formant of the target V1 (/e, ɛ, o, ɔ/) in relation to the height of the V2 trigger /+high/ (i.e. mid-high and high) vs. /-high/ (i.e. mid-low and low). Results show an effect of prosodic position (but no effect of V1 duration) on V2-to-V1 coarticulation: V1 is more resistant to coarticulation when initial in an IP.

An Acoustic Analysis of /r/ in Tyrolean

Vincenzo Galatà, Lorenzo Spreafico, Alessandro Vietti, Constantijn Kaland; Libera Università di Bolzano, Italy

Fri-P-3-1-6, Time: 17:00

This paper offers a preliminary contribution to the phonetic description and acoustic characterization of /r/ allophony in Tyrolean dialect, an under-researched South Bavarian Dialect spoken in the North of Italy. The analysis of target words containing /r/ in different phonotactic contexts, produced by six Tyrolean female speakers, confirms the high degree of intra-speaker variation in the production of /r/ with a uvular place of articulation. The distributional analysis of the allophones in our sample shows a preference among all the speakers for a fricative manner of articulation followed by approximants and taps and, to a lesser extent, by trills (with a very small amount of vocalized variants). These results are in line with previous research in the South-Tyrolean community. Due to the high variability of rhotic sounds, we further investigate and report on some of their shared acoustic features such as duration across the different phonotactic contexts and Harmonics-to-Noise Ratio for the different allophones attested.

Hyperarticulated Production of Korean Glides by Age Group

Seung-Eun Chang, Minsook Kim; University of California at Berkeley, USA

Fri-P-3-1-7, Time: 17:00

This research uses the hyperspace effect (Johnson, Flemming, & Wright, 1993; Lindblom, 1990) of Korean glides to address the issues triggered by the diachronic sound change of some Korean vowels. Specifically, we examine whether there is any difference between Korean 'wae [wɛ]' versus 'oe [we]' by speech style (casual and clear speech) and speakers' age. Twenty adults from Seoul and the Kyunggi area participated: (i) a younger group (21–34 years old) and (ii) an older group (44–71 years old). The first and second formant frequencies (Hz) were measured at two time points: (i) onset of test syllable and (ii) vowel midpoint. The results showed that the transitional trait of glides "wae [wɛ]" and "oe [we]" at initial timing of syllable was more enhanced in clear speech than in casual speech, as predicted. However, no phonetic evidence was found for the difference between "wae [wɛ]" and "oe [we]" in terms of F1 and F2, even in clear speech. Also, no systematic difference of age group depending on vowel type was found. Therefore, we argue that the

NOTES

diachronic sound merge between "wae [wɛ]" and "oe [we]" is now completed even in the Seoul area and for older groups.

Coda Stop and Taiwan Min Checked Tone Sound Changes

Ho-hsien Pan, Hsiao-tung Huang, Shao-ren Lyu; National Chiao Tung University, Taiwan

Fri-P-3-1-8, Time: 17:00

This acoustical and Electroglottography (EGG) study investigates the effect of coda deletion and co-articulatory phasing on vowels and final coda stops, [p t k ʔ], in Taiwan Min checked tones 3 and 5 syllables. Vowel duration, f0, spectral tilt (H1*-A3*), cepstral peak prominence (CPP) and glottal contact quotient (CQ_H) were analyzed. Compensatory lengthening, f0 lowering and increasing periodic phonation during the production of vowels after coda deletion were observed. During gradual phasing when codas were produced as energy damping, the vowels were found to be shorter in duration and less periodic in voicing than vowels abruptly phased with coda that were produced as full stop closure. However, spectral tilt H1*-A3* was not affected by either coda deletion or co-articulatory phasing. Therefore, these findings suggest that H1*-A3* may play a salient role in checked tone identification, and, as a result, is unaffected by sound change.

Fri-P-3-2 : Prosody, Phonation and Voice Quality

Pacific Concourse – Poster B, 17:00–19:00, Friday, 9 Sept. 2016
Chair: Irene Vogel

The Influence of Modality and Speaking Style on the Assimilation Type and Categorization Consistency of Non-Native Speech

Sarah E. Fenwick, Catherine T. Best, Chris Davis, Michael D. Tyler; Western Sydney University, Australia

Fri-P-3-2-1, Time: 17:00

The Perceptual Assimilation Model [1] proposes that non-native contrast discrimination accuracy can be predicted by perceptual assimilation type. However, assimilation types have been based just on auditory-only (AO) citation speech. Since auditory-visual (AV) and clear speech can benefit non-native speech perception [2, 3], we reasoned that modality and speaking style could influence assimilation. This was tested by presenting English monolinguals Sindhi consonants in a categorization task. Results showed that, across speaking styles, consonants were assimilated the same way in AV and AO. For consonants that were uncategorized in visual-only (VO) conditions: 1) their AO counterpart was more consistently categorized than AV; and 2) citation speech was also more consistently categorized than clear. Interestingly, this set of results was reversed for consonants that were assimilated to the same native category across modalities; participants were able to use the visual articulatory information to make more consistent categorization judgments for AV than AO. This was also the case for speaking style: clear speech was more consistently categorized than citation. Together these results demonstrate that the extent to which AV and clear speech is beneficial for cross-language perception may depend on the similarities between the articulatory characteristics of native and non-native consonants.

Prosodic Convergence with Spoken Stimuli in Laboratory Data

Margaret Zellers; Universität Stuttgart, Germany

Fri-P-3-2-2, Time: 17:00

Accommodation or convergence between speakers has been shown to occur on a variety of levels of linguistic structure. Phonetic convergence appears to be a very variable phenomenon in conversation, with social roles strongly influencing who accommodates to whom. Since phonetic convergence appears to be strongly under speaker control, it is unclear whether speakers might converge phonetically in a laboratory setting. The current study investigates accommodation of pitch and duration features in data collected in a laboratory setting. While speakers in the study did not converge to spoken stimuli in terms of duration features, they did converge to an extent on pitch features. However, only some information-structure contexts led to convergence, suggesting that even in a laboratory setting, speakers are aware of the discourse implications of their production.

Effects of Stress on Fricatives: Evidence from Standard Modern Greek

Charalambos Themistocleous [1], Angelandria Savva [2], Andrie Aristodemou [2]; [1] University of Gothenburg, Sweden; [2] University of Cyprus, Cyprus

Fri-P-3-2-3, Time: 17:00

This study investigates the effects of stress on the spectral properties of fricative noise in Standard Modern Greek (SMG). Twenty female speakers of SMG participated in the study. Fricatives were produced in stressed and unstressed positions in two vowel place positions: back and front vowels. Acoustic measurements were taken and the temporal and spectral properties of fricatives — using spectral moments — were calculated. Stressed fricatives are produced with increased duration, center of gravity, standard deviation, and normalized intensity. The machine learning and classification algorithm C5.0 has been employed to estimate the contribution of the temporal and spectral parameters for the classification of fricatives. Overall, duration and center of gravity contribute the most to the classification of stressed vs. unstressed fricatives.

Analysis of Chinese Syllable Durations in Running Speech of Japanese L2 Learners

Yue Sun [1], Shudon Hsiao [1], Yoshinori Sagisaka [1], Jinsong Zhang [2]; [1] Waseda University, Japan; [2] BLCU, China

Fri-P-3-2-4, Time: 17:00

Aiming at better understanding of prosody generation by native Japanese learners of Mandarin as a second language (L2), we analyzed the syllable duration differences between tone types. By comparing the mean syllable durations and the variation of normalized syllable durations across tone types and speakers, significant differences were found between tone types as well as between speakers. Native Chinese speakers generate tone 1 and tone 2 with relatively long durations but smaller variations, contrary to tone 3 and tone 4. Japanese L2 learners generate tone 3 with relatively high variations compared to the other tones, while the mean duration of tone 4 was remarkably different from natives. Compared with native speakers, the variations of both tone 3 and tone 4 are significantly smaller. Furthermore, the neutral tone caused a significant increase of the mean variation across tones for the Japanese L2 learners.

NOTES

The results suggest that native Chinese speakers control syllable durations adaptively with tones, especially for tone 3 and tone 4, in running speech while Japanese L2 learners tend to pronounce them in isolated syllable fashion.

Automatic Paragraph Segmentation with Lexical and Prosodic Features

Catherine Lai[1], *Mireia Farrús*[2], *Johanna D. Moore*[1]; [1]*University of Edinburgh, UK;* [2]*Universitat Pompeu Fabra, Spain*

Fri-P-3-2-5, Time: 17:00

As long-form spoken documents become more ubiquitous in everyday life, so does the need for automatic discourse segmentation in spoken language processing tasks. Although previous work has focused on broad topic segmentation, detection of finer-grained discourse units, such as paragraphs, is highly desirable for presenting and analyzing spoken content. To better understand how different aspects of speech cue these subtle discourse transitions, we investigate automatic paragraph segmentation of TED talks. We build lexical and prosodic paragraph segmenters using Support Vector Machines, AdaBoost, and Long Short Term Memory (LSTM) recurrent neural networks. In general, we find that induced cue words and supra-sentential prosodic features outperform features based on topical coherence, syntactic form and complexity. However, our best performance is achieved by combining a wide range of individually weak lexical and prosodic features, with the sequence modelling LSTM generally outperforming the other classifiers by a large margin. Moreover, we find that models that allow lower level interactions between different feature types produce better results than treating lexical and prosodic contributions as separate, independent information sources.

Automatic Glottal Inverse Filtering with Non-Negative Matrix Factorization

Manu Airaksinen[1], *Lauri Juvela*[1], *Tom Bäckström*[2], *Paavo Alku*[1]; [1]*Aalto University, Finland;* [2]*FAU Erlangen-Nürnberg, Germany*

Fri-P-3-2-6, Time: 17:00

This study presents an automatic glottal inverse filtering (GIF) technique based on separating the effect of the glottal main excitation from the impulse response of the vocal tract. The proposed method is based on a non-negative matrix factorization (NMF) based decomposition of an ultra short-term spectrogram of the analyzed signal. Unlike other state-of-the-art GIF techniques, the proposed method does not require estimation of glottal closure instants.

The proposed method was objectively evaluated with two test sets of continuous synthetic speech created with a glottal vocoding analysis/synthesis procedure. When compared to a set of reference GIF methods, the proposed NMF technique shows improved estimation accuracy especially for male voices.

Speaker Identity and Voice Quality: Modeling Human Responses and Automatic Speaker Recognition

Soo Jin Park[1], *Caroline Sigouin*[2], *Jody Kreiman*[1], *Patricia Keating*[1], *Jinxi Guo*[1], *Gary Yeung*[1], *Fang-Yu Kuo*[1], *Abeer Alwan*[1]; [1]*University of California at Los Angeles, USA;* [2]*Université Laval, Canada*

Fri-P-3-2-7, Time: 17:00

Despite recent breakthroughs in automatic speaker recognition (ASpR), system performance still degrades when utterances are short and/or when within-speaker variability is large. This study used short test utterances (2–3sec) to investigate the effect of within-speaker variability on state-of-the-art ASpR system performance. A subset of a newly-developed UCLA database is used, which contains multiple speech tasks per speaker. The short utterances combined with a speaking-style mismatch between read sentences and spontaneous affective speech degraded system performance, for 25 female speakers, by 36%. Because humans are more robust to utterance length or within-speaker variability, understanding human perception might benefit ASpR systems. Perception experiments were conducted with recorded read sentences from 3 female speakers, and a model is proposed to predict the perceptual dissimilarity between tokens. Results showed that a set of voice quality features including F0, F1, F2, F3, H1*-H2*, H2*-H4*, H4*-H2k*, H2k*-H5k, and CPP provides information that complements MFCCs. By fusing the feature set with MFCCs, human response prediction RMS error was .12, which represents a 12% relative error reduction compared to using MFCCs alone. In ASpR experiments with short utterances from 50 speakers, the voice quality feature set decreased the error rate by 11% when fused with MFCCs.

Analysis of Glottal Stop in Assam Sora Language

Sishir Kalita, Luke Horo, Priyankoo Sarmah, S.R. Mahadeva Prasanna, S. Dandapat; IIT Guwahati, India

Fri-P-3-2-8, Time: 17:00

The objective of this work is to characterize the intervocalic glottal stops in Assam Sora. Assam Sora is a low resource language of the South Munda language family. Glottal stops are produced with gestures in the deep laryngeal level; hence, the estimated excitation source signal is used in this study to characterize the source dynamics during the production of Assam Sora glottal stops. From that, temporal domain voice source features, Quasi-Open Quotient (QOQ) and Normalized Amplitude Quotient (NAQ) are extracted along with spectral features such as H1-H2 ratio and Harmonic Richness Factor (HRF). One excitation source feature is extracted from the zero frequency filtered version of the speech signal to characterize the variations within the glottal cycles in glottal stop region. A recently proposed wavelet based voice source feature, Maxima Dispersion Quotient (MDQ) is also used to characterize the abrupt glottal closure during glottal stop production. From the analysis, it is observed that the features are salient enough to uniquely characterize glottal stops from the adjacent vowel sounds and may also be used in continuous speech. A Mann-Whitney U test confirmed the statistical significance of the differences between glottal stops and their adjacent vowels.

NOTES

Acoustic Differences Between English /t/ Glottalization and Phrasal Creak

Marc Garellek, Scott Seyfarth; University of California at San Diego, USA

Fri-P-3-2-9, Time: 17:00

In American English, the presence of creaky voice can derive from distinct linguistic processes, including phrasal creak (prolonged irregular voicing, often at edges of prosodic phrases) and coda /t/ glottalization (when the alveolar closure for syllable-final /t/ is replaced by or produced simultaneously with glottal constriction). Previous work has shown that listeners can differentiate words in phrasal creak from those with /t/ glottalization, which suggests that there are acoustic differences between the creaky voice derived from phrasal creak and /t/ glottalization. In this study, we analyzed vowels preceding syllable-final /t/ in the Buckeye Corpus, which includes audio recordings of spontaneous speech from 40 speakers of American English. Tokens were coded for presence of phrasal creak (prolonged irregular voicing extending beyond the target syllable) and /t/ glottalization (whether the /t/ was produced only with glottal constriction). Eleven spectral measures of voice quality, including both harmonic and noise measures, were extracted automatically and discriminant analyses were performed. The results indicate that the discriminant functions can classify these sources of creaky voice above chance, and that Cepstral Peak Prominence, a measure of harmonics-to-noise ratio, is important for distinguishing phrasal creak from glottalization.

The Acoustics of Lexical Stress in Italian as a Function of Stress Level and Speaking Style

Anders Eriksson[1], Pier Marco Bertinetto[2], Mattias Heldner[1], Rosalba Nodari[2], Giovanna Lenoci[2]; [1]Stockholm University, Sweden; [2]Scuola Normale Superiore, Italy

Fri-P-3-2-10, Time: 17:00

The study is part of a series of studies, describing the acoustics of lexical stress in a way that should be applicable to any language. The present database of recordings includes Brazilian Portuguese, English, Estonian, German, French, Italian and Swedish. The acoustic parameters examined are *F_0-level, F_0-variation, Duration,* and *Spectral Emphasis.* Values for these parameters, computed for all vowels (a little over 24000 vowels for Italian), are the data upon which the analyses are based. All parameters are examined with respect to their correlation with *Stress* (primary, secondary, unstressed) and speaking *Style* (wordlist reading, phrase reading, spontaneous speech) and *Sex* of the speaker (female, male). For Italian *Duration* was found to be the dominant factor by a wide margin, in agreement with previous studies. *Spectral Emphasis* was the second most important factor. *Spectral Emphasis* has not been studied previously for Italian but intensity, a related parameter, has been shown to correlate with stress. *F_0-level* was also significantly correlated but not to the same degree. Speaker *Sex* turned out as significant in many comparisons. The differences were, however, mainly a function of the degree to which a given parameter was used, not how it was used to signal lexical stress contrasts.

Cross-Gender and Cross-Dialect Tone Recognition for Vietnamese

Antje Schweitzer, Ngoc Thang Vu; Universität Stuttgart, Germany

Fri-P-3-2-11, Time: 17:00

We investigate tone recognition in Vietnamese across gender and dialects. In addition to well-known parameters such as single fundamental frequency (F0) values and energy features, we explore the impact of harmonicity on recognition accuracy, as well as that of the PaIntE parameters, which quantify the shape of the F0 contour over complete syllables instead of providing more local single values. Using these new features for tone recognition in the GlobalPhone database, we observe significant improvements of approx. 1% in recognition accuracy when adding harmonicity, and of another approx. 4% when adding the PaIntE parameters. Furthermore, we analyze the influence of gender and dialect on recognition accuracy. The results show that it is easier to recognize tones for female than for male speakers, and easier for the Northern dialect than for the Southern dialect. Moreover, we achieve reasonable results testing models across gender, while the performance drops strongly when testing across dialects.

Prosody Modification Using Allpass Residual of Speech Signals

Karthika Vijayan, K. Sri Rama Murty; IIT Hyderabad, India

Fri-P-3-2-12, Time: 17:00

In this paper, we attempt to signify the role of phase spectrum of speech signals in acquiring an accurate estimate of excitation source for prosody modification. The phase spectrum is parametrically modeled as the response of an allpass (AP) filter, and the filter coefficients are estimated by considering the linear prediction (LP) residual as the output of the AP filter. The resultant residual signal, namely AP residual, exhibits unambiguous peaks corresponding to epochs, which are chosen as pitch markers for prosody modification. This strategy efficiently removes ambiguities associated with pitch marking, required for pitch synchronous overlap-add (PSOLA) method. The prosody modification using AP residual is advantageous than time domain PSOLA (TD-PSOLA) using speech signals, as it offers fewer distortions due to its flat magnitude spectrum. Windowing centered around unambiguous peaks in AP residual is used for segmentation, followed by pitch/duration modification of AP residual by mapping of pitch markers. The modified speech signal is obtained from modified AP residual using synthesis filters. The mean opinion scores are used for performance evaluation of the proposed method, and it is observed that the AP residual-based method delivers equivalent performance as that of LP residual-based method using epochs, and better performance than the linear prediction PSOLA (LP-PSOLA).

Analyzing the Contribution of Top-Down Lexical and Bottom-Up Acoustic Cues in the Detection of Sentence Prominence

Sofoklis Kakouros[1], Joris Pelemans[2], Lyan Verwimp[2], Patrick Wambacq[2], Okko Räsänen[1]; [1]Aalto University, Finland; [2]Katholieke Universiteit Leuven, Belgium

Fri-P-3-2-13, Time: 17:00

Recent work has suggested that prominence perception could be

NOTES

driven by the predictability of the acoustic prosodic features of speech. On the other hand, lexical predictability and part of speech information are also known to correlate with prominence. In this paper, we investigate how the bottom-up acoustic and top-down lexical cues contribute to sentence prominence by using both types of features in unsupervised and supervised systems for automatic prominence detection. The study is conducted using a corpus of Dutch continuous speech with manually annotated prominence labels. Our results show that unpredictability of speech patterns is a consistent and important cue for prominence at both the lexical and acoustic levels, and also that lexical predictability and part-of-speech information can be used as efficient features in supervised prominence classifiers.

A Longitudinal Study of Children's Intonation in Narrative Speech

Jeffrey Kallay, Melissa A. Redford; University of Oregon, USA

Fri-P-3-2-14, Time: 17:00

Adults' narratives are hierarchically structured. This structure is evident in the linguistic and prosodic domains. Children's narratives have a flatter structure. This structure is evident in the linguistic domain, but less is known about the prosodic domain. Here, we report results from a longitudinal study of children's narratives that enhance our understanding of the development of discourse prosody. Spontaneous narratives were obtained from 60 children (aged 5 to 7) over a 3-year period. F0 was tracked to obtain absolute measures of slope steepness and linearity for every utterance of each narrative. These measures are known correlates of syntactic and semantic complexity. Slope direction and inter-utterance continuity in F0 were also calculated. These measures are known correlates of event boundaries in adult discourse. The results indicated systematic developmental changes related to age and year for all measures except slope steepness, consistent with developmental increases in linguistic complexity and the production of more adult-like narratives. The evidence also indicates that developmental change is most pronounced between the ages of 5 and 7 years, and levels out afterwards.

Fri-P-3-3 : Speech Production Analysis and Modeling

Pacific Concourse – Poster C, 17:00–19:00, Friday, 9 Sept. 2016
Chair: Gayeon Son

Velum Control for Oral Sounds

Reed Blaylock, Louis Goldstein, Shrikanth S. Narayanan; University of Southern California, USA

Fri-P-3-3-1, Time: 17:00

Velum position during speech shows systematic variability within and across speakers, but has a binary phonological contrast (nasal and oral). Velum lowering is often thought to constitute an independent phonological unit, partly because of its robust prosodically-conditioned timing during nasal stops. Velum raising, on the other hand, is usually considered to be a non-phonological consequence of other vocal tract movements. Moreover, velum raising has almost always been observed in the context of nasals, and has rarely been studied in purely oral contexts. This experiment directly contrasts velum movement in oral and nasal contexts. The results show that temporal coordination of velum raising during oral stops resembles the temporal coordination of velum lowering during nasals, suggesting that velum position and movement are controlled for both raising and lowering. The results imply that some revisions to the Articulatory Phonology model may be appropriate, specifically with regards to the treatment of velum raising as an independent phonological unit.

F0 Development in Acquiring Korean Stop Distinction

Gayeon Son; University of Pennsylvania, USA

Fri-P-3-3-2, Time: 17:00

A number of studies have investigated the role of Voice Onset Time (VOT) on acquisition of stop voicing contrast. Korean stop contrasts (lenis, fortis, and aspirated), however, cannot be differentiated only by VOT since they are all pulmonic egressive voiceless stops. For this three-way distinction, another acoustic parameter, fundamental frequency (F0), critically operates. The present study explores how F0 is perceptually acquired and phonetically operates for Korean stop contrast over age. In order to reveal the relationship between F0 developmental patterns and age, a quantitative acoustic model dealt with word-initial stop productions by 58 Korean young children aged 20 months to 47 months. The results showed that phonetic accuracy depends on the perceptual thresholds in F0, and the significant phonetic differentiation with F0 between lenis and aspirated stops was significantly related to age. These findings suggest that acquisition of F0 plays a crucial role in the formation of phonemic categories for lenis and aspirated stops and this process significantly affects articulatory distinction.

Phonetic Reduction Can Lead to Lengthening, and Enhancement Can Lead to Shortening

Clara Cohen, Matt Carlson; Pennsylvania State University, USA

Fri-P-3-3-3, Time: 17:00

Contextually probable, high-frequency, or easily accessible words tend to be phonetically reduced, a pattern usually attributed to faster lexical access. In principle, word forms that are frequent in their inflectional paradigms should also enjoy faster lexical access, leading again to phonetic reduction. Yet research has found evidence of both reduction and enhancement on paradigmatically probable inflectional affixes. The current corpus study uses pronunciation data from conversationally produced English verbs and nouns to test the predictions of two accounts. In an exemplar account, paradigmatically probable forms seem enhanced because their denser exemplar clouds resist influence from related word forms on the average production target. A second pressure reduces such forms because they are, after all, more easily accessed. Under this account, paradigmatically probable forms should have longer affixes but shorter stems. An alternative account proposes that paradigmatically probable forms are produced in such a way as to enhance not articulation, but *contrasts* between related word forms. This account predicts lengthening of suffixed forms, and shortening of unsuffixed forms.

The results of the corpus study support the second account, suggesting that characterizing pronunciation variation in terms of phonetic reduction and enhancement oversimplifies the relationship between lexical storage, retrieval, and articulation.

Mechanical Production of [b], [m] and [w] Using Controlled Labial and Velopharyngeal Gestures

Takayuki Arai; Sophia University, Japan

Fri-P-3-3-4, Time: 17:00

As an extension of a series of models we have developed, a mechanical bent vocal-tract model with nasal cavity was proposed for educational and clinical applications, as well as for understanding human speech production. Although our recent studies have focused on flap and approximant sounds, this paper introduced a new model for the consonants [b], [m] and [w]. Because the articulatory gesture of approximants is slow compared to the more rapid movement of plosives, in our [b] and [m] model, the elastic force of a spring is applied to affect the movement of the lower lip block, as was done for flap sounds in our previous studies. The main difference between [b] and [m] is in the velopharyngeal port, which is closed for [b] and open for [m]. In this study, we concluded that 1) a slower manipulation of the lip block is needed for [w], while 2) [b] and [m] require a faster movement, and finally, 3) close-open coordination of the lip and velopharyngeal gestures is important for [m].

An Improved 3D Geometric Tongue Model

Qiang Fang[1], Yun Chen[2], Haibo Wang[1], Jianguo Wei[2], Jianrong Wang[2], Xiyu Wu[3], Aijun Li[1]; [1]Chinese Academy of Social Sciences, China; [2]Tianjin University, China; [3]Peking University, China

Fri-P-3-3-5, Time: 17:00

This study describes an improved geometric articulatory model based on MRI and CBCT (Cone Beam Computer Tomography) data. The basic idea is to improve the coherence of the vertices of tongue meshes so as to obtain more accurate tongue model. This is conducted in two aspects: i) The representative vertices of tongue surface are depicted in Cartesian coordinate system rather than in a semi-polar gridline coordinate system. ii) tongue surface meshes are modeled with reference to anatomical landmarks. Then, guided PCA is used to extract the control components based on MRI data. The average reconstruction error is less than 1.0 mm. Both qualitative and quantitative evaluation indicates that the proposed method surpasses the conventional semi-polar gridline system based method.

Congruency Effect Between Articulation and Grasping in Native English Speakers

Mikko Tiainen[1], Fatima M. Felisberti[2], Kaisa Tiippana[1], Martti Vainio[1], Juraj Simko[1], Jiri Lukavsky[3], Lari Vainio[1]; [1]University of Helsinki, Finland; [2]Kingston University London, UK; [3]Czech Academy of Sciences, Czech Republic

Fri-P-3-3-6, Time: 17:00

Previous studies have shown congruency effects between specific speech articulations and manual grasping actions. For example, uttering the syllable [kɑ] facilitates power grip responses in terms of reaction time and response accuracy. A similar association of the syllable [ti] with precision grip has also been observed. As these congruency effects have been to date shown only for Finnish native speakers, this study explored whether the congruency effects generalize to native speakers of another language. The original experiments were therefore replicated with English participants (N=16). Several previous findings were reproduced, namely the association of syllables [kɑ] and [ke] with power grip and of [ti] and [te] with precision grip. However, the association of vowels [ɑ] and [i] with power and precision grip, respectively, previously found for Finnish participants, was not significant for English speakers. This difference could be related to ambiguities of English orthography and pronunciation variations. It is possible that for English speakers seeing a certain written vowel activates several different phonological representations associated with that letter. If the congruency effects are based on interactions between specific phonological representations and grasp actions, this ambiguity might lead to weakening of the effects in the manner demonstrated here.

Emergence of Vocal Developmental Sequences in a Predictive Coding Model of Speech Acquisition

Shamima Najnin, Bonny Banerjee; University of Memphis, USA

Fri-P-3-3-7, Time: 17:00

Learning temporal patterns among primitive speech sequences and being able to control the motor apparatus for effective production of the learned patterns are imperative for speech acquisition in infants. In this paper, we develop a predictive coding model whose objective is to minimize the sensory (auditory) and proprioceptive prediction errors. Temporal patterns are learned by minimizing the former while control is learned by minimizing the latter. The model is learned using a set of synthetically generated syllables, as in other contemporary models. We show that the proposed model outperforms existing ones in learning vocalization classes. It also computes the control/muscle activation which is useful for determining the degree of easiness of vocalization.

Categorization of Natural Spanish Whistled Vowels by Naïve Spanish Listeners

Julien Meyer[1], Laure Dentel[2], Fanny Meunier[3]; [1]GIPSA, France; [2]World Whistles Research Association, France; [3]L2C2, France

Fri-P-3-3-8, Time: 17:00

Whistled speech in a non tonal language consists of the natural emulation of vocalic and consonantal qualities in a simple modulated whistled signal. This special speech register represents a natural telecommunication system that enables high levels of sentence intelligibility by trained speakers. It is not directly intelligible to naïve listeners. Yet, it is easily learned by speakers of the language that is being whistled, as attested by current efforts of revitalization of whistled Spanish in the Canary Islands. To understand better the relation between whistled and spoken speech perception, we looked here at how Spanish native speakers knowing nothing about whistled speech categorized four Spanish whistled vowels. The results show that naïve participants were able to categorize these vowels, although not as accurately as a native whistler.

Between- and Within-Speaker Effects of Bilingualism on F0 Variation

Rob Voigt, Dan Jurafsky, Meghan Sumner; Stanford University, USA

Fri-P-3-3-9, Time: 17:00

To what extent is prosody shaped by cultural and social factors? Existing research has shown that an individual bilingual speaker exhibits differences in framing, ideology, and personality when speaking their two languages. To understand whether these differences extend to prosody we study F0 variation in a corpus of interviews with German-Italian and German-French bilingual speakers. We find two primary effects. First, a *between-speaker* effect: these two groups of bilinguals make different use of F0 even when they are all speaking German. Second, a *within-speaker* effect: bilinguals use F0 differently depending on which language they are speaking, differences that are consistent across speakers. These effects are modulated strongly by gender, suggesting that language-specific social positioning may play a central role. These results have important implications for our understanding of bilingualism and cross-cultural linguistic difference in general. Prosody appears to be a moving target rather than a stable feature, as speakers use prosodic variation to position themselves on cultural and social axes like linguistic context and gender.

Vowel Characteristics in the Assessment of L2 English Pronunciation

Calbert Graham, Paula Buttery, Francis Nolan; University of Cambridge, UK

Fri-P-3-3-10, Time: 17:00

There is considerable need to utilise linguistically meaningful measures of second language (L2) proficiency that are based on perceptual cues used by humans to assess pronunciation. Previous research on non-native acquisition of vowel systems suggests a strong link between vowel production accuracy and speech intelligibility. It is well known that the acoustic and perceptual identification of vowels rely on formant frequencies. However, formant analysis may not be viable in large-scale corpus research, given the need for manual correction of tracking errors. Spectral analysis techniques have been shown to be a robust alternative to formant tracking. This paper explores the use of one such technique — the discrete cosine transform (DCT) — for modelling English vowel spectra in the productions of non-native English speakers. Mel-scaled DCT coefficients were calculated over a frequency band of 200–4000 Hz. Results show a statistically significant correlation between coefficients and the proficiency level of speakers, and suggest that this technique holds some promise in automated L2 pronunciation teaching and assessment.

Kulning (Swedish Cattle Calls): Acoustic, EGG, Stroboscopic and High-Speed Video Analyses of an Unusual Singing Style

Ahmed Geneid[1], Anne-Maria Laukkanen[2], Anita McAllister[3], Robert Eklund[4]; [1]Helsinki University Hospital, Finland; [2]University of Tampere, Finland; [3]Karolinska Institute, Sweden; [4]Linköping University, Sweden

Fri-P-3-3-11, Time: 17:00

The Swedish cattle call singing style ‘kulning’ is surprisingly understudied, despite its mythical status in folklore. While some acoustic and physiological aspects have been addressed previously [1,2], a more detailed analysis is still lacking. Previous work [2] showed that sound pressure level (SPL) in kulning tapered off less than in head register as a function of distance, which warrants a study of underlying physiological mechanisms responsible for this. In the present paper, the same singer, singing the same song — in kulning and in head register (“falsetto”) mode — was recorded indoors. Electroglottographic (EGG), stroboscopic, high-speed endoscopic and audio registrations were made. Analyses examined differences between kulning and head register. Results show somewhat higher SPL in kulning than in head register confirming the previous findings. EGG showed longer relative glottal closed time and higher amplitude of the signal in kulning. This suggests better vocal fold contact in kulning. Flexible nasofiberoscopy and high-speed recordings during kulning showed medial and antero-posterior narrowing of the laryngeal inlet, a clear approximation of the false vocal folds and marked adduction of the vocal folds.

Glottal Squeaks in VC Sequences

Míša Hejná[1], Pertti Palo[2], Scott Moisik[3]; [1]Newcastle University, UK; [2]Queen Margaret University, UK; [3]NTU, Singapore

Fri-P-3-3-12, Time: 17:00

This paper reports results related to the phenomenon referred to as a “glottal squeak” (coined by [1]). At present, nothing is known about the conditioning and the articulation of this feature of speech. Our qualitative acoustic analyses of the conditioning of squeaks (their frequency of occurrence, duration, and f_0) found in Aberystwyth English and Manchester English suggest that squeaking may be a result of intrinsically tense vocal fold state associated with thyroarytenoid (TA) muscle recruitment [2] required for epilaryngeal constriction and vocal-ventricular fold contact (VVFC) needed to produce glottalisation [3]. In this interpretation, we hypothesise that squeaks occasionally occur during constriction disengagement: at the point when VVFC suddenly releases but the TAs have not yet fully relaxed. Extralinguistic conditioning identified in this study corroborates findings reported by [1]: females are more prone to squeaking and the phenomenon is individual-dependent.

Automatic Pronunciation Generation by Utilizing a Semi-Supervised Deep Neural Networks

Naoya Takahashi[1], Tofigh Naghibi[2], Beat Pfister[2]; [1]Sony, Japan; [2]ETH Zürich, Switzerland

Fri-P-3-3-13, Time: 17:00

Phonemic or phonetic sub-word units are the most commonly used atomic elements to represent speech signals in modern ASRs. However they are not the optimal choice due to several reasons such as: large amount of effort required to handcraft a pronunciation dictionary, pronunciation variations, human mistakes and under-resourced dialects and languages. Here, we propose a data-driven pronunciation estimation and acoustic modeling method which only takes the orthographic transcription to jointly estimate a set of sub-word units and a reliable dictionary. Experimental results show that the proposed method which is based on semi-supervised training of a deep neural network largely outperforms phoneme based continuous speech recognition on the TIMIT dataset.

Notes

Fri-P-3-4 : Spoken Dialogue Systems

Pacific Concourse – Poster D, 17:00–19:00, Friday, 9 Sept. 2016
Chair: Ruhi Sarikaya

Personalized Natural Language Understanding

Xiaohu Liu, Ruhi Sarikaya, Liang Zhao, Yong Ni, Yi-Cheng Pan; Microsoft, USA

Fri-P-3-4-1, Time: 17:00

Natural language understanding (NLU) is one of the critical components of dialog systems. Its aim is to extract semantic meaning from typed text input or the spoken text coming out of the speech recognizer. Traditionally, NLU systems are built in a user-independent fashion, where the system behavior does not adapt to the user. However, personal information can be very useful for language understanding tasks, if it is made available to the system. With personal digital assistant (PDA) systems, many forms of personal data are readily available for the NLU systems to make the models and the system more personal. In this paper, we propose a method to personalize language understanding models by making use of the personal data with privacy respected and protected. We report experiments on two domains for intent classification and slot tagging, where we achieve significant accuracy improvements compared to the baseline models that are trained in a user independent manner.

A Sequence-to-Sequence Model for User Simulation in Spoken Dialogue Systems

Layla El Asri, Jing He, Kaheer Suleman; Maluuba Research, Canada

Fri-P-3-4-2, Time: 17:00

User simulation is essential for generating enough data to train a statistical spoken dialogue system. Previous models for user simulation suffer from several drawbacks, such as the inability to take dialogue history into account, the need of rigid structure to ensure coherent user behaviour, heavy dependence on a specific domain, the inability to output several user intentions during one dialogue turn, or the requirement of a summarized action space for tractability. This paper introduces a data-driven user simulator based on an encoder-decoder recurrent neural network. The model takes as input a sequence of dialogue contexts and outputs a sequence of dialogue acts corresponding to user intentions. The dialogue contexts include information about the machine acts and the status of the user goal. We show on the Dialogue State Tracking Challenge 2 (DSTC2) dataset that the sequence-to-sequence model outperforms an agenda-based simulator and an n-gram simulator, according to F-score. Furthermore, we show how this model can be used on the original action space and thereby models user behaviour with finer granularity.

Root Cause Analysis of Miscommunication Hotspots in Spoken Dialogue Systems

Spiros Georgiladakis[1], Georgia Athanasopoulou[1], Raveesh Meena[2], José Lopes[2], Arodami Chorianopoulou[3], Elisavet Palogiannidi[3], Elias Iosif[1], Gabriel Skantze[2], Alexandros Potamianos[1]; [1]NTUA, Greece; [2]KTH, Sweden; [3]Technical University of Crete, Greece

Fri-P-3-4-3, Time: 17:00

A major challenge in Spoken Dialogue Systems (SDS) is the detection of problematic communication (hotspots), as well as the classification of these hotspots into different types (root cause analysis). In this work, we focus on two classes of root cause, namely, erroneous speech recognition vs. other (e.g., dialogue strategy). Specifically, we propose an automatic algorithm for detecting hotspots and classifying root causes in two subsequent steps. Regarding hotspot detection, various lexico-semantic features are used for capturing repetition patterns along with affective features. Lexico-semantic and repetition features are also employed for root cause analysis. Both algorithms are evaluated with respect to the Let's Go dataset (bus information system). In terms of classification unweighted average recall, performance of 80% and 70% is achieved for hotspot detection and root cause analysis, respectively.

Making Personal Digital Assistants Aware of What They Do Not Know

Omar Zia Khan, Ruhi Sarikaya; Microsoft, USA

Fri-P-3-4-4, Time: 17:00

Personal digital assistants (PDAs) are spoken (and typed) dialog systems that are expected to assist users without being constrained to a particular domain. Typically, it is possible to construct deep multi-domain dialog systems focused on a narrow set of head domains. For the long tail (or when the speech recognition is not correct) the PDA does not know what to do. Two common fallback approaches are to either acknowledge its limitation or show web search results. Either approach can severely undermine the user's trust in the PDA's intelligence if invoked at the wrong time. In this paper, we propose features that are helpful in predicting the right fallback response. We then use these features to construct dialog policies such that the PDA is able to correctly decide between invoking web search or acknowledging its limitation. We evaluate these dialog policies on real user logs gathered from a PDA, deployed to millions of users, using both offline (judged) and online (user-click) metrics. We demonstrate that our hybrid dialog policy significantly increases the accuracy of choosing the correct response, measured by analyzing click-rate in logs, and also enhances user satisfaction, measured by human evaluations of the replayed experience.

NOTES

Implementing Acoustic-Prosodic Entrainment in a Conversational Avatar

Rivka Levitan[1], Štefan Beňuš[2], Ramiro H. Gálvez[3], Agustín Gravano[3], Florencia Savoretti[3], Marian Trnka[4], Andreas Weise[1], Julia Hirschberg[5]; [1]CUNY Brooklyn College, USA; [2]UKF, Slovak Republic; [3]Universidad de Buenos Aires, Argentina; [4]Slovak Academy of Sciences, Slovak Republic; [5]Columbia University, USA

Fri-P-3-4-5, Time: 17:00

Entrainment, aka accommodation or alignment, is the phenomenon by which conversational partners become more similar to each other in behavior. While there has been much work on some behaviors there has been little on entrainment in speech and even less on how Spoken Dialogue Systems (SDS) which entrain to their users' speech can be created. We present an architecture and algorithm for implementing acoustic-prosodic entrainment in SDS and show that speech produced under this algorithm conforms to the feature targets, satisfying the properties of entrainment behavior observed in human-human conversations. We present results of an extrinsic evaluation of this method, comparing whether subjects are more likely to ask advice from a conversational avatar that entrains vs. one that does not, in English, Spanish and Slovak SDS.

Perceived Usability and Cognitive Demand of Secondary Tasks in Spoken Versus Visual-Manual Automotive Interaction

Annika Silvervarg[1], Sofia Lindvall[1], Jonatan Andersson[1], Ida Esberg[2], Christian Jernberg[2], Filip Frumerie[2], Arne Jönsson[1]; [1]Linköping University, Sweden; [2]Volvo, Sweden

Fri-P-3-4-6, Time: 17:00

We present results from a study of truck drivers' experience of using two different interfaces; spoken interaction and visual-manual interaction, to perform secondary tasks while driving. The instruments used to measure their experience are based on three popular questionnaires, measuring different aspects of usability and cognitive load: SASSI, SUS and DALI. Our results show that the speech interface is preferred both regarding usability and cognitive demand.

Fri-S&T-3 : Show & Tell Session 3

Market Street Foyer, 17:00–19:00, Friday, 9 Sept. 2016
Chairs: Shiva Sundaram, Nicolas Scheffer

Zara: An Empathetic Interactive Virtual Agent

Pascale Fung, Anik Dey, Farhad Bin Siddique, Ruixi Lin, Yang Yang, Wan Yan, Ricky Ho Yin Chan; HKUST, China

Fri-S&T-3-1, Time: 17:00

Zara, or 'Zara the Supergirl', is a virtual robot that can show empathy while interacting with an user, and at the end of a 5–10 minute conversation, it can give a personality analysis based on the user responses. It can display and share emotions with the aid of its built in sentiment analysis, facial and emotion recognition, and speech module. Being the first of its kind, it has successfully integrated an empathetic system along with the human emotion recognition and sharing, into an augmented human-robot interaction system. Zara was also displayed at the World Economic Forum held at Dalian in September 2015.

Measuring Pronunciation Improvement in Users of CAPT Tool TipTopTalk!

Cristian Tejedor-García, David Escudero-Mancebo, Enrique Cámara-Arenas, César González-Ferreras, Valentín Cardeñoso-Payo; Universidad de Valladolid, Spain

Fri-S&T-3-2, Time: 17:00

We present a L2 pronunciation training serious game based on the minimal-pairs technique, incorporating sequences of exposure, discrimination and production, and using text-to-speech and speech recognition systems. We have measured the quality of users' production during a period of time in order to assess improvement after using the application. Substantial improvement is found among users with poorer initial performance levels. The program's gamification resources manage to engage a high percentage of users. A need is felt to include feedback for users in future versions with the purpose of increasing their performance and avoiding the performance drop detected after protracted use of the tool.

SparkNG: Interactive MATLAB Tools for Introduction to Speech Production, Perception and Processing Fundamentals and Application of the Aliasing-Free L-F Model Component

Hideki Kawahara; Wakayama University, Japan

Fri-S&T-3-3, Time: 17:00

This article introduces a set of interactive tools for studying fundamentals of speech production, perception and processing. In addition to this voice production simulator, it consists of interactive time-frequency representation, auditory representation visualizer and a vocal tract shape visualizer for introductory materials. It consists of compiled executables for Windows and Mac environment, which do not require MATLAB license. The MATLAB sources of the tools and their constituent functions are publicly accessible under open source license.

Real-Time Tracking of Speakers' Emotions, States, and Traits on Mobile Platforms

Erik Marchi, Florian Eyben, Gerhard Hagerer, Björn Schuller; audEERING, Germany

Fri-S&T-3-4, Time: 17:00

We demonstrate audEERING's sensAI technology running natively on low-resource mobile devices applied to emotion analytics and speaker characterisation tasks. A show-case application for the Android platform is provided, where audEERING's highly noise robust voice activity detection based on LSTM-RNN is combined with our core emotion recognition and speaker characterisation engine natively on the mobile device. This eliminates the need for network connectivity and allows to perform robust speaker state and trait recognition efficiently in real-time without network transmission lags. Real-time factors are benchmarked for a popular mobile device to demonstrate the efficiency, and average response times are compared to a server based approach. The output of the emotion analysis is visualized graphically in the arousal and valence space alongside the emotion category and further speaker characteristics.

NOTES

Sat-SE-1 : Special Event: Mindfulness

Grand Ballroom ABC, 08:00–08:25, Saturday, 10 Sept. 2016
Chair: David Suendermann-Oeft

Mindfulness Special Event

Nikki Mirghafori; ICSI, USA

Sat-SE-1, Time: 08:00

Mindfulness has entered the cultural mainstream in recent years, with classes and workshops offered on the topic at many universities and companies (including Google, Facebook, etc.). Mindfulness can be thought of as a way to train our mind to be fully present with this moment's experience with curiosity, kindness, and equanimity. The training can serve as a refuge in our busy professional lives and help build resilience. This special event will be in the form of a guided meditation and serve as an introduction for those who are new to this practice, and a chance to practice in community for those who have previous experience. Everyone is welcome.

Keynote 2: Edward Chang

Grand Ballroom ABC, 08:30–09:30, Saturday, 10 Sept. 2016
Chairs: Andreas Stolcke

The Human Speech Cortex

Edward Chang; University of California at San Francisco, USA

Sat-Keynote-2, Time: 08:30

A unique and defining trait of human behavior is our ability to communicate through speech. The fundamental organizational principles of the neural circuits within speech brain areas are largely unknown. In this talk, I will present new results from our research on the functional organization of the human higher-order auditory cortex, known as Wernicke's area. I will focus on how neural populations in the superior temporal lobe encode acoustic-phonetic representations of speech, and also how they integrate influences of linguistic context to achieve perceptual robustness.

Sat-SE-2 : Special Event: Speaker Comparison for Forensic and Investigative Applications II

Grand Ballroom A, 10:00–12:00, Saturday, 10 Sept. 2016
Chair: Joe Campbell

Speaker Comparison for Forensic and Investigative Applications II

Jean-François Bonastre [1], Joseph P. Campbell [2], Anders Eriksson [3], Hiro Nakasone [4], Reva Schwartz [5]; [1]LIA, France; [2]MIT Lincoln Laboratory, USA; [3]Stockholm University, Sweden; [4]FBI, USA; [5]NIST, USA

Sat-SE-2, Time: 10:00

The aim of this special event is to have several structured discussions on speaker comparison for forensic and investigative applications, where many international experts will present their views and participate in the free exchange of ideas. In speaker comparison, speech samples are compared by humans and/or machines for use in investigations or in court to address questions that are of interest to the legal system. Speaker comparison is a high-stakes application that can change people's lives and it demands the best that science has to offer; however, methods, processes, and practices vary widely. These variations are not necessarily for the better and, although recognized, are not generally appreciated and acted upon. Methods, processes, and practices grounded in science are critical for the proper application (and nonapplication) of speaker comparison to a variety of international investigative and forensic applications. This event follows the successful Interspeech 2015 special event of the same name.

Sat-O-4-2 : Special Session: Clinical and Neuroscience-Inspired Vocal Biomarkers of Neurological and Psychiatric Disorders

Grand Ballroom BC, 10:00–12:00, Saturday, 10 Sept. 2016
Chairs: Nicholas Cummins, Julien Epps, Emily Mower Provost, Thomas Quatieri, Stefan Scherer

Acoustic-Prosodic and Turn-Taking Features in Interactions with Children with Neurodevelopmental Disorders

Daniel Bone [1], Somer Bishop [2], Rahul Gupta [1], Sungbok Lee [1], Shrikanth S. Narayanan [1]; [1]University of Southern California, USA; [2]University of California at San Francisco, USA

Sat-O-4-2-1, Time: 10:00

Atypical speech prosody is a hallmark feature of autism spectrum disorder (ASD) that presents across the lifespan, but is difficult to reliably characterize qualitatively. Given the great heterogeneity of symptoms in ASD, an acoustic-based objective measure would be vital for clinical assessment and interventions. In this study, we investigate speech features in child-psychologist conversational samples, including: segmental and suprasegmental pitch dynamics, speech rate, coordination of prosodic attributes, and turn-taking. Data consist of 95 children with ASD as well as 81 controls with non-ASD developmental disorders. We demonstrate significant predictive performance using these features as well as interpret feature correlations of both interlocutors. The most robust finding is that segmental and suprasegmental prosodic variability increases for both participants in interactions with children having higher ASD severity. Recommendations for future research towards a fully-automatic quantitative measure of speech prosody in neurodevelopmental disorders are discussed.

Automatic Detection of Parkinson's Disease Based on Modulated Vowels

Daria Hemmerling [1], Juan Rafael Orozco-Arroyave [2], Andrzej Skalski [1], Janusz Gajda [1], Elmar Nöth [3]; [1]AGH UST, Poland; [2]Universidad de Antioquia, Colombia; [3]FAU Erlangen-Nürnberg, Germany

Sat-O-4-2-2, Time: 10:15

In this paper we present a novel approach of automatic detection of phonatory and articulatory impairments caused by Parkinson's disease (PD). Modulated (varying between low and high pitch) and

Notes

sustained vowels are considered and analysed. The fundamental frequency of the phonations and its range are computed using the Hilbert-Huang transformation. Additionally, a set with "standard" measures are calculated to model phonatory and articulatory deficits exhibited by Parkinson's patients. Kernel Principal Component Analysis was also applied in order to reduce the dimensionality of the representation space. The automatic discrimination between speakers with PD and healthy controls (HC) is performed using decision trees. According to the results, modulated vowels are suitable to evaluate phonatory and articulatory deficits observed in PD speech.

Towards Automatic Detection of Amyotrophic Lateral Sclerosis from Speech Acoustic and Articulatory Samples

Jun Wang[1], Prasanna V. Kothalkar[1], Beiming Cao[1], Daragh Heitzman[2]; [1]University of Texas at Dallas, USA; [2]Texas Neurology, USA

Sat-O-4-2-3, Time: 10:30

Amyotrophic lateral sclerosis (ALS) is a rapid neurodegenerative disease that affects the speech motor functions of patients, thus causes dysarthria. There is no definite marker for the diagnosis of ALS. Currently, the diagnosis of ALS is primarily based on clinical observations of upper and lower motor neuron damage in the absence of other causes, which is time-consuming, of high cost, and often delayed. Timely diagnosis and assessment for ALS are crucial. Automatic detection of ALS from speech samples would advance the diagnosis of ALS. In this paper, we investigated the automatic detection of ALS from short, pre-symptom speech acoustic and articulatory samples using machine learning approaches (support vector machine and deep neural network). A data set of more than 2,500 speech samples collected from eleven patients with ALS and eleven healthy speakers was used. Leave-subjects-out cross validation experimental results indicate the feasibility of the automatic detection of ALS from speech samples. Adding articulatory motion information (from tongue and lips) further improved the detection performance.

Neurophysiological Vocal Source Modeling for Biomarkers of Disease

Gregory Ciccarelli[1], Thomas F. Quatieri[1], Satrajit S. Ghosh[2]; [1]MIT Lincoln Laboratory, USA; [2]MIT, USA

Sat-O-4-2-4, Time: 10:45

Speech is potentially a rich source of biomarkers for detecting and monitoring neuropsychological disorders. Current biomarkers typically comprise acoustic descriptors extracted from behavioral measures of source, filter, prosodic and linguistic cues. In contrast, in this paper, we extract vocal features based on a neurocomputational model of speech production, reflecting latent or internal motor control parameters that may be more sensitive to individual variation under neuropsychological disease. These features, which are constrained by neurophysiology, may be resilient to artifacts and provide an articulatory complement to acoustic features. Our features represent a mapping from a low-dimensional acoustics-based feature space to a high-dimensional space that captures the underlying neural process including articulatory commands and auditory and somatosensory feedback errors. In particular, we demonstrate a neurophysiological vocal source model that generates biomarkers of disease by modeling vocal source control. By using the fundamental frequency contour and a biophysical representation of the vocal source, we infer two neuromuscular time series whose coordination provides vocal features that are applied to depression and Parkinson's disease as examples. These vocal source coordination features alone, on a single held vowel, outperform or are comparable to other features sets and reflect a significant compression of the feature space.

Relation of Automatically Extracted Formant Trajectories with Intelligibility Loss and Speaking Rate Decline in Amyotrophic Lateral Sclerosis

Rachelle L. Horwitz-Martin[1], Thomas F. Quatieri[1], Adam C. Lammert[1], James R. Williamson[1], Yana Yunusova[2], Elizabeth Godoy[1], Daryush D. Mehta[1], Jordan R. Green[3]; [1]MIT Lincoln Laboratory, USA; [2]University of Toronto, Canada; [3]Harvard-MIT SHBT, USA

Sat-O-4-2-5, Time: 11:00

Effective monitoring of bulbar disease progression in persons with amyotrophic lateral sclerosis (ALS) requires rapid, objective, automatic assessment of speech loss. The purpose of this work was to identify acoustic features that aid in predicting intelligibility loss and speaking rate decline in individuals with ALS. Features were derived from statistics of the first (F_1) and second (F_2) formant frequency trajectories and their first and second derivatives. Motivated by a possible link between components of formant dynamics and specific articulator movements, these features were also computed for low-pass and high-pass filtered formant trajectories. When compared to clinician-rated intelligibility and speaking rate assessments, F_2 features, particularly mean F_2 speed and a novel feature, mean F_2 acceleration, were most strongly correlated with intelligibility and speaking rate, respectively (Spearman correlations > 0.70, p < 0.0001). These features also yielded the best predictions in regression experiments (r > 0.60, p < 0.0001). Comparable results were achieved using low-pass filtered F_2 trajectory features, with higher correlations and lower prediction errors achieved for speaking rate over intelligibility. These findings suggest information can be exploited in specific frequency components of formant trajectories, with implications for automatic monitoring of ALS.

Automatic Analysis of Typical and Atypical Encoding of Spontaneous Emotion in the Voice of Children

Fabien Ringeval[1], Erik Marchi[1], Charline Grossard[2], Jean Xavier[2], Mohamed Chetouani[2], David Cohen[2], Björn Schuller[1]; [1]audEERING, Germany; [2]UPMC, France

Sat-O-4-2-6, Time: 11:15

Children with Autism Spectrum Disorders (ASD) present significant difficulties to understand and express emotions. Systems have thus been proposed to provide objective measurements of acoustic features used by children suffering from ASD to encode emotion in speech. However, only a few studies have exploited such systems to compare different groups of children in their ability to express emotions, and even less have focused on the analysis of spontaneous emotion. In this contribution, we provide insights by extensive evaluations carried out on a new database of spontaneous speech inducing three emotion categories of valence (positive, neutral, and negative). We evaluate the potential of using an automatic

NOTES

recognition system to differentiate groups of children, i.e., pervasive developmental disorders, pervasive developmental disorders not-otherwise specified, specific language impairments, and typically developing, in their abilities to express spontaneous emotion in a common unconstrained task. Results show that all groups of children can be differentiated directly (diagnosis recognition) and indirectly (emotion recognition) by the proposed system.

Recognition of Depression in Bipolar Disorder: Leveraging Cohort and Person-Specific Knowledge

Soheil Khorram, John Gideon, Melvin McInnis, Emily Mower Provost; University of Michigan, USA

Sat-O-4-2-7, Time: 11:30

Individuals with bipolar disorder typically exhibit changes in the acoustics of their speech. Mobile health systems seek to model these changes to automatically detect and correctly identify current states in an individual and to ultimately predict impending mood episodes. We have developed a program, PRIORI (Predicting Individual Outcomes for Rapid Intervention), that analyzes acoustics of speech as predictors of mood states from mobile smartphone data. Mood prediction systems generally assume that the symptomatology of an individual can be modeled using patterns common in a cohort population due to limitations in the size of available datasets. However, individuals are unique. This paper explores person-level systems that can be developed from the current PRIORI database of an extensive and longitudinal collection composed of two subsets: a smaller labeled portion and a larger unlabeled portion. The person-level system employs the unlabeled portion to extract i-vectors, which characterize single individuals. The labeled portion is then used to train person-level and population-level supervised classifiers, operating on the i-vectors and on speech rhythm statistics, respectively. The unification of these two approaches results in a significant improvement over the baseline system, demonstrating the importance of a multi-level approach to capturing depression symptomatology.

Diagnosing People with Dementia Using Automatic Conversation Analysis

Bahman Mirheidari[1], Daniel Blackburn[1], Markus Reuber[2], Traci Walker[1], Heidi Christensen[1]; [1]University of Sheffield, UK; [2]Royal Hallamshire Hospital, UK

Sat-O-4-2-8, Time: 11:45

A recent study using Conversation Analysis (CA) has demonstrated that communication problems may be picked up during conversations between patients and neurologists, and that this can be used to differentiate between patients with (progressive neurodegenerative dementia) ND and those with (nonprogressive) functional memory disorders (FMD). This paper presents a novel automatic method for transcribing such conversations and extracting CA-style features. A range of acoustic, syntactic, semantic and visual features were automatically extracted and used to train a set of classifiers. In a proof-of-principle style study, using data recording during real neurologist-patient consultations, we demonstrate that automatically extracting CA-style features gives a classification accuracy of 95%when using verbatim transcripts. Replacing those transcripts with automatic speech recognition transcripts, we obtain a classification accuracy of 79% which improves to 90% when feature selection is applied. This is a first and encouraging step towards replacing inaccurate, potentially stressful cognitive tests with a test based on monitoring conversation capabilities that could be conducted in e.g. the privacy of the patient's own home.

Sat-O-4-3 : Special Session: Singing Synthesis Challenge: Fill-In the Gap

Bayview A, 10:00–12:00, Saturday, 10 Sept. 2016
Chairs: Christophe d'Alessandro, Axel Roebel, Olivier Deroo

SERAPHIM: A Wavetable Synthesis System with 3D Lip Animation for Real-Time Speech and Singing Applications on Mobile Platforms

*Paul Yaozhu Chan[1], Minghui Dong[1], Grace Xue Hui Ho[2], Haizhou Li[1]; [1]A*STAR, Singapore; [2]NTU, Singapore*

Sat-O-4-3-1, Time: 10:00

Singing synthesis is a rising musical art form gaining popularity amongst composers and end-listeners alike. To date, this art form is largely confined to offline boundaries of the music studio, whereas a large part music is about live performances. This calls for a real-time synthesis system readily deployable for onstage applications.

SERAPHIM is a wavetable synthesis system that is lightweight and deployable on mobile platforms. Apart from conventional offline studio applications, SERAPHIM also supports real-time synthesis applications, enabling live control inputs for on-stage performances. It also provides for easy lip animation control. SERAPHIM will be made available as a toolbox on Unity 3D for easy adoption into game development across multiple platforms. A readily compiled version will also be deployed as a VST studio plugin, directly addressing end users. It currently supports Japanese (singing only) and Mandarin (speech and singing) languages. This paper describes our work on SERAPHIM and discusses its capabilities and applications.

Expressive Singing Synthesis Based on Unit Selection for the Singing Synthesis Challenge 2016

Jordi Bonada, Martí Umbert, Merlijn Blaauw; Universitat Pompeu Fabra, Spain

Sat-O-4-3-2, Time: 10:15

Sample and statistically based singing synthesizers typically require a large amount of data for automatically generating expressive synthetic performances. In this paper we present a singing synthesizer that using two rather small databases is able to generate expressive synthesis from an input consisting of notes and lyrics. The system is based on unit selection and uses the Wide-Band Harmonic Sinusoidal Model for transforming samples. The first database focuses on expression and consists of less than 2 minutes of free expressive singing using solely vowels. The second one is the timbre database which for the English case consists of roughly 35 minutes of monotonic singing of a set of sentences, one syllable per beat. The synthesis is divided in two steps. First, an expressive vowel singing performance of the target song is generated using the expression database. Next, this performance is used as input control of the synthesis using the timbre database and the target lyrics. A selection of synthetic performances have been submitted to the Interspeech Singing Synthesis Challenge 2016, in which they are compared to other competing systems.

Notes

Vocal Effort Modification for Singing Synthesis

Olivier Perrotin, Christophe d'Alessandro; LIMSI, France

Sat-O-4-3-3, Time: 10:30

Vocal effort modification of natural speech is an asset to various applications, in particular, for adding flexibility to concatenative voice synthesis systems. Although decreasing vocal effort is not particularly difficult, increasing vocal effort is a challenging issue. It requires the generation of artificial harmonics in the voice spectrum, along with transformation of the spectral envelope. After a raw source-filter decomposition, harmonic enrichment is achieved by 1/ increasing the source signal impulsiveness using time distortion, 2/ mixing the distorted and natural signals' spectra. Two types of spectral envelope transformations are used: spectral morphing and spectral modeling. Spectral morphing is the transplantation of natural spectral envelopes. Spectral modeling focuses on spectral tilt, formant amplitudes and first formant position modifications. The effectiveness of source enrichment, spectrum morphing, and spectrum modeling for vocal effort modification of sung vowels was evaluated with the help of a perceptive experiment. Results showed a significant positive influence of harmonic enrichment on vocal effort perception with both spectral envelope transformations. Spectral envelope morphing and harmonic enrichment applied on soft voices were perceptively close to natural loud voices. Automatic spectral envelope modeling did not match the results of spectral envelope morphing, but it significantly increased the perception of vocal effort.

Bertsokantari: a TTS Based Singing Synthesis System

Eder del Blanco, Inma Hernaez, Eva Navas, Xabier Sarasola, D. Erro; Universidad del País Vasco, Spain

Sat-O-4-3-4, Time: 10:45

This paper describes the implementation of the Aholab entry for the Singing Synthesis Challenge: Fill-in the Gap. Our approach in this work makes use of an HTS based Text-to-Speech (TTS) synthesizer for Basque to generate the singing voice. The prosody related parameters provided by the TTS system for a spoken version of the score are modified to adapt them to the requirements of the music score concerning syllables duration and tone, while the spectral parameters are basically maintained. The paper describes the processing details developed to improve the quality of the output signal: the syllable timing, the generation of the intonation with vibrato and the manipulation of the model states. In this entry, the lyrics have been freely translated into Basque and the rhythm has been adapted to a Basque traditional rhythm.

Evaluation of Singing Synthesis: Methodology and Case Study with Concatenative and Performative Systems

Lionel Feugère[1], Christophe d'Alessandro[1], Samuel Delalez[1], Luc Ardaillon[2], Axel Roebel[2]; [1]LIMSI, France; [2]IRCAM, France

Sat-O-4-3-5, Time: 11:00

The special session Singing Synthesis Challenge: Fill-In the Gap aims at comparative evaluation of singing synthesis systems. The task is to synthesize a new couplet for two popular songs. This paper address the methodology needed for quality assessment of singing synthesis systems and reports on a case study using 2 systems with a total of 6 different configurations. The two synthesis systems are: a concatenative Text-to-Chant (TTC) system, including a parametric representation of the melodic curve; a Singing Instrument (SI), allowing for real-time interpretation of utterances made of flat-pitch natural voice or diphone concatenated voice. Absolute Category Rating (ACR) and Paired Comparison (PC) tests are used. Natural and natural-degraded reference conditions are used for calibration of the ACR test. The MOS obtained using ACR shows that the TTC (resp. the SI) ranks below natural voice but above (resp. in between) degraded conditions. Then singing synthesis quality is judged better than auto-tuned or distorted natural voice in some cases. PC results show that: 1/ signal processing is an important quality issue, making the difference between systems; 2/ diphone concatenation degrades the quality compared to flat-pitch natural voice; 3/ Automatic melodic modelling is preferred to gestural control for off-line synthesis.

Expressive Control of Singing Voice Synthesis Using Musical Contexts and a Parametric *F0* Model

Luc Ardaillon, Celine Chabot-Canet, Axel Roebel; IRCAM, France

Sat-O-4-3-6, Time: 11:15

Expressive singing voice synthesis requires an appropriate control of both prosodic and timbral aspects. While it is desirable to have an intuitive control over the expressive parameters, synthesis systems should be able to produce convincing results directly from a score. As countless interpretations of a same score are possible, the system should also target a particular singing style, which implies to mimic the various strategies used by different singers. Among the control parameters involved, the pitch (*F0*) should be modeled in priority. In previous work, a parametric *F0* model with intuitive controls has been proposed, but no automatic way to choose the model parameters was given. In the present work, we propose a new approach for modeling singing style, based on parametric templates selection. In this approach, the *F0* parameters and phonemes durations are extracted from annotated recordings, along with a rich description of contextual informations, and stored to form a database of parametric templates. This database is then used to build a model of the singing style using decision-trees. At the synthesis stage, appropriate parameters are then selected according to the target contexts. The results produced by this approach have been evaluated by means of a listening test.

Optimal Unit Stitching in a Unit Selection Singing Synthesis System

Marius Cotescu; Acapela Group, Belgium

Sat-O-4-3-7, Time: 11:30

Unit Selection based speech synthesis systems are currently the best performing, producing natural sounding speech with minimal CPU load. One of the important reasons behind their success is the amount of recordings that are now commonly used in synthesis applications. However, in the case of singing applications, it is quite hard for a database to cover a large phonetic space due to the relative inefficiency of the recording process. Thus, due to the reduced catalogue of units, singing unit selection systems are more likely to produce spectral discontinuity artefacts. Taking advantage of the quasi stable nature of articulation during singing, we propose a novel unit stitching method. The method was implemented into the system that was used for the "Fill-In the Gap" Singing Synthesis Challenge.

NOTES

Sat-O-4-4 : Conversation and Interaction

Bayview B, 10:00–12:00, Saturday, 10 Sept. 2016
Chairs: Margaret Zellers, Julia Hirschberg

The Perception of Overlapping Speech: Effects of Speaker Prosody and Listener Attitudes

Katherine Hilton; Stanford University, USA

Sat-O-4-4-1, Time: 10:00

Speakers use overlapping speech to achieve a range of interactional moves. Competitive overlaps, or interruptions, challenge an interlocutor's control of the conversational floor, while non-competitive overlaps, like back-channeling and co-constructed discourse, communicate engagement with the conversation and ratify the interlocutor's right to be speaking. Being able to evaluate the intentions behind moments of overlap is critical for interlocutors, as well as researchers seeking to model human-human interaction. Researchers have analyzed the acoustics of overlapping speech in order to understand what determines whether an overlap is heard as competitive or non-competitive. They have overwhelmingly found that prosodic prominence plays an important role; incoming overlaps with higher pitch and intensity are more competitive or interruptive. However, no research has directly tested whether and how listeners use prosodic cues to evaluate moments of overlap. Furthermore, much of the current research on classifying overlapping speech ignores listener variability. The present study uses a perception experiment with 500 participants to test the effects of speaker prosody and listener attitudes on the evaluation of overlapping speech. The results demonstrate that prosodic prominence does significantly affect evaluations of overlapping speech, but it is mediated by the listener's own interactional style and attitudes toward overlapping speech.

Who Do You Think Will Speak Next? Perception of Turn-Taking Cues in Slovak and Argentine Spanish

Agustín Gravano[1], Pablo Brusco[1], Štefan Beňuš[2]; [1]Universidad de Buenos Aires, Argentina; [2]UKF, Slovak Republic

Sat-O-4-4-2, Time: 10:20

We investigate perceptual cues in human-human dialogue management related to signalling the change of speaker and the interlocutor's wish to backchannel or contribute with propositional content. We are interested primarily in the relevance of prosodic cues in relation to textual ones, and their cross-linguistic validity by comparing unrelated languages Slovak and Argentine Spanish. Results of a perception study indicate that 1) in addition to textual cues, prosodic cues also play a clear role in perceiving how the dialogue will unfold; and 2) there exists a non-empty intersection of temporal and intonational prosodic turn-taking cues in the two languages, despite their belonging to separate families.

Disentrainment may be a Positive Thing: A Novel Measure of Unsigned Acoustic-Prosodic Synchrony, and its Relation to Speaker Engagement

Juan M. Pérez, Ramiro H. Gálvez, Agustín Gravano; Universidad de Buenos Aires, Argentina

Sat-O-4-4-3, Time: 10:40

Synchrony is a form of entrainment which consists in a relative coordination between two speakers, who throughout conversation simultaneously vary some properties of their speech. We describe two novel measures of acoustic-prosodic synchrony that are derived from a time-series analysis of the speech signal. Both of these measures reward positive synchrony (entrainment) and, while one penalizes negative synchrony (disentrainment), the other one rewards it. We describe significant correlations between the second measure and a number of positive social characteristics of the conversations, such as degree of speaker engagement, in a corpus of task-oriented dialogues in Standard American English. Since these correlations are not found to be significant for the first measure, our results suggest that disentrainment may sometimes have a positive effect on the development of conversation.

Respiratory Turn-Taking Cues

Marcin Włodarczak, Mattias Heldner; Stockholm University, Sweden

Sat-O-4-4-4, Time: 11:00

This paper investigates to what extent breathing can be used as a cue to turn-taking behaviour. The paper improves on existing accounts by considering all possible transitions between speaker states (silent, speaking, backchanneling) and by not relying on global speaker models. Instead, all features (including breathing range and resting expiratory level) are estimated in an incremental fashion using the left-hand context. We identify several inhalatory features relevant to turn-management, and assess the fit of models with these features as predictors of turn-taking behaviour.

The Discourse Marker "so" in Turn-Taking and Turn-Releasing Behavior

Emma Rennie[1], Rebecca Lunsford[2], Peter A. Heeman[2]; [1]Reed College, USA; [2]Oregon Health & Science University, USA

Sat-O-4-4-5, Time: 11:20

Although *so* is a recognized discourse marker, little work has explored its uses in turn-taking, especially when it is not followed by additional speech. In this paper we explore the use of the discourse marker *so* as it pertains to turn-taking and turn-releasing. Specifically, we compare the duration and intensity of *so* when used to take a turn, mid-utterance, and when releasing a turn. We found that durations of turn-retaining tokens are generally shorter than turn-releases; we also found that turn-retaining tokens tend to be lower in intensity than the following speech. These trends of turn-taking behavior alongside certain lexical and prosodic features may prove useful for the development of speech-recognition software.

Acoustic Properties of Formality in Conversational Japanese

Ethan Sherr-Ziarko; University of Oxford, UK

Sat-O-4-4-6, Time: 11:40

This paper examines potential acoustic cues for level of formality in Japanese conversational speech using speech data gathered outside the laboratory, with the objective of using any significant cues to develop a model to predict level of formality in spoken Japanese. Based on previous work on the phonetic properties of formality in Japanese [1],[2] and other languages [3], and on a pilot study of informal geminate contractions in Japanese (section 2), the

NOTES

study examined the mean f_0, articulation rate, and f_0 range (the difference between the minimum and maximum f_0 in an utterance) via direct examination of the data and a functional data analysis [4],[5]. Analysis of the speech data shows significant relationships between all three variables and level of formality, and a binary logistic regression indicates that the variables have some potential as predictors of formality independent of lexical cues, although further refinement of any model will be necessary.

Sat-O-4-5 : Automatic Learning of Representations

Seacliff BCD, 10:00–12:00, Saturday, 10 Sept. 2016
Chair: Carolina Parada

Inferring Phonemic Classes from CNN Activation Maps Using Clustering Techniques

Thomas Pellegrini, Sandrine Mouysset; IRIT, France

Sat-O-4-5-1, Time: 10:00

Today's state-of-art in speech recognition involves deep neural networks (DNN). These last years, a certain research effort has been invested in characterizing the feature representations learned by DNNs. In this paper, we focus on convolutional neural networks (CNN) trained for phoneme recognition in French. We report clustering experiments performed on activation maps extracted from the different layers of a CNN comprised of two convolution and sub-sampling layers followed by three dense layers. Our goal was to get insights into phone separability and phonemic categories inferred by the network, and how they vary according to the successive layers. Two directions were explored with both linear and non-linear clustering techniques. First, we imposed a number of 33 classes equal to the number of context-independent phone models for French, in order to assess the phoneme separability power of the different layers. As expected, we observed that this power increases with the layer depth in the network: from 34% to 74% in F-measure from the first convolution to the last dense layers, when using spectral clustering. Second, optimal numbers of classes were automatically inferred through inter- and intra-cluster measure criteria. We analyze these classes in terms of standard French phonological features.

Joint Learning of Speaker and Phonetic Similarities with Siamese Networks

Neil Zeghidour[1], Gabriel Synnaeve[1], Nicolas Usunier[1], Emmanuel Dupoux[2]; [1]Facebook, France; [2]ENS, France

Sat-O-4-5-2, Time: 10:20

Recent work has demonstrated, on small datasets, the feasibility of jointly learning specialized speaker and phone embeddings, in a weakly supervised siamese DNN architecture using word and speaker identity as side information. Here, we scale up these architectures to the 360 hours of the Librispeech corpus by implementing a sampling method to efficiently select pairs of words from the dataset and improving the loss function. We also compare the standard siamese networks fed with same (AA) or different (AB) pairs, to a 'triamese' network fed with AAB triplets. We use ABX discrimination tasks to evaluate the discriminability and invariance properties of the obtained joined embeddings, and compare these results with mono-embeddings architectures. We find that the joined embeddings architectures succeed in effectively disentangling speaker from phoneme information, with around 10% errors for the matching tasks and embeddings (speaker task on speaker embeddings, and phone task on phone embedding) and near chance for the mismatched task. Furthermore, the results carry over in out-of-domain datasets, even beating the best results obtained with similar weakly supervised techniques.

Unsupervised Learning of Acoustic Units Using Autoencoders and Kohonen Nets

Vikramjit Mitra, Dimitra Vergyri, Horacio Franco; SRI International, USA

Sat-O-4-5-3, Time: 10:40

Often, prior knowledge of subword units is unavailable for low-resource languages. Instead, a global subword unit description, such as a universal phone set, is typically used in such scenarios. One major bottleneck for existing speech-processing systems is their reliance on transcriptions. Unfortunately, the preponderance of data becoming available everyday is only worsening the problem, as properly transcribing, and hence making this data useful for training speech-processing models, is impossible. This work investigates learning acoustic units in an unsupervised manner from real-world speech data by using a cascade of an autoencoder and a Kohonen net. For this purpose, a deep autoencoder with a bottleneck layer at the center was trained with multiple languages. Once trained, the bottleneck-layer output was used to train a Kohonen net, such that state-level ids can be assigned to the bottleneck outputs. To ascertain how consistent such state-level ids are with respect to the acoustic units, phone-alignment information was used for a part of the data to qualify if indeed a functional relationship existed between the phone ids and the Kohonen state ids and, if yes, whether such relationship can be generalized to data that are not transcribed.

Learning Multiscale Features Directly from Waveforms

Zhenyao Zhu, Jesse H. Engel, Awni Hannun; Baidu Research, USA

Sat-O-4-5-4, Time: 11:00

Deep learning has dramatically improved the performance of speech recognition systems through learning hierarchies of features optimized for the task at hand. However, true end-to-end learning, where features are learned directly from waveforms, has only recently reached the performance of hand-tailored representations based on the Fourier transform. In this paper, we detail an approach to use convolutional filters to push past the inherent tradeoff of temporal and frequency resolution that exists for spectral representations. At increased computational cost, we show that increasing temporal resolution via reduced stride and increasing frequency resolution via additional filters delivers significant performance improvements. Further, we find more efficient representations by simultaneously learning at multiple scales, leading to an overall decrease in word error rate on a difficult internal speech test set by 20.7% relative to networks with the same number of parameters trained on spectrograms.

NOTES

Supervised Learning of Acoustic Models in a Zero Resource Setting to Improve DPGMM Clustering

Michael Heck, Sakriani Sakti, Satoshi Nakamura; NAIST, Japan

Sat-O-4-5-5, Time: 11:20

In this work we utilize a supervised acoustic model training pipeline without supervision to improve Dirichlet process Gaussian mixture model (DPGMM) based feature vector clustering. We exploit methods common in supervised acoustic modeling to unsupervisedly learn feature transformations for application to the input data prior to clustering. The idea is to automatically find mappings of feature vectors into sub-spaces that are more robust to channel, context and speaker variability. The need of labels for these techniques makes it difficult to use them in a zero resource setting. To overcome this issue we utilize a first iteration of DPGMM clustering to generate frame based class labels for the target data. The labels serve as basis for learning an acoustic model in the form of hidden Markov models (HMMs) using linear discriminant analysis (LDA), maximum likelihood linear transform (MLLT) and speaker adaptive training (SAT). We show that the learned transformations lead to features that consistently outperform untransformed features on the ABX sound class discriminability task. We also demonstrate that the combination of multiple clustering runs is a suitable method to further enhance sound class discriminability.

Semi-Supervised and Cross-Lingual Knowledge Transfer Learnings for DNN Hybrid Acoustic Models Under Low-Resource Conditions

*Haihua Xu[1], Hang Su[2], Chongjia Ni[3], Xiong Xiao[1], Hao Huang[4], Eng Siong Chng[1], Haizhou Li[1]; [1] TL@NTU, Singapore; [2] University of California at Berkeley, USA; [3] A*STAR, Singapore; [4] Xinjiang University, China*

Sat-O-4-5-6, Time: 11:40

Semi-supervised and cross-lingual knowledge transfer learnings are two strategies for boosting performance of low-resource speech recognition systems. In this paper, we propose a unified knowledge transfer learning method to deal with these two learning tasks. Such a knowledge transfer learning is realized by fine-tuning of Deep Neural Network (DNN). We demonstrate its effectiveness in both monolingual based semi-supervised learning task and cross-lingual knowledge transfer learning task. We then combine these two learning strategies to obtain further performance improvement.

Sat-O-4-6 : Language Modeling for Conversational Speech and Confidence Measures

Seacliff A, 10:00–12:00, Saturday, 10 Sept. 2016

Chairs: Renato de Mori, Dimitra Vergyri

Recurrent Out-of-Vocabulary Word Detection Using Distribution of Features

Taichi Asami[1], Ryo Masumura[1], Yushi Aono[1], Koichi Shinoda[2]; [1] NTT, Japan; [2] Tokyo Institute of Technology, Japan

Sat-O-4-6-1, Time: 10:00

The repeated use of out-of-vocabulary (OOV) words in a spoken document seriously degrades a speech recognizer's performance. This paper provides a novel method for accurately detecting such recurrent OOV words. Standard OOV word detection methods classify each word segment into in-vocabulary (IV) or OOV. This word-by-word classification tends to be affected by sudden vocal irregularities in spontaneous speech, triggering false alarms. To avoid this sensitivity to the irregularities, our proposal focuses on consistency of the repeated occurrence of OOV words. The proposed method preliminarily detects recurrent segments, segments that contain the same word, in a spoken document by open vocabulary spoken term discovery using a phoneme recognizer. If the recurrent segments are OOV words, features for OOV detection in those segments should exhibit consistency. We capture this consistency by using the mean and variance (distribution) of features (DOF) derived from the recurrent segments, and use the DOF for IV/OOV classification. Experiments illustrate that the proposed method's use of the DOF significantly improves its performance in recurrent OOV word detection.

Investigation of Semi-Supervised Acoustic Model Training Based on the Committee of Heterogeneous Neural Networks

Naoyuki Kanda, Shoji Harada, Xugang Lu, Hisashi Kawai; NICT, Japan

Sat-O-4-6-2, Time: 10:20

This paper investigates the semi-supervised training for deep neural network-based acoustic models (AM). In the conventional self-learning approach, a "seed-AM" is first trained by using a small transcribed data set. Then, a large untranscribed data set is decoded by using the seed-AM to create a transcription, which is finally used to train a new AM on the entire data. Our investigation in this paper focuses on the different approach that uses additional complementary AMs to form a committee of label creation for untranscribed data. Especially, we investigate the case of using heterogeneous neural networks as complementary AMs, and the case of intentional exclusion of the primary seed-AM from the committee, both of which could enhance the chance to find more informative training samples for the seed-AM. We investigated those approaches based on Japanese lecture recognition experiments with 50-hours of transcribed data and 190-hours of untranscribed data. In our experiment, the committee-based approach showed significant improvements in the word error rate, and the best method finally recovered 75.2% of the oracle improvement with full manual transcription, while the conventional self-learning approach recovered only 32.7% of the oracle gain.

NOTES

Acoustic Word Embeddings for ASR Error Detection

Sahar Ghannay, Yannick Estève, Nathalie Camelin, Paul deléglise; LIUM, France

Sat-O-4-6-3, Time: 10:40

This paper focuses on error detection in Automatic Speech Recognition (ASR) outputs. A neural network architecture is proposed, which is well suited to handle continuous word representations, like word embeddings. In a previous study, the authors explored the use of linguistic word embeddings, and more particularly their combination. In this new study, the use of acoustic word embeddings is explored. Acoustic word embeddings offer the opportunity of an *a priori* acoustic representation of words that can be compared, in terms of similarity, to an embedded representation of the audio signal.

First, we propose an approach to evaluate the intrinsic performances of acoustic word embeddings in comparison to orthographic representations in order to capture discriminative phonetic information. Since French language is targeted in experiments, a particular focus is made on homophone words. Then, the use of acoustic word embeddings is evaluated for ASR error detection. The proposed approach gets a classification error rate of 7.94% while the previous state-of-the-art CRF-based approach gets a CER of 8.56% on the outputs of the ASR system which won the ETAPE evaluation campaign on speech recognition of French broadcast news.

Combining Semantic Word Classes and Sub-Word Unit Speech Recognition for Robust OOV Detection

Axel Horndasch, Anton Batliner, Caroline Kaufhold, Elmar Nöth; FAU Erlangen-Nürnberg, Germany

Sat-O-4-6-4, Time: 11:00

Out-of-vocabulary words (OOVs) are often the main reason for the failure of tasks like automated voice searches or human-machine dialogs. This is especially true if rare but task-relevant content words, e.g. person or location names, are not in the recognizer's vocabulary. Since applications like spoken dialog systems use the result of the speech recognizer to extract a semantic representation of a user utterance, the detection of OOVs as well as their (semantic) word class can support to manage a dialog successfully. In this paper we suggest to combine two well-known approaches in the context of OOV detection: semantic word classes and OOV models based on sub-word units. With our system, which builds upon the widely used Kaldi speech recognition toolkit, we show on two different data sets that — compared to other methods — such a combination improves OOV detection performance for open word classes at a given false alarm rate. Another result of our approach is a reduction of the word error rate (WER).

Web Data Selection Based on Word Embedding for Low-Resource Speech Recognition

Chuandong Xie [1], Wu Guo [1], Guoping Hu [2], Junhua Liu [3]; [1] USTC, China; [2] Ministry of Public Security, China; [3] iFLYTEK, China

Sat-O-4-6-5, Time: 11:20

The lack of transcription files will lead to a high out-of-vocabulary (OOV) rate and a weak language model in low-resource speech recognition systems. This paper presents a web data selection method to augment these systems. After mapping all the vocabularies or short sentences to vectors in a low-dimensional space through a word embedding technique, the similarities between the web data and the small pool of training transcriptions are calculated. Then, the web data with high similarity are selected to expand the pronunciation lexicon or language model. Experiments are conducted on the NIST Open KWS15 Swahili VLLP recognition task. Compared with the baseline system, our methods can achieve a 5.23% absolute reduction in word error rate (WER) using the expanded pronunciation lexicon and a 9.54% absolute WER reduction using both the expanded lexicon and language model.

Colloquialising Modern Standard Arabic Text for Improved Speech Recognition

Sarah Al-Shareef, Thomas Hain; University of Sheffield, UK

Sat-O-4-6-6, Time: 11:40

Modern standard Arabic (MSA) is the official language of spoken and written Arabic media. Colloquial Arabic (CA) is the set of spoken variants of modern Arabic that exist in the form of regional dialects. CA is used in informal and everyday conversations while MSA is formal communication. An Arabic speaker switches between the two variants according to the situation. Developing an automatic speech recognition system always requires a large collection of transcribed speech or text, and for CA dialects this is an issue. CA has limited textual resources because it exists only as a spoken language, without a standardised written form unlike MSA. This paper focuses on the data sparsity issue in CA textual resources and proposes a strategy to emulate a native speaker in colloquialising MSA to be used in CA language models (LMs) by use of a machine translation (MT) framework. The empirical results in Levantine CA show that using LMs estimated from colloquialised MSA data outperformed MSA LMs with a perplexity reduction up to 68% relative. In addition, interpolating colloquialised MSA LMs with a CA LMs improved speech recognition performance by 4% relative.

Sat-P-4-1 : Topics in Speech Perception

Pacific Concourse – Poster A, 10:00–12:00, Saturday, 10 Sept. 2016
Chair: Bernd Meyer

Pitch-Range Perception: The Dynamic Interaction Between Voice Quality and Fundamental Frequency

Jianjing Kuang, Mark Liberman; University of Pennsylvania, USA

Sat-P-4-1-1, Time: 10:00

Effective pitch-range normalization is important to uncover intended linguistic pitch targets in continuous speech. Our previous study demonstrated that voice quality plays a role in pitch-range perception: "tense voice", implemented as stimuli with spectral balance tilted towards higher frequency, was perceived as higher in pitch. This psychoacoustic effect is consistent with the co-variation between pitch and tense voice in production. However, a spectral balance tilted towards higher frequency is also one of the properties of creaky voice, which is often associated with low pitch in production. Therefore, this raises the possibility that manipulating the f0 range of the stimuli or changing the sex of the speaker of the stimuli can reverse the direction of the shift. This current study replicates the previous experiment with the same forced-choice

NOTES

pitch classification experiment with four spectral conditions, but uses a female voice to create the stimuli. In addition, two f0 ranges are used in the current experiments, which resemble the lower range and the higher range of a female voice. Overall, the results show that spectral balance interacts with f0 range: the presence of voice quality cues affect the perception of pitch range; but, the spectrum with greater energy in the high-frequency range can be interpreted as either creaky or tense depending on the f0 range. This current study enriches our understanding of the interaction between voice quality and pitch.

Comparing the Contributions of Amplitude and Phase to Speech Intelligibility in a Vocoder-Based Speech Synthesis Model

Fei Chen [1], Benson C.L. Chiao [2]; [1] SUSTC, China; [2] University of Hong Kong, China

Sat-P-4-1-2, Time: 10:00

Vocoder-based speech synthesis model has been long used to assess the contribution of acoustic cue for speech recognition. This study compared the perceptual contributions of amplitude and phase by using two types of stimuli, i.e., amplitude- and phase-based vocoded stimuli. The amplitude-based vocoded stimuli were synthesized by preserving amplitude fluctuation cue but discarding phase cue (i.e., setting phase to zero), while the phase-based vocoded stimuli were synthesized by preserving phase cue and discarding amplitude cue (i.e., setting amplitude to unit). Listening experiments with normal-hearing participants showed consistent findings with earlier studies that the intelligibility scores of both amplitude- and phase-based vocoded stimuli increased when using a large number of channels in vocoder-based speech synthesis. In addition, at all tested conditions, the intelligibility scores of amplitude-based vocoded stimuli were significantly larger than those of phase-based vocoded stimuli, suggesting that amplitude might carry more perceptual contribution than phase. This intelligibility advantage of amplitude over phase may be attributed to the difference in the amount of envelope information contained in the two types of vocoded stimuli.

Modeling Noise Influence to Speech Intelligibility Non-Intrusively by Reduced Speech Dynamic Range

Fei Chen; SUSTC, China

Sat-P-4-1-3, Time: 10:00

The noise influence to speech signal waveform can be characterized by reduced speech dynamic range (rDR). This motivated the present work to propose an rDR-based intelligibility measure (denoted as rDRm) that could be used to non-intrusively (i.e., do not require clean reference speech signal) predict speech intelligibility in noise and is computed only using the dynamic range extracted from the noise-corrupted speech. The rDRm indices were evaluated with intelligibility scores obtained from normal-hearing listeners presented with sentences corrupted by four types of maskers in a total of 22 conditions. High correlation (r=0.93) was obtained between rDRm values and listeners' sentence recognition scores, and this correlation was comparable to those computed with existing intrusive and non-intrusive intelligibility measures. This suggests that the dynamic range of speech signal may work as a simple but efficient predictor of speech intelligibility in noise, whose computation does not need access to the clean reference speech signal.

Do GMM Phoneme Classifiers Perceive Synthetic Sibilants as Humans Do?

Gábor Pintér, Hiroki Watanabe; Kobe University, Japan

Sat-P-4-1-4, Time: 10:00

This study presents a psycholinguistically motivated evaluation method for phoneme classifiers by using non-categorical perceptual data elicited in a Japanese sibilant matching 2AFC task. Probability values of a perceptual [s]-[ʃ] boundary, obtained from 42 speakers over a 7-step synthetic [s]-[ʃ] continuum, were compared to probability estimates of Gaussian mixture models (GMMs) of Japanese [s] and [ʃ]. The GMMs, trained on the Corpus of Spontaneous Japanese, differed in feature vectors (MFCC, PLP, acoustic features), covariance matrix types (full, tied, diagonal, spherical), and numbers of mixtures (1–20). Using ten-fold cross validation, it was found that GMMs trained on MFCC features had the best sibilant classification accuracies (87.4–90.4%), but their correlations with human perceptual data were non-conclusive (0.35–0.98). Acoustic feature-based GMMs with tied covariance matrices had near human-like synthetic stimuli perception (0.957–0.996), but their classification performance was poor (71.3–80.4%). Models trained on perceptual linear prediction (PLP) features were on par with the acoustic feature-based models in terms correlation to the perceptual experiment (0.884–0.995), while losing slightly on classification performance (86.1–88.9%) compared to MFCC models. Across the board correlation tests and mixture-effect models confirmed that GMMs with better sibilant classifying performance produced more human-like probability estimations on the synthetic sibilant continuum.

Neural Responses to Speech-Specific Modulations Derived from a Spectro-Temporal Filter Bank

Marina Frye [1], Cristiano Micheli [1], Inga M. Schepers [1], Gerwin Schalk [2], Jochem W. Rieger [1], Bernd T. Meyer [3]; [1] Carl von Ossietzky Universität Oldenburg, Germany; [2] NCAN, USA; [3] Johns Hopkins University, USA

Sat-P-4-1-5, Time: 10:00

This paper analyzes the application of methods developed in automatic speech recognition (ASR) to better understand neural activity measured with electrocorticography (ECoG) during the presentation of speech. ECoG data is collected from temporal cortex in two subjects listening to a matrix sentence test. We investigate the relation of ECoG signals and acoustic speech that has been processed with spectro-temporal filters, which have been shown to produce robust and reliable representations for speech applications. The organization of spectro-temporal filters into a filter bank allows for a straight-forward separation into spectral or temporal only, as well as true spectro-temporal components. We find electrodes positioned over the superior temporal gyrus that is associated with the auditory cortex to show significant specific high gamma activity to fine temporal and spectro-temporal patterns present in speech. This indicates that representations developed in machine listening are a suitable tool for the analysis of biosignals.

Comparing Different Methods for Analyzing ERP Signals

Kimberley Mulder, Louis ten Bosch, Lou Boves; Radboud Universiteit Nijmegen, The Netherlands

Sat-P-4-1-6, Time: 10:00

Event-Related Potential (ERP) signals obtained from EEG recordings are widely used for studying cognitive processes in spoken language

NOTES

processing. The computation of ERPs involves averaging over multiple participants and multiple stimuli. Especially with speech stimuli, which also evoke substantial exogenous excitation, even averaging within conditions results in pooling many sources of variance. This raises questions about the statistical processing needed to uncover reliable differences between conditions. In this study we investigate differences between ERPs when participants listened to full and reduced pronunciations of verb forms in Dutch, in isolation and in mid-sentence position. Conventional statistical analysis uncovers some (but not all) differences between full and reduced forms in isolation, but not in mid-sentence position. In this paper, we show that linear mixed models (lmer) and generalized additive models (gam), which are able to account for participant- and stimulus-related variance may uncover more effects than conventional statistical models. However, depending on the complexity of the data, lmer and gam models may not be able to fit the data closely enough to warrant blind interpretation of the summary output. We discuss opportunities and threats of these approaches to analyzing ERP signals.

Supplementary Motor Area Activation in Disfluency Perception: An fMRI Study of Listener Neural Responses to Spontaneously Produced Unfilled and Filled Pauses

Robert Eklund [1], Martin Ingvar [2]; [1] Linköping University, Sweden; [2] Karolinska Institute, Sweden

Sat-P-4-1-7, Time: 10:00

Spontaneously produced Unfilled Pauses (UPs) and Filled Pauses (FPs) were played to subjects in an fMRI experiment. For both stimuli increased activity was observed in the Primary Auditory Cortex (PAC). However, FPs, but not UPs, elicited modulation in the Supplementary Motor Area (SMA), Brodmann Area 6. Our results provide neurocognitive confirmation of the alleged difference between FPs and other kinds of speech disfluency and could also provide a partial explanation for the previously reported beneficial effect of FPs on reaction times in speech perception. Our results also have potential implications for two of the suggested functions of FPs: the “floor-holding” and the “help-me-out” hypotheses.

Vowel Fundamental and Formant Frequency Contributions to English and Mandarin Sentence Intelligibility

Daniel Fogerty [1], Fei Chen [2]; [1] University of South Carolina, USA; [2] SUSTC, China

Sat-P-4-1-8, Time: 10:00

The current study investigated spectral components of vowels that contribute to Mandarin and English sentence intelligibility. Sentences were processed to preserve various amounts of vowel information. Processing parameters ensured similar proportions of speech preserved between the two languages. In the first experiment, speech segments, primarily containing vocalic cues, were processed to flatten fundamental frequency (F0) cues. In the second experiment, sine-wave speech synthesis was used to coarsely code speech to retain only amplitude and frequency variation associated with the first three formants. Results demonstrated remarkable similarity between Mandarin and English sentence intelligibility with flattened F0 sentences. In contrast, the intelligibility of English sentences surpassed that of Mandarin sentences for sine-wave speech. Combined with earlier reports of superior intelligibility of Mandarin sentences with full spectrum vowels, these results highlight significant contributions of Mandarin F0 information, likely related to lexical tone. In contrast, English listeners may rely more on frequency and/or amplitude variation of the formants.

Sat-P-4-2 : Behavioral Signal Processing and Speaker State and Traits Analytics

Pacific Concourse – Poster B, 10:00–12:00, Saturday, 10 Sept. 2016
Chair: Chung-Hsien Wu

Attention Assisted Discovery of Sub-Utterance Structure in Speech Emotion Recognition

Che-Wei Huang, Shrikanth S. Narayanan; University of Southern California, USA

Sat-P-4-2-1, Time: 10:00

Recently, attention mechanism based deep learning has gained much popularity in speech recognition and natural language processing due to its flexibility at the decoding phase. Through the attention mechanism, the relevant encoding context vectors contribute a majority portion to the construction of the decoding context, while the effect of the irrelevant ones is minimized. Inspired by this idea, a speech emotion recognition system is proposed in this work for an active selection of sub-utterance representations to better compose a discriminative utterance representation. Compared to the baseline of a model based on the uniform attention, i.e. no attention at all, an attention based model improves the weighted accuracy by an absolute of 1.46% (and relative 57.87% to 59.33%) on the emotion classification task. Moreover, the selection distribution leads to a better understanding of the sub-utterance structure in an emotional utterance.

Combining CNN and BLSTM to Extract Textual and Acoustic Features for Recognizing Stances in Mandarin Ideological Debate Competition

Linchuan Li, Zhiyong Wu, Mingxing Xu, Helen Meng, Lianhong Cai; Tsinghua University, China

Sat-P-4-2-2, Time: 10:00

Recognizing stances in ideological debates is a relatively new and challenging problem in opinion mining. While previous work mainly focused on text modality, in this paper, we try to recognize stances from both text and acoustic modalities, where how to derive more representative textual and acoustic features still remains the research problem. Inspired by the promising performances of neural network models in natural language understanding and speech processing, we propose a unified framework named C-BLSTM by combining convolutional neural network (CNN) and bidirectional long short-term memory (BLSTM) recurrent neural network (RNN) for feature extraction. In C-BLSTM, CNN is utilized to extract higher-level local features of text (n-grams) and speech (emphasis, intonation), while BLSTM is used to extract bottleneck features for context-sensitive feature compression and target-related feature representation. Maximum entropy model is then used to recognize stances from the bimodal textual acoustic bottleneck features. Experiments on four debate datasets show C-BLSTM outperforms all challenging baseline methods, and specifically, acoustic intonation and emphasis features further improve F1-measure by 6% as compared to textual features only.

NOTES

Inter-Speech Clicks in an Interspeech Keynote

Jürgen Trouvain, Zofia Malisz; Universität des Saarlandes, Germany

Sat-P-4-2-3, Time: 10:00

Clicks are usually described as phoneme realisations in some African languages or as paralinguistic vocalisations, e.g. to signal disapproval or as sound imitation. A more recent discovery is that clicks are, presumably unintentionally, used as discourse markers indexing a new sequence in a conversation or before a word search. In this single-case study, we investigated more than 300 apical clicks of an experienced speaker during a keynote address at an Interspeech conference. The produced clicks occurred only in inter-speech intervals and were often combined with either hesitation particles like "uhm" or audible inhalation. Our observations suggest a link between click production and ingressive airflow as well as indicate that clicks are used as hesitation markers. The rather high frequency of clicks in the analysed sections from the 1-hour-talk shows that in larger discourse, the time between articulatory phases consists of more than silence, audible inhalation and typical hesitation particles. The rather large variation in the intensity and duration and particularly the number of bursts of the observed clicks indicates that this prosodic discourse marker seems to be a rather acoustically inconsistent phonetic category.

Speaker Age Classification and Regression Using i-Vectors

Joanna Grzybowska, Stanisław Kacprzak; AGH UST, Poland

Sat-P-4-2-4, Time: 10:00

In this paper, we examine the use of i-vectors both for age regression as well as for age classification. Although i-vectors have been previously used for age regression task, we extend this approach by applying fusion of i-vectors and acoustic features regression to estimate the speaker age. By our fusion we obtain a relative improvement of 12.6% comparing to solely i-vector system.

We also use i-vectors for age classification, which to our knowledge is the first attempt to do so. Our best results reach unweighted accuracy 62.9%, which is a relative improvement of 16.7% comparing to the best results obtained in age classification task at *Age Sub-Challenge* at Interspeech 2010.

Sparsely Connected and Disjointly Trained Deep Neural Networks for Low Resource Behavioral Annotation: Acoustic Classification in Couples' Therapy

Haoqi Li[1], Brian Baucom[2], Panayiotis Georgiou[1]; [1]University of Southern California, USA; [2]University of Utah, USA

Sat-P-4-2-5, Time: 10:00

Observational studies are based on accurate assessment of human state. A behavior recognition system that models interlocutors' state in real-time can significantly aid the mental health domain. However, behavior recognition from speech remains a challenging task since it is difficult to find generalizable and representative features because of noisy and high-dimensional data, especially when data is limited and annotated coarsely and subjectively. Deep Neural Networks (DNN) have shown promise in a wide range of machine learning tasks, but for Behavioral Signal Processing (BSP) tasks their application has been constrained due to limited quantity of data.

We propose a Sparsely-Connected and Disjointly-Trained DNN (SD-DNN) framework to deal with limited data. First, we break the acoustic feature set into subsets and train multiple distinct classifiers. Then, the hidden layers of these classifiers become parts of a deeper network that integrates all feature streams. The overall system allows for full connectivity while limiting the number of parameters trained at any time and allows convergence possible with even limited data. We present results on multiple behavior codes in the couples' therapy domain and demonstrate the benefits in behavior classification accuracy. We also show the viability of this system towards live behavior annotations.

Automatically Classifying Self-Rated Personality Scores from Speech

Guozhen An[1], Sarah Ita Levitan[2], Rivka Levitan[3], Andrew Rosenberg[4], Michelle Levine[2], Julia Hirschberg[2]; [1]CUNY Graduate Center, USA; [2]Columbia University, USA; [3]CUNY Brooklyn College, USA; [4]CUNY Queens College, USA

Sat-P-4-2-6, Time: 10:00

Automatic personality recognition is useful for many computational applications, including recommendation systems, dating websites, and adaptive dialogue systems. There have been numerous successful approaches to classify the "Big Five" personality traits from a speaker's utterance, but these have largely relied on judgments of personality obtained from external raters listening to the utterances in isolation. This work instead classifies personality traits based on self-reported personality tests, which are more valid and more difficult to identify. Our approach, which uses lexical and acoustic-prosodic features, yields predictions that are between 6.4% and 19.2% more accurate than chance. This approach predicts Openness-to-Experience and Neuroticism most successfully, with less accurate recognition of Extroversion. We compare the performance of classification and regression techniques, and also explore predicting personality clusters.

Estimation of Children's Physical Characteristics from Their Voices

Jill Fain Lehman, Rita Singh; Disney Research, USA

Sat-P-4-2-7, Time: 10:00

To date, multiple strategies have been proposed for the estimation of speakers' physical parameters such as height, weight, age, gender etc. from their voices. These employ various types of feature measurements in conjunction with different regression and classification mechanisms. While some are quite effective for adults, they are not so for children's voices. This is presumably because in children, the relationship between voice and physical parameters is relatively more complex. The vocal tracts of adults, and the processes that accompany speech production, are fully mature and do not undergo changes within small age differentials. In children, however, these factors change continuously with age, causing variations in style, content, enunciation, rate and quality of their speech. Strategies for the estimation of children's physical parameters from their voice must take this variability into account. In this paper, using different formant-related measurements as exemplary analysis features generated within articulatory-phonetic guidelines, we demonstrate the nonlinear relationships of children's physical parameters to

NOTES

their voice. We also show how such analysis can help us focus on the specific sounds that relate well to each parameter, which can be useful in obtaining more accurate estimates of the physical parameters.

Talking to a System and Talking to a Human: A Study from a Speech-to-Speech, Machine Translation Mediated Map Task

Hayakawa Akira[1], Saturnino Luz[2], Nick Campbell[1]; [1]Trinity College Dublin, Ireland; [2]University of Edinburgh, UK

Sat-P-4-2-8, Time: 10:00

This study focuses on the properties of Human-to-Human (H2H) communication in spontaneous dialogues in two different settings. Direct H2H dialogues are compared to the ones that are mediated by a Speech-to-Speech machine translation system. For the analysis, dialogues from the HCRC Map Task Corpus, for direct H2H conversations, and dialogues from the ILMT-s2s Corpus, for computer mediated conversations, were used. In the conversations speakers take the roles of information giver and follower and all the utterances are labelled as instructions, questions or statement, etc. While direct H2H communication enables speakers also to benefit from non-verbal acts, gestures and facial expressions, machine mediated conversation is more complex for the interlocutors. Due to errors made by speech recognition system, speakers adapt their speaking style and also apply repair strategies in order to accomplish the tasks successfully. Comparing the two corpora showed that in the case of computer mediated communication the utterances of the speakers contained less words than in the case of direct H2H interaction where utterances were longer. Also, different word count was found depending on the role of the speaker as well as on the type of the utterance.

Predicting Affective Dimensions Based on Self Assessed Depression Severity

Rahul Gupta, Shrikanth S. Narayanan; University of Southern California, USA

Sat-P-4-2-9, Time: 10:00

Depression is a state of severe despondency and affects a person's thoughts and behavior. Depression leads to several psychiatric symptoms such as fatigue, restlessness, insomnia as well as other mood disorders (e.g. anxiety and irritation). These symptoms have a resultant impact on the subject's emotional expression. In this work, we address the problem of predicting the emotional dimensions of valence, arousal and dominance in subjects suffering from variable levels of depression, as quantified by the Beck Depression Inventory-II (BDI-II) index. We investigate the relationship between depression severity and affect, and propose a novel method for incorporating the BDI-II index in affect prediction. We validate our models on two datasets recorded as a part of the AViD (Audio-Visual Depressive language) corpus: Freeform and Northwind. Using the depression severity and a set of audio-visual cues, we obtain an average correlation coefficient of .33/.52 for affective dimension prediction in the Freeform/Northwind datasets, against baseline performances of .24/.48 based on using the audio-visual cues only. Our experiments suggest that the knowledge of depression severity significantly improves the emotion dimension prediction, however the BDI-II score incorporation scheme varies between the two datasets of interest.

Enhancement of Automatic Oral Presentation Assessment System Using Latent N-Grams Word Representation and Part-of-Speech Information

Wen-Yu Huang[1], Shan-Wen Hsiao[1], Hung-Ching Sun[1], Ming-Chuan Hsieh[2], Ming-Hsueh Tsai[2], Chi-Chun Lee[1]; [1]National Tsing Hua University, Taiwan; [2]NAER, Taiwan

Sat-P-4-2-10, Time: 10:00

The development of an automatic oral presentation assessment system is important for the educational researchers to assess and train the communication ability of school leaders. In this work, we aim at enhancing the performance of the existing pre-service school principals' presentation scoring system by including lexical information as an additional modality. We propose to use latent n-grams distributed word representations and weighted counts of part-of-speech tag to derive features from the speech transcripts in the National Academy for Educational Research (NAER) oral presentation database. We carry out two different experiments: Exp I is a binary classification task between high versus low performing speech, and Exp II is a continuous scoring on the entire dataset. In Exp I, the proposed framework achieves a competitive accuracy of 0.79, and in Exp II, by fusing this text-based system to the existing audio-video based system, we obtain a spearman correlation of 0.641 (18.05% relative improvement). The two experiments demonstrate the modeling power of our proposed framework and signify the substantial complementary information in the lexical modality while assessing the quality of an oral presentation.

Use of Vowels in Discriminating Speech-Laugh from Laughter and Neutral Speech

Sri Harsha Dumpala, P. Gangamohan, Suryakanth V. Gangashetty, B. Yegnanarayana; IIIT Hyderabad, India

Sat-P-4-2-11, Time: 10:00

In natural conversations, significant part of laughter co-occurs with speech which is referred to as speech-laugh. Hence, speech-laugh will have characteristics of both laughter and neutral speech. But it is not clearly evident how acoustic properties of neutral speech are influenced by its co-occurring laughter. The objective of this study is to analyze the acoustic variations between vowel regions of laughter, speech-laugh and neutral speech. The features based on excitation source characteristics extracted at epochs are considered in this study. Features extracted in the vowel regions of speech-laugh exhibit deviations from that of laughter and neutral speech. These deviations in feature values are exploited to discriminate speech-laugh from laughter and neutral speech. Two different datasets consisting of conversational speech and meeting recordings are used in this analysis. Experimental results show that the discrimination between the three classes obtained by considering vowel regions is better than that of considering the complete utterance.

A Convex Model for Linguistic Influence in Group Conversations

Kan Kawabata, Visar Berisha, Anna Scaglione, Amy LaCross; Arizona State University, USA

Sat-P-4-2-12, Time: 10:00

Conversational partners can influence each other's speaking patterns. In this paper, we aim to develop a computational model that infers influence levels directly from language samples. We propose

NOTES

a new approach to modeling linguistic influence in conversations based on a well-accepted model of social influence. Very generally, this approach assumes that an individual's language model can be expressed as a convex combination of language models from individuals with whom that person interacts. We propose an optimization criterion to estimate the pairwise influence between conversational partners directly from speech and language data. We evaluate the model on three different corpora: (1) a synthetic corpus where the language influence is experimentally set; (2) a corpus that tracks a child's interaction with her family during the early stages of language development; (3) a corpus of Supreme Court cases analyzing interactions between judges and attorneys.

A Deep Learning Approach to Modeling Empathy in Addiction Counseling

James Gibson [1], Doğan Can [1], Bo Xiao [1], Zac E. Imel [2], David C. Atkins [3], Panayiotis Georgiou [1], Shrikanth S. Narayanan [1]; [1] University of Southern California, USA; [2] University of Utah, USA; [3] University of Washington, USA

Sat-P-4-2-13, Time: 10:00

Motivational interviewing is a goal-oriented psychotherapy, employed in cases such as addiction, that aims to help clients explore and resolve their ambivalence about their problem. In motivational interviewing, it is desirable for the counselor to communicate empathy towards the client to promote better therapy outcomes. In this paper, we propose a deep neural network (DNN) system for predicting counselors' session level empathy ratings from transcripts of the interactions. First, we train a recurrent neural network mapping the text of each speaker turn to a set of task-specific behavioral acts that represent local dynamics of the client-counselor interaction. Subsequently, this network is used to initialize lower layers of a deep network predicting session level counselor empathy rating. We show that this method outperforms training the DNN end-to-end in a single stage and also outperforms a baseline neural network model that attempts to predict empathy ratings directly from text without modeling turn level behavioral dynamics.

Unipolar Depression vs. Bipolar Disorder: An Elicitation-Based Approach to Short-Term Detection of Mood Disorder

Kun-Yi Huang [1], Chung-Hsien Wu [1], Yu-Ting Kuo [1], Fong-Lin Jang [2]; [1] National Cheng Kung University, Taiwan; [2] Chi Mei Medical Center, Taiwan

Sat-P-4-2-14, Time: 10:00

Mood disorders include unipolar depression (UD) and bipolar disorder (BD). In this work, an elicitation-based approach to short-term detection of mood disorder based on the elicited speech responses is proposed. First, a long-short term memory (LSTM)-based classifier was constructed to generate the emotion likelihood for each segment in the elicited speech responses. The emotion likelihoods were then clustered into emotion codewords using the K-means algorithm. Latent semantic analysis (LSA) was then adopted to model the latent relationship between the emotion codewords and the elicited responses. The structural relationships among the emotion codewords in the LSA-based matrix were employed to construct a latent affective structure model (LASM) for characterizing each mood. For mood disorder detection, the similarity between the input speech LASM and each of the mood-specific LASMs was estimated. Finally, the mood with its LASM most similar to the input speech LASM is regarded as the detected mood. Experimental results show that the proposed LASM-based method achieved 73.3%, improving the detection accuracy by 13.3% compared to the commonly used SVM-based classifiers.

Sat-P-4-3 : Speech Synthesis Poster

Pacific Concourse – Poster C, 10:00–12:00, Saturday, 10 Sept. 2016
Chair: Ingmar Steiner

Conditional Random Fields for the Tunisian Dialect Grapheme-to-Phoneme Conversion

Abir Masmoudi [1], Mariem Ellouze [2], Fethi Bougares [1], Yannick Esètve [1], Lamia Belguith [2]; [1] LIUM, France; [2] University of Sfax, Tunisia

Sat-P-4-3-1, Time: 10:00

Conditional Random Fields (CRFs) represent an effective approach for monotone string-to-string translation tasks. In this work, we apply the CRF model to perform grapheme-to-phoneme (G2P) conversion for the Tunisian Dialect. This choice is motivated by the fact that CRFs give a long term prediction and assume relaxed state independence conditions compared to HMMs [7]. The CRF model needs to be trained on a 1-to-1 alignement between graphemes and phonemes. Alignments are generated using Joint-Multigram Model (JMM) and GIZA++ toolkit. We trained CRF model for each generated alignment. We then compared our models to state-of-the-art G2P systems based on Sequitur G2P and Phonetisaurus toolkit. We also investigate the CRF prediction quality with different training size. Our results show that CRF perform slightly better using JMM alignment and outperform both Sequitur and Phonetisaurus systems with different training size. At the end, our system gets a phone error rate of 14.09%.

Efficient Thai Grapheme-to-Phoneme Conversion Using CRF-Based Joint Sequence Modeling

Sittipong Saychum, Sarawoot Kongyoung, Anocha Rugchatjaroen, Patcharika Chootrakool, Sawit Kasuriya, Chai Wutiwiwatchai; NECTEC, Thailand

Sat-P-4-3-2, Time: 10:00

This paper presents the successful results of applying joint sequence modeling in Thai grapheme-to-phoneme conversion. The proposed method utilizes Conditional Random Fields (CRFs) in two-stage prediction. The first CRF is used for textual syllable segmentation and syllable type prediction. Graphemes and their corresponding phonemes are then aligned using well-designed many-to-many alignment rules and outputs given by the first CRF. The second CRF, modeling the jointly aligned sequences, efficiently predicts phonemes. The proposed method obviously improves the prediction of *linking syllables*, normally hidden from their textual graphemes. Evaluation results show that the prediction word error rate (WER) of the proposed method reaches 13.66%, which is 11.09% lower than that of the baseline system.

NOTES

An Articulatory-Based Singing Voice Synthesis Using Tongue and Lips Imaging

Aurore Jaumard-Hakoun[1], Kele Xu[1], Clémence Leboullenger[1], Pierre Roussel-Ragot[2], Bruce Denby[3]; [1]UPMC, France; [2]Institut Langevin, France; [3]Tianjin University, China

Sat-P-4-3-3, Time: 10:00

Ultrasound imaging of the tongue and videos of lips movements can be used to investigate specific articulation in speech or singing voice. In this study, tongue and lips image sequences recorded during singing performance are used to predict vocal tract properties via Line Spectral Frequencies (LSF). We focused our work on traditional Corsican singing "Cantu in paghjella". A multimodal Deep Autoencoder (DAE) extracts salient descriptors directly from tongue and lips images. Afterwards, LSF values are predicted from the most relevant of these features using a multilayer perceptron. A vocal tract model is derived from the predicted LSF, while a glottal flow model is computed from a synchronized electroglottographic recording. Articulatory-based singing voice synthesis is developed using both models. The quality of the prediction and singing voice synthesis using this method outperforms the state of the art method.

Phoneme Embedding and its Application to Speech Driven Talking Avatar Synthesis

Xu Li[1], Zhiyong Wu[1], Helen Meng[1], Jia Jia[1], Xiaoyan Lou[2], Lianhong Cai[1]; [1]Tsinghua University, China; [2]Beijing Samsung Telecom R&D Center, China

Sat-P-4-3-4, Time: 10:00

Word embedding has made great achievements in many natural language processing tasks. However, the attempt to apply word embedding to the field of speech got few breakthroughs. The reason is that word vectors mainly contain semantic and syntactic information. Such high level features are difficult to be directly incorporated in speech related tasks compared to acoustic or phoneme related features. In this paper, we investigate the method for phoneme embedding to generate phoneme vectors carrying acoustic information for speech related tasks. One-hot representations of phoneme labels are fed into embedding layer to generate phoneme vectors that are then passed through bidirectional long short-term memory (BLSTM) recurrent neural network to predict acoustic features. Weights in embedding layer are updated through backpropagation during training. Analyses indicate that phonemes with similar acoustic pronunciations are close to each other in cosine distance in the generated phoneme vector space, and tend to be in the same category after k-means clustering. We evaluate the phoneme embedding by applying the generated phoneme vector into speech driven talking avatar synthesis. Experimental results indicate that adding phoneme vector as features can achieve 10.2% relative improvement in objective test.

Expressive Speech Driven Talking Avatar Synthesis with DBLSTM Using Limited Amount of Emotional Bimodal Data

Xu Li[1], Zhiyong Wu[1], Helen Meng[1], Jia Jia[1], Xiaoyan Lou[2], Lianhong Cai[1]; [1]Tsinghua University, China; [2]Beijing Samsung Telecom R&D Center, China

Sat-P-4-3-5, Time: 10:00

One of the essential problems in synthesizing expressive talking avatar is how to model the interactions between emotional facial expressions and lip movements. Traditional methods either simplify such interactions through separately modeling lip movements and facial expressions, or require substantial high quality emotional audio-visual bimodal training data which are usually difficult to collect. This paper proposes several methods to explore different possibilities in capturing the interactions using a large-scale neutral corpus in addition to a small size emotional corpus with limited amount of data. To incorporate contextual influences, deep bidirectional long short-term memory (DBLSTM) recurrent neural network is adopted as the regression model to predict facial features from acoustic features, emotional states as well as contexts. Experimental results indicate that the method by concatenating neutral facial features with emotional acoustic features as the input of DBLSTM model achieves the best performance in both objective and subjective evaluations.

Audio-to-Visual Speech Conversion Using Deep Neural Networks

Sarah Taylor[1], Akihiro Kato[1], Iain Matthews[2], Ben Milner[1]; [1]University of East Anglia, UK; [2]Disney Research, USA

Sat-P-4-3-6, Time: 10:00

We study the problem of mapping from acoustic to visual speech with the goal of generating accurate, perceptually natural speech animation automatically from an audio speech signal. We present a sliding window deep neural network that learns a mapping from a window of acoustic features to a window of visual features from a large audio-visual speech dataset. Overlapping visual predictions are averaged to generate continuous, smoothly varying speech animation. We outperform a baseline HMM inversion approach in both objective and subjective evaluations and perform a thorough analysis of our results.

Generative Acoustic-Phonemic-Speaker Model Based on Three-Way Restricted Boltzmann Machine

Toru Nakashika, Yasuhiro Minami; University of Electro-Communications, Japan

Sat-P-4-3-7, Time: 10:00

In this paper, we argue the way of modeling speech signals based on three-way restricted Boltzmann machine (3WRBM) for separating phonetic-related information and speaker-related information from an observed signal automatically. The proposed model is an energy-based probabilistic model that includes three-way potentials of three variables: acoustic features, latent phonetic features, and speaker-identity features. We train the model so that it automatically captures the undirected relationships among the three variables. Once the model is trained, it can be applied to many tasks in speech signal processing. For example, given a speech signal, estimating speaker-identity features is equivalent to speaker recognition; on the other hand, estimated latent phonetic features may be helpful for speech recognition because they contain more phonetic-related information than the acoustic features. Since the model is generative, we can also apply it to voice conversion; i.e., we just estimate acoustic features from the phonetic features that were estimated given the source speakers acoustic features along with the desired speaker-identity features. In our experiments, we discuss the effectiveness of the speech modeling through a speaker recognition, a speech (continuous phone) recognition, and a voice conversion tasks.

NOTES

Articulatory Synthesis Based on Real-Time Magnetic Resonance Imaging Data

Asterios Toutios, Tanner Sorensen, Krishna Somandepalli, Rachel Alexander, Shrikanth S. Narayanan; University of Southern California, USA

Sat-P-4-3-8, Time: 10:00

This paper presents a methodology for articulatory synthesis of running speech in American English driven by real-time magnetic resonance imaging (rtMRI) mid-sagittal vocal-tract data. At the core of the methodology is a time-domain simulation of the propagation of sound in the vocal tract developed previously by Maeda. The first step of the methodology is the automatic derivation of air-tissue boundaries from the rtMRI data. These articulatory outlines are then modified in a systematic way in order to introduce additional precision in the formation of consonantal vocal-tract constrictions. Other elements of the methodology include a previously reported set of empirical rules for setting the time-varying characteristics of the glottis and the velopharyngeal port, and a revised sagittal-to-area conversion. Results are promising towards the development of a full-fledged text-to-speech synthesis system leveraging directly observed vocal-tract dynamics.

Deep Neural Network Based Acoustic-to-Articulatory Inversion Using Phone Sequence Information

Xurong Xie, Xunying Liu, Lan Wang; Chinese Academy of Sciences, China

Sat-P-4-3-9, Time: 10:00

In recent years, neural network based acoustic-to-articulatory inversion approaches have achieved the state-of-the-art performance. One major issue associated with these approaches is the lack of phone sequence information during inversion. In order to address this issue, this paper proposes an improved architecture hierarchically concatenating phone classification and articulatory inversion component DNNs to improve articulatory movement generation. On a Mandarin Chinese speech inversion task, the proposed technique consistently outperformed a range of baseline DNN and RNN inversion systems constructed using no phone sequence information, a mixture density parameter output layer, additional phone features at the input layer, or multi-task learning with additional monophone output layer target labels, measured in terms of electromagnetic articulography (EMA) root mean square error (RMSE) and correlation. Further improvements were obtained using the bottleneck features extracted from the proposed hierarchical articulatory inversion systems as auxiliary features in generalized variable parameter HMMs (GVP-HMMs) based inversion systems.

Articulatory-to-Acoustic Conversion with Cascaded Prediction of Spectral and Excitation Features Using Neural Networks

Zheng-Chen Liu, Zhen-Hua Ling, Li-Rong Dai; USTC, China

Sat-P-4-3-10, Time: 10:00

This paper presents an articulatory-to-acoustic conversion method using electromagnetic midsagittal articulography (EMA) measurements as input features. Neural networks, including feed-forward deep neural networks (DNNs) and recurrent neural networks (RNNs) with long short-term term memory (LSTM) cells, are adopted to map EMA features towards not only spectral features (i.e. mel-cepstra) but also excitation features (i.e. power, U/V flag and F0). Then speech waveforms are reconstructed using the predicted spectral and excitation features. A cascaded prediction strategy is proposed to utilize the predicted spectral features as auxiliary input to boost the prediction of excitation features. Experimental results show that LSTM-RNN models can achieve better objective and subjective performance in articulatory-to-spectral conversion than DNNs and Gaussian mixture models (GMMs). The strategy of cascaded prediction can increase the accuracy of excitation feature prediction and the neural network-based methods also outperform the GMM-based approach when predicting power features.

Generating Gestural Scores from Acoustics Through a Sparse Anchor-Based Representation of Speech

Christopher Liberatore, Ricardo Gutierrez-Osuna; Texas A&M University, USA

Sat-P-4-3-11, Time: 10:00

We present a procedure for generating gestural scores from speech acoustics. The procedure is based on our recent SABR (sparse, anchor-based representation) algorithm, which models the speech signal as a linear combination of acoustic anchors. We present modifications to SABR that encourage temporal smoothness by restricting the number of anchors that can be active over an analysis window. We propose that peaks in the SABR weights can be interpreted as "keyframes" that determine when vocal tract articulations occur. We validate the approach in two ways. First, we compare SABR keyframes to maxima in the velocity of electromagnetic articulography (EMA) pellets from an articulatory corpus. Second, we use keyframes and SABR weights to build a gestural score for the VocalTractLab (VTL) model, and compare synthetic EMA trajectories generated by VTL against those in the articulatory corpus. We find that SABR keyframes occur within 15–20 ms (on average) of EMA maxima, suggesting that SABR keyframes can be used to identify articulatory phenomena. However, comparison of synthetic and real EMA pellets show moderate correlation on tongue pellets but weak correlation on lip pellets, a result that may be due to differences between the VTL speaker model and the source speaker in our corpus.

On the Suitability of Vocalic Sandwiches in a Corpus-Based TTS Engine

David Guennec, Damien Lolive; IRISA, France

Sat-P-4-3-12, Time: 10:00

Unit selection speech synthesis systems generally rely on target and concatenation costs for selecting the best unit sequence. The role of the concatenation cost is to insure that joining two voice segments will not cause any acoustic artefact to appear. For this task, acoustic distances (MFCC, F_0) are typically used but in many cases, this is not enough to prevent concatenation artefacts. Among other strategies, the improvement of corpus covering by favoring units that naturally support well the joining process (vocalic sandwiches) seems to be effective on TTS. In this paper, we investigate if vocalic sandwiches can be used directly in the unit selection engine when the corpus was not created using that principle. First, the sandwich approach is directly transposed in the unit selection engine with a penalty that greatly favors concatenation on sandwich boundaries. Second, a derived fuzzy version is proposed to relax the penalty based on the concatenation cost, with respect to the cost distribution.

Notes

We show that the sandwich approach, very efficient at the corpus creation step, seems to be inefficient when directly transposed in the unit selection engine. However, we observe that the fuzzy approach enhances synthesis quality, especially on sentences with high concatenation costs.

Unsupervised Stress Information Labeling Using Gaussian Process Latent Variable Model for Statistical Speech Synthesis

Decha Moungsri, Tomoki Koriyama, Takao Kobayashi; Tokyo Institute of Technology, Japan

Sat-P-4-3-13, Time: 10:00

In Thai language, stress is an important prosodic feature that not only affects naturalness but also has a crucial role in meaning of phrase-level utterance. It is seen that a speech synthesis model that is trained with lack of stress and phrase-level information causes incorrect tones and ambiguity in meaning of synthetic speech. Our previous work has shown that manually annotated stress information improves naturalness of synthetic speech. However, a high time consumption is a drawback of the manual annotation. In this paper, we utilize an unsupervised learning technique called Bayesian Gaussian process latent variable model (Bayesian GP-LVM) to automatically put stress annotation on the given training data. Stress related features are projected onto a latent space in which syllables are easier classified into stressed/unstressed classes. We use the stressed/unstressed information as an additional context in GPR-based speech synthesis. Experimental results show that the proposed technique improves naturalness of synthetic speech as well as accuracy of stressed/unstressed classification. Moreover, the proposed technique enables us to avoid ambiguity in meaning of synthetic speech by providing intended stress position into context label sequence to be synthesized.

Using Zero-Frequency Resonator to Extract Multilingual Intonation Structure

Jinfu Ni, Yoshinori Shiga, Hisashi Kawai; NICT, Japan

Sat-P-4-3-14, Time: 10:00

Human uses expressive intonation to convey linguistic and paralinguistic meaning, especially making focal prominence to give emphasis that highlights the focus of speech. Automatic extraction of dynamic intonation feature from a speech corpus and representing it in a continuous form are desired in multilingual speech synthesis. This paper presents a method to extract dynamic prosodic structure from speech signal using zero-frequency resonator to detect glottal cycle epoch and filter both voice amplitude and fundamental frequency (F0) contours. We choose stable voice F0 segments free from micro-prosodic effect to recover relevant F0 trajectory of an utterance, taking into consideration of intercorrelation of micro-prosody with phonetic segments and syllable structure of the utterance, and further filter out long-term global pitch movements. The method is evaluated by objective tests upon multilingual speech corpora including Chinese, Japanese, Korean, and Myanmar. Our experiment results show that the extracted intonation contour can match F0 contour by conventional approach in very high accuracy and the estimated long-term pitch movements demonstrate regular characteristics of intonation across languages. The proposed method is language-independent and robust to noisy speech.

Sat-P-4-4 : Resources and Annotation of Resources

Pacific Concourse – Poster D, 10:00–12:00, Saturday, 10 Sept. 2016
Chair: Dilek Hakkani-Tur

A DNN-HMM Approach to Story Segmentation

Jia Yu[1], Xiong Xiao[2], Lei Xie[1], Eng Siong Chng[2], Haizhou Li[2]; [1]Northwestern Polytechnical University, China; [2]TL@NTU, Singapore

Sat-P-4-4-1, Time: 10:00

Hidden Markov model (HMM) is one of the popular techniques for story segmentation, where hidden Markov states represent the topics, and the emission distributions of n-gram language model (LM) are dependent on the states. Given a text document, a Viterbi decoder finds the hidden story sequence, with a change of topic indicating a story boundary. In this paper, we propose a discriminative approach to story boundary detection. In the HMM framework, we use deep neural network (DNN) to estimate the posterior probability of topics given the bag-of-words in the local context. We call it the DNN-HMM approach. We consider the topic dependent LM as a generative modeling technique, and the DNN-HMM as the discriminative solution. Experiments on topic detection and tracking (TDT2) task show that DNN-HMM outperforms traditional n-gram LM approach significantly and achieves state-of-the-art performance.

The SIWIS Database: A Multilingual Speech Database with Acted Emphasis

Jean-Philippe Goldman[1], Pierre-Edouard Honnet[2], Rob Clark[3], Philip N. Garner[2], Maria Ivanova[1], Alexandros Lazaridis[2], Hui Liang[4], Tiago Macedo[1], Beat Pfister[4], Manuel Sam Ribeiro[3], Eric Wehrli[1], Junichi Yamagishi[3]; [1]Université de Genève, Switzerland; [2]Idiap Research Institute, Switzerland; [3]University of Edinburgh, UK; [4]ETH Zürich, Switzerland

Sat-P-4-4-2, Time: 10:00

We describe here a collection of speech data of bilingual and trilingual speakers of English, French, German and Italian. In the context of speech to speech translation (S2ST), this database is designed for several purposes and studies: training CLSA systems (cross-language speaker adaptation), conveying emphasis through S2ST systems, and evaluating TTS systems. More precisely, 36 speakers judged as accentless (22 bilingual and 14 trilingual speakers) were recorded for a set of 171 prompts in two or three languages, amounting to a total of 24 hours of speech. These sets of prompts include 100 sentences from news, 25 sentences from Europarl, the same 25 sentences with one acted emphasised word, 20 semantically unpredictable sentences, and finally a 240-word long text. All in all, it yielded 64 bilingual session pairs of the six possible combinations of the four languages. The database is freely available for non-commercial use and scientific research purposes.

NOTES

Open Source Speech and Language Resources for Frisian

Emre Yılmaz[1], Henk van den Heuvel[1], Jelske Dijkstra[2], Hans Van de Velde[2], Frederik Kampstra[3], Jouke Algra[3], David Van Leeuwen[1]; [1]Radboud Universiteit Nijmegen, The Netherlands; [2]Fryske Akademy, The Netherlands; [3]Omrop Fryslân, The Netherlands

Sat-P-4-4-3, Time: 10:00

In this paper, we present several open source speech and language resources for the under-resourced Frisian language. Frisian is mostly spoken in the province of Fryslân which is located in the north of the Netherlands. The native speakers of Frisian are Frisian-Dutch bilingual and often code-switch in daily conversations. The resources presented in this paper include a code-switching speech database containing radio broadcasts, a phonetic lexicon with more than 70k words and a language model trained on a text corpus with more than 38M words. With this contribution, we aim to share the Frisian resources we have collected in the scope of the FAME! project, in which a spoken document retrieval system is built for the disclosure of the regional broadcaster's radio archives. These resources enable research on code-switching and longitudinal speech and language change. Moreover, a sample automatic speech recognition (ASR) recipe for the Kaldi toolkit will also be provided online to facilitate the Frisian ASR research.

The SRI CLEO Speaker-State Corpus

Andreas Kathol, Elizabeth Shriberg, Massimilano de Zambotti; SRI International, USA

Sat-P-4-4-4, Time: 10:00

We introduce the SRI CLEO (Conversational Language about Everyday Objects) Speaker-State Corpus of speech, video, and biosignals. The goal of the corpus is providing insight on the speech and physiological changes resulting from subtle, context-based influences on affect and cognition. Speakers were prompted by collections of pictures of neutral everyday objects and were instructed to provide speech related to any subset of the objects for a preset period of time (120 or 180 seconds depending on task).

The corpus provides signals for 43 speakers under four different speaker-state conditions: (1) neutral and emotionally charged audiovisual background; (2) cognitive load; (3) time pressure; and (4) various acted emotions. Unlike previous studies that have linked speaker state to the content of the speaking task itself, the CLEO prompts remain largely pragmatically, semantically, and affectively neutral across all conditions. This framework enables for more direct comparisons across both conditions and speakers. The corpus also includes more traditional speaker tasks involving reading and free-form reporting of neutral and emotionally charged content. The explored biosignals include skin conductance, respiration, blood pressure, and ECG. The corpus is in the final stages of processing and will be made available to the research community.

SingaKids-Mandarin: Speech Corpus of Singaporean Children Speaking Mandarin Chinese

*Nancy F. Chen[1], Rong Tong[1], Darren Wee[2], Peixuan Lee[2], Bin Ma[1], Haizhou Li[1]; [1]A*STAR, Singapore; [2]NUS, Singapore*

Sat-P-4-4-5, Time: 10:00

We present *SingaKids-Mandarin*, a speech corpus of 255 Singaporean children aged 7 to 12 reading Mandarin Chinese, for a total of 125 hours of data (75 hours of speech) and 79,843 utterances. This corpus is phonetically balanced and detailed in human annotations, including phonetic transcriptions, lexical tone markings, and proficiency scoring at the utterance level. The reading scripts span a diverse set of utterance styles, covering syllable-level minimal pairs, words, phrases, sentences, and short stories. We analyze the acoustic properties of Singaporean children. We also observe that while the lack of the neutral tone is the same for Singaporean adults and children, the phonetic pronunciation patterns in these two age groups differ: although Singaporean adults tend to front their retroflex, nasal, and palatal consonants, Singaporean children show both fronting and backing in these consonants. For future work, we plan to develop computer-assisted pronunciation training (CAPT) systems with *SingaKids-Mandarin*.

The SRI Speech-Based Collaborative Learning Corpus

Colleen Richey, Cynthia D'Angelo, Nonye Alozie, Harry Bratt, Elizabeth Shriberg; SRI International, USA

Sat-P-4-4-6, Time: 10:00

We introduce the SRI speech-based collaborative learning corpus, a novel collection designed for the investigation and measurement of how students collaborate together in small groups. This is a multi-speaker corpus containing high-quality audio recordings of middle school students working in groups of three to solve mathematical problems. Each student was recorded via a head-mounted noise-cancelling microphone. Each group was also recorded via a stereo microphone placed nearby. A total of 80 sessions were collected with the participation of 134 students. The average duration of a session was 20 minutes. All students spoke English; for some students, English was a second language. Sessions have been annotated with time stamps to indicate which mathematical problem the students were solving and which student was speaking. Sessions have also been hand annotated with common indicators of collaboration for each speaker (e.g., inviting others to contribute, planning) and the overall collaboration quality for each problem. The corpus will be useful to education researchers interested in collaborative learning and to speech researchers interested in children's speech, speech analytics, and speech diarization. The corpus, both audio and annotation, will be made available to researchers.

An Expectation Maximization Approach to Joint Modeling of Multidimensional Ratings Derived from Multiple Annotators

Anil Ramakrishna[1], Rahul Gupta[1], Ruth B. Grossman[2], Shrikanth S. Narayanan[1]; [1]University of Southern California, USA; [2]Emerson College, USA

Sat-P-4-4-7, Time: 10:00

Ratings from multiple human annotators are often pooled in applications where the ground truth is hidden. Examples include annotating perceived emotions and assessing quality metrics for speech and image. These ratings are not restricted to a single dimension and can be multidimensional. In this paper, we propose an Expectation-Maximization based algorithm to model such ratings. Our model assumes that there exists a latent multidimensional ground truth that can be determined from the observation features and that the ratings provided by the annotators are noisy versions of the ground truth. We test our model on a study conducted on children with autism to predict a four dimensional rating of

Notes

expressivity, naturalness, pronunciation goodness and engagement. Our goal in this application is to reliably predict the individual annotator ratings which can be used to address issues of cognitive load on the annotators as well as the rating cost. We initially train a baseline directly predicting annotator ratings from the features and compare it to our model under three different settings assuming: (i) each entry in the multidimensional rating is independent of others, (ii) a joint distribution among rating dimensions exists, (iii) a partial set of ratings to predict the remaining entries is available.

Voting Detector: A Combination of Anomaly Detectors to Reveal Annotation Errors in TTS Corpora

Jindřich Matoušek, Daniel Tihelka; University of West Bohemia, Czech Republic

Sat-P-4-4-8, Time: 10:00

Anomaly detection techniques were shown to help in detecting word-level annotation errors in read-speech corpora for text-to-speech synthesis. In this framework, correctly annotated words are considered as normal examples on which the detection methods are trained. Misannotated words are then taken as anomalous examples which do not conform to normal patterns of the trained detection models. In this paper we propose a concept of a voting detector — a combination of anomaly detectors in which each "single" detector "votes" on whether a testing word is annotated correctly or not. The final decision is then made by aggregating the votes. Our experiments show that voting detector has a potential to overcome each of the single anomaly detectors.

Sat-S&T-4 : Show & Tell Session 4

Market Street Foyer, 10:00–12:00, Saturday, 10 Sept. 2016
Chairs: Nicolas Scheffer, Shiva Sundaram

The Magic Stone: A Video Game to Improve Communication Skills of People with Intellectual Disabilities

Mario Corrales-Astorgano[1], David Escudero-Mancebo[1], César González-Ferreras[1], Yurena Gutiérrez-González[2], Valle Flores-Lucas[1], Valentín Cardeñoso-Payo[1], Lourdes Aguilar-Cuevas[2]; [1]Universidad de Valladolid, Spain; [2]Universidad Autónoma de Barcelona, Spain

Sat-S&T-4-1, Time: 10:00

"The Magic Stone" is a video game whose main aim is to help people with Down syndrome to improve communication skills that have been affected due to their disability, especially those related with prosody. The interface of the video game includes a number of elements to motivate the users to practice and train their pronunciation. The usability tests of the system have reported high degrees of satisfaction of users and trainers. Perception tests have permitted to confirm that players improve the use of prosody with the use.

Identifying Perceptually Similar Voices with a Speaker Recognition System Using Auto-Phonetic Features

Finnian Kelly[1], Anil Alexander[1], Oscar Forth[1], Samuel Kent[1], Jonas Lindh[2], Joel Åkesson[2]; [1]Oxford Wave Research, UK; [2]Voxalys, Sweden

Sat-S&T-4-2, Time: 10:00

Assessing the perceptual similarity of voices is necessary for the creation of voice parades, along with media applications such as voice casting. These applications are normally prohibitively expensive to administer, requiring significant amounts of 'expert listening'. The ability to automatically assess voice similarity could benefit these applications by increasing efficiency and reducing subjectivity, while enabling the use of a much larger search space of candidate voices. In this paper, the use of automatically extracted phonetic features within an i-vector speaker recognition system is proposed as a means of identifying cohorts of perceptually similar voices. Features considered include formants (F1-F4), fundamental frequency (F0), semitones of F0, and their derivatives. To demonstrate the viability of this approach, a subset of the Interspeech 2016 special session 'Speakers In The Wild' (SITW) dataset is used in a pilot study comparing subjective listener ratings of similarity with the output of the automatic system. It is observed that the automatic system can locate cohorts of male voices with good perceptual similarity. In addition to these experiments, this proposal will be demonstrated with an application allowing a user to retrieve voices perceptually similar to their own from a large dataset.

A Real-Time Framework for Visual Feedback of Articulatory Data Using Statistical Shape Models

Kristy James[1], Alexander Hewer[1], Ingmar Steiner[1], Stefanie Wuhrer[2]; [1]Universität des Saarlandes, Germany; [2]Inria, France

Sat-S&T-4-3, Time: 10:00

We present a novel open-source framework for visualizing electromagnetic articulography (EMA) data in real-time, with a modular framework and anatomically accurate tongue and palate models derived by multilinear subspace learning.

Flexible, Rapid Authoring of Goal-Orientated, Multi-Turn Dialogues Using the Task Completion Platform

Alex Marin, Paul Crook, Omar Zia Khan, Vasiliy Radostev, Khushboo Aggarwal, Ruhi Sarikaya; Microsoft, USA

Sat-S&T-4-4, Time: 10:00

The Task Completion Platform (TCP) is a multi-domain, multi-modal dialogue system that can host and execute large numbers of goal-orientated dialogue tasks. TCP is comprised of a task configuration language, TaskForm, and a task-independent dialogue runtime, allowing task definitions to be decoupled from the global dialogue policy used by the platform to execute the tasks. This separation enables scenario developers to rapidly develop new dialogue systems, by eliminating the need to re-implement the policy from scratch for each new task. In this paper, we introduce support for authoring tasks in a variety of dialogue styles, ranging from entirely flexible to fully system-initiative. This flexibility is enabled by a set of task-level policy override constructs, which augment or constrain the default platform-level policy to achieve the desired system behavior. We demonstrate the use of the TaskForm language to define complex, multi-turn tasks in a variety of domains and add different task-specific policy constructs to demonstrate the flexibility of the task authoring process.

Sat-O-5-1 : Acoustic Model Adaptation

Grand Ballroom A, 13:30–15:30, Saturday, 10 Sept. 2016
Chairs: Brian Mak, Kai Yu

Context Adaptive Neural Network for Rapid Adaptation of Deep CNN Based Acoustic Models

Marc Delcroix, Keisuke Kinoshita, Atsunori Ogawa, Takuya Yoshioka, Dung T. Tran, Tomohiro Nakatani; NTT, Japan

Sat-O-5-1-1, Time: 13:30

Using auxiliary input features has been seen as one of the most effective ways to adapt deep neural network (DNN)-based acoustic models to speaker or environment. However, this approach has several limitations. It only performs compensation of the bias term of the hidden layer and therefore does not fully exploit the network capabilities. Moreover, it may not be well suited for certain types of architectures such as convolutional neural networks (CNNs) because the auxiliary features have different time-frequency structures from speech features. This paper resolves these problems by extending the recently proposed context adaptive DNN (CA-DNN) framework to CNN architectures. A CA-DNN is a DNN with one or several layers factorized in sub-layers associated with an acoustic context class representing speaker or environment. The output of the factorized layer is obtained as the weighted sum of the contributions of each sub-layer, weighted by acoustic context weights that are derived from auxiliary features such as i-vectors. Importantly, a CA-DNN can compensate both bias and weight matrices. In this paper, we investigate the use of CA-DNN for deep CNN-based architectures. We demonstrate consistent performance gains for utterance level rapid adaptation on the AURORA4 task over a strong network-in-network based deep CNN architecture.

Transfer Learning with Bottleneck Feature Networks for Whispered Speech Recognition

*Boon Pang Lim[1], Faith Wong[2], Yuyao Li[2], Jia Wei Bay[2]; [1]A*STAR, Singapore; [2]Nanyang Girls' High School, Singapore*

Sat-O-5-1-2, Time: 13:50

Previous work on whispered speech recognition has shown that acoustic models (AM) trained on whispered speech can somewhat classify unwhispered (neutral) speech sounds, but not vice versa. In fact, AMs trained purely on neutral speech completely fail to recognize whispered speech. Meanwhile, recipes used to train neutral AMs will work just as well for whispered speech, but such methods require a large volume of transcribed whispered speech which is expensive to gather. In this work, we propose and investigate the use of bottleneck feature networks to normalize differences between whispered and neutral speech modes. Our extensive experiments show that this type of speech variability can be effectively normalized. We also show that it is possible to transfer this knowledge from two source languages with whispered speech (Mandarin and English), to a new target language (Malay) without whispered speech. Furthermore, we report a substantial reduction in word error rate for cross-mode speech recognition, effectively demonstrate that it is possible to train acoustic models capable of classifying both types of speech without needing any additional whispered speech.

Adaptation of Neural Networks Constrained by Prior Statistics of Node Co-Activations

Tasha Nagamine, Zhuo Chen, Nima Mesgarani; Columbia University, USA

Sat-O-5-1-3, Time: 14:10

We propose a novel unsupervised model adaptation framework in which a neural network uses prior knowledge of the statistics of its output and hidden layer activations to update its parameters online to improve performance in mismatched environments. This idea is inspired by biological neural networks, which use feedback to dynamically adapt their computation when faced with unexpected inputs. Here, we introduce an adaptation criterion for deep neural networks based on the observation that in matched testing and training conditions, the node co-activation statistics of each layer in a neural network are relatively stable over time. The proposed method thus adapts the model layer by layer to minimize the distance between the co-activation statistics of nodes in matched versus mismatched conditions. In phoneme classification experiments, we show that such node co-activation constrained adaptation in a deep neural network model significantly improves the recognition accuracy over baseline performance when the system is tested in various novel noises not included in the training.

Domain Adaptation of CNN Based Acoustic Models Under Limited Resource Settings

Masayuki Suzuki[1], Ryuki Tachibana[1], Samuel Thomas[2], Bhuvana Ramabhadran[2], George Saon[2]; [1]IBM, Japan; [2]IBM, USA

Sat-O-5-1-4, Time: 14:30

Adaptation of Automatic Speech Recognition (ASR) systems to a new domain (channel, speaker, topic, etc.) remains a significant challenge, as often, only a limited amount of target domain data for adaptation of Acoustic Models (AMs) is available. However, unlike GMMs, to date, there has not been an established, efficient method for adapting current state-of-the-art Convolutional Neural Network (CNN)-based AMs. In this paper, we explore various training algorithms for domain adaptation of CNN based speech recognition systems with limited acoustic training data resources. Our investigations illustrate the following three main contributions. First, introducing a weight decay based regularizer along with the standard cross entropy criteria can significantly improve recognition performances with as little as one hour of adaptation data. Second, the observed gains can be improved further with the state-level Minimum Bayes Risk (sMBR) based sequence training technique. In addition to supervised training with limited amounts of data, we also study the effect of introducing unsupervised data at both the initial cross-entropy and subsequent sequence training stages. Our experiments show that unsupervised data helps with cross-entropy and sequence training criteria. Third, the effect of speaker diversity in the adaptation data is also investigated where our experiments show that although there can be large variance in final performance

NOTES

depending on the speakers selected, regularization is required to obtain significant gains. Overall, we demonstrate that with adaptation of neural network based acoustic models, we can obtain performance improvements of up to 24.8% relative.

Subspace LHUC for Fast Adaptation of Deep Neural Network Acoustic Models

Lahiru Samarakoon, Khe Chai Sim; NUS, Singapore

Sat-O-5-1-5, Time: 14:50

Recently, the learning hidden unit contributions (LHUC) method is proposed for the adaptation of deep neural network (DNN) based acoustic models for automatic speech recognition (ASR). In LHUC, a set of speaker dependent (SD) parameters is estimated to linearly recombine the hidden units in an unsupervised fashion. Although LHUC performs considerably well, the gains diminish when the availability of the adaptation data amount decreases. Moreover, the per-speaker footprint of LHUC adaptation is in thousands and it is not desirable. Therefore, in this work, we propose the subspace LHUC, where the SD parameters are estimated in a subspace and connected to various layers through a new set of adaptively trained weights. We evaluate the subspace LHUC in the Aurora4 and AMI IHM tasks. Experimental results show that the subspace LHUC outperforms standard LHUC adaptation. With utterance-level fast adaptation, the subspace LHUC achieved 11.3% and 4.5% relative improvements over the standard LHUC for the Aurora4 and AMI IHM tasks respectively. Furthermore, the subspace LHUC reduces the per-speaker footprint by 94% over the standard LHUC adaptation.

Improving Children's Speech Recognition Through Out-of-Domain Data Augmentation

Joachim Fainberg[1], Peter Bell[1], Mike Lincoln[2], Steve Renals[1]; [1]University of Edinburgh, UK; [2]Quorate Technology, UK

Sat-O-5-1-6, Time: 15:10

Children's speech poses challenges to speech recognition due to strong age-dependent anatomical variations and a lack of large, publicly-available corpora. In this paper we explore data augmentation for children's speech recognition using stochastic feature mapping (SFM) to transform out-of-domain adult data for both GMM-based and DNN-based acoustic models. We performed experiments on the English PF-STAR corpus, augmenting using WSJCAM0 and ABI. Our experimental results indicate that a DNN acoustic model for childrens speech can make use of adult data, and that out-of-domain SFM is more accurate than in-domain SFM.

Sat-O-5-2 : Special Session: Sharing Research and Education Resources for Understanding Speech Processing

Grand Ballroom BC, 13:30–15:30, Saturday, 10 Sept. 2016
Chairs: Eric Fosler-Lussier, Rebecca Bates, Florian Metze

Virtual Machines and Containers as a Platform for Experimentation

Florian Metze[1], Eric Riebling[1], Anne S. Warlaumont[2], Elika Bergelson[3]; [1]Carnegie Mellon University, USA; [2]University of California at Merced, USA; [3]University of Rochester, USA

Sat-O-5-2-1, Time: 13:30

Research on computational speech processing has traditionally relied on the availability of a relatively large and complex infrastructure, which encompasses data (text and audio), tools (feature extraction, model training, scoring, possibly on-line and off-line, etc.), glue code, and computing. Traditionally, it has been very hard to move experiments from one site to another, and to replicate experiments. With the increasing availability of shared platforms such as commercial cloud computing platforms or publicly funded supercomputing centers, there is a need and an opportunity to abstract the experimental environment from the hardware, and distribute complete setups as a virtual machine, a container, or some other shareable resource, that can be deployed and worked with anywhere.

In this paper, we discuss our experience with this concept and present some tools that the community might find useful. We outline, as a case study, how such tools can be applied to a naturalistic language acquisition audio corpus.

CloudCAST — Remote Speech Technology for Speech Professionals

Phil Green[1], Ricard Marxer[1], Stuart Cunningham[1], Heidi Christensen[1], Frank Rudzicz[2], Maria Yancheva[3], André Coy[4], Massimiliano Malavasi[5], Lorenzo Desideri[5], Fabio Tamburini[6]; [1]University of Sheffield, UK; [2]Toronto Rehabilitation Institute, Canada; [3]University of Toronto, Canada; [4]University of West Indies, Jamaica; [5]AIAS Onlus Bologna, Italy; [6]Università di Bologna, Italy

Sat-O-5-2-2, Time: 13:45

Recent advances in speech technology are potentially of great benefit to the professionals who help people with speech problems: therapists, pathologists, educators and clinicians. There are 3 obstacles to progress which we seek to address in the CloudCAST project: • the design of applications deploying the technology should be user-driven; • the computing resource should be available remotely; • the software should be capable of personalisation: clinical applications demand individual solutions.

CloudCAST aims to provide such a resource, and in addition to gather the data produced as the applications are used, to underpin the machine learning required for further progress.

Notes

webASR 2 — Improved Cloud Based Speech Technology

Thomas Hain, Jeremy Christian, Oscar Saz, Salil Deena, Madina Hasan, Raymond W.M. Ng, Rosanna Milner, Mortaza Doulaty, Yulan Liu; University of Sheffield, UK

Sat-O-5-2-3, Time: 14:00

This paper presents the most recent developments of the webASR service (www.webasr.org), the world's first web-based fully functioning automatic speech recognition platform for scientific use. Initially released in 2008, the functionalities of webASR have recently been expanded with 3 main goals in mind: Facilitate access through a RESTful architecture, that allows for easy use through either the web interface or an API; allow the use of input metadata when available by the user to improve system performance; and increase the coverage of available systems beyond speech recognition. Several new systems for transcription, diarisation, lightly supervised alignment and translation are currently available through webASR. The results in a series of well-known benchmarks (RT'09, IWSLT'12 and MGB'15 evaluations) show how these webASR systems provides state-of-the-art performances across these tasks.

Sharing Speech Synthesis Software for Research and Education Within Low-Tech and Low-Resource Communities

Andrew R. Plummer, Mary E. Beckman; Ohio State University, USA

Sat-O-5-2-4, Time: 14:15

Parametric speech synthesis has played an integral role in speech research since the 1950s. However, software sharing is unwieldy, making replication of experiments difficult, creating obstacles to communication between laboratories, and hindering entry into research. This paper describes our use of the Speech Recognition Virtual Kitchen environment (www.speechkitchen.org) to develop an infrastructure for sharing synthesis software for research and education. We tested the infrastructure by using it in teaching a seminar on "the speech science of speech synthesis" to students from several of the graduate programs in linguistics at the Ohio State University. Using the virtual machines that we developed for Klatt's formant synthesis program and Kawahara's STRAIGHT speech analysis, modification, and synthesis system enabled the students to advance much further in their understanding of the basic principles underlying these acoustic-domain models by comparison to the students enrolled in a similar seminar that we taught previously without the virtual machines. At the same time, implementing these and two other virtual machines for the course did not live up to our expectations for the course, in ways that highlight the need to adapt both the Speech Kitchen environment and the synthesis software systems to the needs of low-tech, low-resource users.

The Berkeley Phonetics Machine

Ronald L. Sprouse, Keith Johnson; University of California at Berkeley, USA

Sat-O-5-2-5, Time: 14:30

The Berkeley Phonetics Machine is a Linux virtual machine image produced and used by the UC Berkeley Phonology Lab as a platform for phonetic research. It contains a full data analysis stack based on Python and R and also specialized tools for phonetic research. The machine is designed as a flexible and productive platform for established and novel research agendas that can be easily shared and reproduced. We list the software available in the machine, which includes many command-line tools for acoustic analysis and media file manipulation, as well as specialized Python libraries. We also discuss the use of this machine in the Phonology Lab and in phonetics courses. The overall experience with the machine has been positive, as faculty and graduate students are able to share and execute scripts in a common working environment. Undergraduate students have less opportunity to master the virtual machine environment but benefit from simplified instructions and fewer installation and operating problems. The primary difficulty that we have encountered has been with a few underpowered student computers that cannot run the virtual machine or do not run it well.

Experiences with Shared Resources for Research and Education in Speech and Language Processing

Rebecca Bates [1], Eric Fosler-Lussier [2], Florian Metze [3], Martha Larson [4], Gina-Anne Levow [5], Emily Mower Provost [6]; [1] Minnesota State University, USA; [2] Ohio State University, USA; [3] Carnegie Mellon University, USA; [4] Technische Universiteit Delft, The Netherlands; [5] University of Washington, USA; [6] University of Michigan, USA

Sat-O-5-2-6, Time: 14:45

Resource barriers can prevent capable researchers from participating in the speech and language community and can make it difficult to support learning and participation in our field at a wide variety of institutions. Sharing resources, whether software, processed data, experimental methodologies or virtual machines, can reduce the barrier to entry and potentially broaden participation in speech and language research and improve workforce development. As an introduction to the special session on Sharing Research and Education Resources for Understanding Speech Processing, we outline current trends and requirements for expanding participation in speech processing research. A qualitative research approach was used. Faculty at a variety of institutions have been interviewed and have participated in reflection writing about needs, tools, challenges, and successes. Themes from reflections were generated using a grounded theory approach and were used to code interviews for related evidence. This paper describes the educational and research challenges experienced by faculty as users of resources, rather than the details of specific resources provided. The goal is to engage in a stronger dialog between users and providers so that needs and resources are better aligned. A case study of a shared resource used at several universities highlights this dialog.

Panel and Audience Discussion: How do we Develop, Disseminate, and Sustain Shared Resources from User and Developer Perspectives?

Sat-O-5-2-7, Time: 15:00

(No abstract available at the time of publication)

NOTES

Sat-O-5-3 : Special Session: Voice Conversion Challenge

Bayview A, 13:30–15:30, Saturday, 10 Sept. 2016

Chairs: Tomoki Toda, Junichi Yamagishi, Fernando Villavicencio, Zhizheng Wu, Ling-Hui Chen, Daisuke Saito, Mirjam Wester

The Voice Conversion Challenge 2016

Tomoki Toda[1], Ling-Hui Chen[2], Daisuke Saito[3], Fernando Villavicencio[4], Mirjam Wester[5], Zhizheng Wu[5], Junichi Yamagishi[4]; [1]Nagoya University, Japan; [2]USTC, China; [3]University of Tokyo, Japan; [4]NII, Japan; [5]University of Edinburgh, UK

Sat-O-5-3-1, Time: 13:30

This paper describes the Voice Conversion Challenge 2016 devised by the authors to better understand different voice conversion (VC) techniques by comparing their performance on a common dataset. The task of the challenge was speaker conversion, i.e., to transform the voice identity of a source speaker into that of a target speaker while preserving the linguistic content. Using a common dataset consisting of 162 utterances for training and 54 utterances for evaluation from each of 5 source and 5 target speakers, 17 groups working in VC around the world developed their own VC systems for every combination of the source and target speakers, i.e., 25 systems in total, and generated voice samples converted by the developed systems. These samples were evaluated in terms of target speaker similarity and naturalness by 200 listeners in a controlled environment. This paper summarizes the design of the challenge, its result, and a future plan to share views about unsolved problems and challenges faced by the current VC techniques.

Analysis of the Voice Conversion Challenge 2016 Evaluation Results

Mirjam Wester, Zhizheng Wu, Junichi Yamagishi; University of Edinburgh, UK

Sat-O-5-3-2, Time: 13:45

The Voice Conversion Challenge 2016 is the first Voice Conversion Challenge in which different voice conversion systems and approaches using the same voice data were compared. This paper describes the design of the evaluation, it presents the results and statistical analyses of the results.

The USTC System for Voice Conversion Challenge 2016: Neural Network Based Approaches for Spectrum, Aperiodicity and F_0 Conversion

Ling-Hui Chen[1], Li-Juan Liu[2], Zhen-Hua Ling[1], Yuan Jiang[2], Li-Rong Dai[1]; [1]USTC, China; [2]iFLYTEK, China

Sat-O-5-3-3, Time: 14:00

This paper introduces the methods we adopt to build our system for the evaluation event of Voice Conversion Challenge (VCC) 2016. We propose to use neural network-based approaches to convert both spectral and excitation features. First, the generatively trained deep neural network (GTDNN) is adopted for spectral envelope conversion after the spectral envelopes have been pre-processed by frequency warping. Second, we propose to use a recurrent neural network (RNN) with long short-term memory (LSTM) cells for F0 trajectory conversion. In addition, we adopt a DNN for band aperiodicity conversion. Both internal tests and formal VCC evaluation results demonstrate the effectiveness of the proposed methods.

A Voice Conversion Mapping Function Based on a Stacked Joint-Autoencoder

Seyed Hamidreza Mohammadi, Alexander Kain; Oregon Health & Science University, USA

Sat-O-5-3-4, Time: 14:15

In this study, we propose a novel method for training a regression function and apply it to a voice conversion task. The regression function is constructed using a Stacked Joint-Autoencoder (SJAE). Previously, we have used a more primitive version of this architecture for pre-training a Deep Neural Network (DNN). Using objective evaluation criteria, we show that the lower levels of the SJAE perform best with a low degree of jointness, and higher levels with a higher degree of jointness. We demonstrate that our proposed approach generates features that do not suffer from the averaging effect inherent in back-propagation training. We also carried out subjective listening experiments to evaluate speech quality and speaker similarity. Our results show that the SJAE approach has both higher quality and similarity than a SJAE+DNN approach, where the SJAE is used for pre-training a DNN, and the fine-tuned DNN is then used for mapping. We also present the system description and results of our submission to Voice Conversion Challenge 2016.

Locally Linear Embedding for Exemplar-Based Spectral Conversion

Yi-Chiao Wu, Hsin-Te Hwang, Chin-Cheng Hsu, Yu Tsao, Hsin-Min Wang; Academia Sinica, Taiwan

Sat-O-5-3-5, Time: 14:30

This paper describes a novel exemplar-based spectral conversion (SC) system developed by the AST (Academia Sinica, Taipei) team for the 2016 voice conversion challenge (vcc2016). The key feature of our system is that it integrates the locally linear embedding (LLE) algorithm, a manifold learning algorithm that has been successfully applied for the super-resolution task in image processing, with the conventional exemplar-based SC method. To further improve the quality of the converted speech, our system also incorporates (1) the maximum likelihood parameter generation (MLPG) algorithm, (2) the postfiltering-based global variance (GV) compensation method, and (3) a high-resolution feature extraction process. The results of subjective evaluation conducted by the vcc2016 organizer show that our LLE-exemplar-based SC system notably outperforms the baseline GMM-based system (implemented by the vcc2016 organizer). Moreover, our own internal evaluation results confirm the effectiveness of the major LLE-exemplar-based SC method and the three additional approaches with improved speech quality.

Applying Spectral Normalisation and Efficient Envelope Estimation and Statistical Transformation for the Voice Conversion Challenge 2016

Fernando Villavicencio[1], Junichi Yamagishi[1], Jordi Bonada[2], Felipe Espic[3]; [1]NII, Japan; [2]Universitat Pompeu Fabra, Spain; [3]University of Edinburgh, UK

Sat-O-5-3-6, Time: 14:45

In this work we present our entry for the Voice Conversion Challenge 2016, denoting new features to previous work on GMM-based

NOTES

voice conversion. We incorporate frequency warping and pitch transposition strategies to perform a normalisation of the spectral conditions, with benefits confirmed by objective and perceptual means. Moreover, the results of the challenge showed our entry among the highest performing systems in terms of perceived naturalness while maintaining the target similarity performance of GMM-based conversion.

ML Parameter Generation with a Reformulated MGE Training Criterion — Participation in the Voice Conversion Challenge 2016

D. Erro, A. Alonso, L. Serrano, D. Tavarez, I. Odriozola, Xabier Sarasola, Eder del Blanco, J. Sanchez, I. Saratxaga, Eva Navas, Inma Hernaez; Universidad del País Vasco, Spain

Sat-O-5-3-7, Time: 15:00

This paper describes our entry to the Voice Conversion Challenge 2016. Based on the maximum likelihood parameter generation algorithm, the method is a reformulation of the minimum generation error training criterion. It uses a GMM for soft classification, a Mel-cepstral vocoder for acoustic analysis and an improved dynamic time warping procedure for source-target alignment. To compensate the oversmoothing effect, the generated parameters are filtered through a speaker-independent postfilter implemented as a linear transform in cepstral domain. The process is completed with mean and variance adaptation of the log- fundamental frequency and duration modification by a constant factor. The results of the evaluation show that the proposed system achieves a high conversion accuracy in comparison with other systems, while its naturalness scores are intermediate.

The NU-NAIST Voice Conversion System for the Voice Conversion Challenge 2016

Kazuhiro Kobayashi[1], Shinnosuke Takamichi[1], Satoshi Nakamura[1], Tomoki Toda[2]; [1]NAIST, Japan; [2]Nagoya University, Japan

Sat-O-5-3-8, Time: 15:15

This paper presents the NU-NAIST voice conversion (VC) system for the Voice Conversion Challenge 2016 (VCC 2016) developed by a joint team of Nagoya University and Nara Institute of Science and Technology. Statistical VC based on a Gaussian mixture model makes it possible to convert speaker identity of a source speaker' voice into that of a target speaker by converting several speech parameters. However, various factors such as parameterization errors and over-smoothing effects usually cause speech quality degradation of the converted voice. To address this issue, we have proposed a direct waveform modification technique based on spectral differential filtering and have successfully applied it to singing voice conversion where excitation features are not necessary converted. In this paper, we propose a method to apply this technique to a standard voice conversion task where excitation feature conversion is needed. The result of VCC 2016 demonstrates that the NU-NAIST VC system developed by the proposed method yields the best conversion accuracy for speaker identity (more than 70% of the correct rate) and quite high naturalness score (more than 3 of the mean opinion score). This paper presents detail descriptions of the NU-NAIST VC system and additional results of its performance evaluation.

Sat-O-5-4 : Intelligibility and Masking

Bayview B, 13:30–15:30, Saturday, 10 Sept. 2016
Chairs: Fei (Felix) Chen, Tan Lee

Release from Energetic Masking Caused by Repeated Patterns of Glimpsing Windows

Maury Lander-Portnoy; University of Southern California, USA

Sat-O-5-4-1, Time: 13:30

The study of auditory masking not only provides data for how healthy and impaired listeners perform in adverse listening conditions, and thereby approximates their ability to perceive speech in the noisy environments of everyday life, but also provides insights into the mechanisms that underly the detection and perception of speech. Previous studies, (Pollack 1955) (Festen & Plomp 1990) (Cooper et al. 2015), have manipulated noise maskers in an attempt to observe the relationship between modulation of the type or characteristics of masking noise to subjects ability to detect or recognize a target signal. In this experiment, long term average spectrum speech shaped noise maskers were modulated to allow either short or long glimpsing (Cooke 2005) windows, during which the target signal was unmasked, in one second long morse code patterns of eight windows. The results from 60 participants with normal hearing showed that subjects performed significantly better on trials of an open set word recognition task when the pattern of glimpsing windows repeated twice before presentation of the masked signal than a control with the same glimpsing windows during the signal but different beforehand and one with the same amount of noise masking in random patterns before and during the target.

Glimpsing Predictions for Natural and Vocoded Sentence Intelligibility During Modulation Masking: Effect of the Glimpse Cutoff Criterion

Bobby Gibbs II, Daniel Fogerty; University of South Carolina, USA

Sat-O-5-4-2, Time: 13:50

This study varied the signal-to-noise ratio (SNR) cutoff criterion for acoustically defining usable perceptual glimpses that contribute to speech intelligibility. Criterion-dependent effects were determined by examining the correlation of three different acoustic glimpse metrics with intelligibility. Glimpse properties change depending on the acoustic interactions between the speech and competing noise. Therefore, these measures were investigated with different rates of competing speech that were varied using time compression or expansion. Finally, effects of temporal modulation masking and spectral segregation were examined by comparison between unprocessed (natural) and vocoded speech. Results revealed a range of SNR cutoffs that were associated with correlations between the different acoustic glimpse metrics and intelligibility. Changing the glimpse criterion strongly influenced the associations between intelligibility and two of the acoustic glimpse metrics for the different masker modulation rates. However, the proportion of target speech above the SNR cutoff was less affected by altering the cutoff criterion. These results suggest that intelligibility models should account for the perceptual contribution of different glimpse metrics or limit glimpse cutoff criteria to an SNR region (1–3 dB based on this data) that captures the perceptual utility of multiple glimpse mechanisms.

Notes

Temporal Envelopes in Sine-Wave Speech Recognition

Li Xu; Ohio University, USA

Sat-O-5-4-3, Time: 14:10

There is a long debate on the relative importance of spectral and temporal cues in speech perception theories. On the one hand, the highly-intelligible sine-wave speech (SWS) has been viewed as a representation of the global spectral structure of the speech signal. On the other hand, there is accumulating evidence showing that the temporal aspects of speech without spectral details provide sufficient speech understanding. The present study explored whether the temporal envelopes imbedded in the SWS contribute to its intelligibility. In the experiments, both SWS and natural speech signals were processed with noise and tone vocoders to remove the spectral details but to preserve the temporal envelopes. Twenty-two normal-hearing, native English-speaking adult listeners participated in sentence recognition tasks. Speech recognition performance of vocoder-processed SWS was slightly inferior to that of vocoder-processed natural speech but both reached plateau performance at 6–8 channels. Acoustic analysis further indicated that the temporal envelopes of the SWS were almost identical to those of the natural speech, with a mean correlation coefficient r = 0.949 across all sentences. The results provide strong evidence that the SWS represents both spectral and temporal structures of the speech and that the temporal envelopes imbedded in SWS carry important information for speech recognition.

Understanding Periodically Interrupted Mandarin Speech

Jing Liu [1], Rosanna H.N. Tong [2], Fei Chen [1]; [1] SUSTC, China; [2] University of Hong Kong, China

Sat-O-5-4-4, Time: 14:30

This study investigated the effects of two parameters (i.e., interruption rate, and duty cycle of interruption) on the perception of periodically interrupted Mandarin speech. Normal-hearing listeners were instructed to identify consonant/vowel/tone/word from isolated Mandarin words and recognize Mandarin sentences when they were temporally interrupted by square wave. Results showed that consistent with earlier findings obtained with English speech, interruption with a large rate or duty cycle favored the perception of periodically interrupted Mandarin speech. In addition, for isolated Mandarin word, the perception of vowel or tone was less affected by periodical interruption than that of consonant, and under periodical interruption the perception of consonant could largely account for the recognition of Mandarin word. For Mandarin sentence, the tonal characteristics and the simpler syllable structure in Mandarin might facilitate spectral-temporal integration of the target words, which contributed to a sentence intelligibility advantage of Mandarin over English under interrupted conditions.

Factors Affecting the Intelligibility of Sine-Wave Speech

Fei Chen [1], Daniel Fogerty [2]; [1] SUSTC, China; [2] University of South Carolina, USA

Sat-O-5-4-5, Time: 14:50

Studies on sine-wave speech (SWS) perception suggest that formants contain sufficient information for sentence intelligibility. This study further investigated the effects of amplitude modulation, number of sine-waves, and vowel resonance in SWS recognition. Results showed that Mandarin sentences synthesized using frequency trajectories of the first two formants were highly intelligible with additional contributions from formant amplitude modulation. However, amplitude modulation significantly contributed to intelligibility when only the vowels were preserved. The present work demonstrates that the intelligibility of Mandarin SWS can be largely attributed to the frequency transition of the first two formants and is susceptible to temporal interruption.

Effects of Urgent Speech and Preceding Sounds on Speech Intelligibility in Noisy and Reverberant Environments

Nao Hodoshima; Tokai University, Japan

Sat-O-5-4-6, Time: 15:10

Public-address (PA) announcements are used to convey emergency information; however, noise and reverberation sometimes make announcements in public spaces unintelligible. Therefore, the present study investigated how combinations of speech spoken in an urgent style and preceding sounds affect speech intelligibility and perceived urgency in noisy and reverberant environments. Sentences were spoken in normal or urgent styles and preceded by either two sounds (siren sound or ocean wave-like sound) or no sounds. Eighteen young participants carried out word identification test and rated perceived urgency on five-point scales in noisy and reverberant environments. The results showed that the urgently spoken speech had significantly higher speech intelligibility than the normal speech. The urgently spoken speech preceded by the wave-like sound showed significantly higher speech intelligibility than normal speech without sounds, normal speech preceded by the siren sound, and urgently spoken speech preceded by the siren sound. The results also demonstrated that the perceived urgency was rated higher for the urgently spoken speech than that for the normal speech, regardless of the types of preceding sounds. These results suggest that appropriate combinations of speaking styles and alerting sounds will increase the intelligibility of emergency PA announcements.

Sat-O-5-5 : Robust Speaker Recognition and Anti-Spoofing

Seacliff BCD, 13:30–15:30, Saturday, 10 Sept. 2016

Chairs: Tomi Kinnunen, Nicholas Evans

Integrated Spoofing Countermeasures and Automatic Speaker Verification: An Evaluation on ASVspoof 2015

Md. Sahidullah [1], Héctor Delgado [2], Massimiliano Todisco [2], Hong Yu [3], Tomi Kinnunen [1], Nicholas Evans [2], Zheng-Hua Tan [3]; [1] University of Eastern Finland, Finland; [2] EURECOM, France; [3] Aalborg University, Denmark

Sat-O-5-5-1, Time: 13:30

It is well known that automatic speaker verification (ASV) systems can be vulnerable to spoofing. The community has responded to the threat by developing dedicated countermeasures aimed at detecting spoofing attacks. Progress in this area has accelerated over recent years, partly as a result of the first standard evaluation,

Notes

ASVspoof 2015, which focused on spoofing detection in isolation from ASV. This paper investigates the integration of state-of-the-art spoofing countermeasures in combination with ASV. Two general strategies to countermeasure integration are reported: cascaded and parallel. The paper reports the first comparative evaluation of each approach performed with the ASVspoof 2015 corpus. Results indicate that, even in the case of varying spoofing attack algorithms, ASV performance remains robust when protected with a diverse set of integrated countermeasures.

Cross-Database Evaluation of Audio-Based Spoofing Detection Systems

Pavel Korshunov, Sébastien Marcel; Idiap Research Institute, Switzerland

Sat-O-5-5-2, Time: 13:50

Since automatic speaker verification (ASV) systems are highly vulnerable to spoofing attacks, it is important to develop mechanisms that can detect such attacks. To be practical, however, a spoofing attack detection approach should have (i) high accuracy, (ii) be well-generalized for practical attacks, and (iii) be simple and efficient. Several audio-based spoofing detection methods have been proposed recently but their evaluation is limited to less realistic databases containing homogeneous data. In this paper, we consider eight existing presentation attack detection (PAD) methods and evaluate their performance using two major publicly available speaker databases with spoofing attacks: AVspoof and ASVspoof. We first show that realistic presentation attacks (speech is replayed to PAD system) are significantly more challenging for the considered PAD methods compared to the so called 'logical access' attacks (speech is presented to PAD system directly). Then, via a cross-database evaluation, we demonstrate that the existing methods generalize poorly when different databases or different types of attacks are used for training and testing. The results question the efficiency and practicality of the existing PAD systems, as well as, call for creation of databases with larger variety of realistic speech presentation attacks.

Investigation of Sub-Band Discriminative Information Between Spoofed and Genuine Speech

Kaavya Sriskandaraja, Vidhyasaharan Sethu, Phu Ngoc Le, Eliathamby Ambikairajah; University of New South Wales, Australia

Sat-O-5-5-3, Time: 14:10

A speaker verification system should include effective precautions against malicious spoofing attacks, and although some initial countermeasures have been recently proposed, this remains a challenging research problem. This paper investigates discrimination between spoofed and genuine speech, as a function of frequency bands, across the speech bandwidth. Findings from our investigation inform some proposed filter bank design approaches for discrimination of spoofed speech. Experiments are conducted on the Spoofing and Anti-Spoofing (SAS) corpus using the proposed frequency-selective approach demonstrates an 11% relative improvement in terms of equal error rate compared with a conventional mel filter bank.

An Investigation of Spoofing Speech Detection Under Additive Noise and Reverberant Conditions

Xiaohai Tian[1], Zhizheng Wu[2], Xiong Xiao[3], Eng Siong Chng[1], Haizhou Li[1]; [1]NTU, Singapore; [2]University of Edinburgh, UK; [3]TL@NTU, Singapore

Sat-O-5-5-4, Time: 14:30

Spoofing detection for automatic speaker verification (ASV), which is to discriminate between live and artificial speech, has received increasing attentions recently. However, the previous studies have been done on the clean data without significant noise. It is still not clear whether the spoofing detectors trained on clean speech can generalise well under noisy conditions. In this work, we perform an investigation of spoofing detection under additive noise and reverberant conditions. In particular, we consider five difference additive noises at three different signal-to-noise ratios (SNR), and a reverberation noise with different reverberation time (RT). Our experimental results reveal that additive noises degrade the spoofing detectors trained on clean speech significantly. However, the reverberation does not hurt the performance too much.

Robust Speaker Recognition with Combined Use of Acoustic and Throat Microphone Speech

Md. Sahidullah[1], Rosa Gonzalez Hautamäki[1], Dennis Alexander Lehmann Thomsen[2], Tomi Kinnunen[1], Zheng-Hua Tan[2], Ville Hautamäki[1], Robert Parts[3], Martti Pitkänen[3]; [1]University of Eastern Finland, Finland; [2]Aalborg University, Denmark; [3]Aplcomp, Finland

Sat-O-5-5-5, Time: 14:50

Accuracy of automatic speaker recognition (ASV) systems degrades severely in the presence of background noise. In this paper, we study the use of additional side information provided by a body-conducted sensor, throat microphone. Throat microphone signal is much less affected by background noise in comparison to acoustic microphone signal. This makes throat microphones potentially useful for feature extraction or speech activity detection. This paper, firstly, proposes a new prototype system for simultaneous data-acquisition of acoustic and throat microphone signals. Secondly, we study the use of this additional information for both speech activity detection, feature extraction and fusion of the acoustic and throat microphone signals. We collect a pilot database consisting of 38 subjects including both clean and noisy sessions. We carry out speaker verification experiments using Gaussian mixture model with universal background model (GMM-UBM) and i-vector based system. We have achieved considerable improvement in recognition accuracy even in highly degraded conditions.

Statistical Modeling of Speaker's Voice with Temporal Co-Location for Active Voice Authentication

Zhong Meng, Biing-Hwang Juang; Georgia Institute of Technology, USA

Sat-O-5-5-6, Time: 15:10

Active voice authentication (AVA) is a new mode of talker authentication, in which the authentication is performed continuously on very short segments of the voice signal, which may have instantaneously undergone change of talker. AVA is necessary in providing real-time monitoring of a device authorized for a particular user. The authentication test thus cannot rely on a long history of the voice data nor any past decisions. Most conventional voice authentication techniques that operate on the assumption that the entire test utterance is from only one talker with a claimed identity (including i-vector) fail to meet this stringent requirement. This paper presents a different signal modeling technique, within a conditional vector-quantization framework and with matching short-time statistics that take into account the co-located speech codes to meet the new challenge. As one variation, the temporally co-located VQ (TC-VQ) associates each codeword with a set of Gaussian mixture models to account for the co-located distributions and a temporally co-located hidden Markov model (TC-HMM) is built upon the TC-VQ. The proposed technique achieves an window-based equal error rate in the range of 3–5% and a relative gain of 4–25% over a baseline system using traditional HMMs on the AVA database.

Sat-O-5-6 : Speech Enhancement and Applications

Seacliff A, 13:30–15:30, Saturday, 10 Sept. 2016
Chairs: Horacio Franco, Omid Sadjadi

Joint Enhancement and Coding of Speech by Incorporating Wiener Filtering in a CELP Codec

Johannes Fischer, Tom Bäckström; FAU Erlangen-Nürnberg, Germany

Sat-O-5-6-1, Time: 13:30

The performance of speech communication applications in the field of mobile devices is often hampered by background noises and distortions. Therefore, noise attenuation methods are commonly used as a pre-processing method, cascaded with the speech-codec. We demonstrate that the performance of such combinations of speech enhancement and coding methods can be improved by joining the two methods into a single block. The proposed method is based on incorporating Wiener filtering into the objective function used for optimization of the quantization in code excited linear prediction (CELP)-based codecs. The benefits are that 1) the non-linear components of CELP codecs, including quantization and error feedback, are taken into account in the joint minimization function thereby improving quality and 2) by merging blocks both delay and computational complexity can be minimized. Our experiments demonstrate that the proposed joint enhancement and coding approach consistently improves subjective and objective quality. The proposed method is compatible with any CELP-based codecs without changing the bit-stream, whereby it can be readily applied in mobile phones or speech communication devices applying the concepts of CELP codecs for improving perceptual quality in adverse conditions.

Multi-Channel Linear Prediction Based on Binaural Coherence for Speech Dereverberation

Hong Liu, Xiuling Wang, Miao Sun, Cheng Pang; Peking University, China

Sat-O-5-6-2, Time: 13:50

It has been shown that the multi-channel linear prediction (MCLP) can achieve blind speech dereverberation effectively. However, it always degrades the binaural cues which are exploited for human sound localization, i.e., interaural time differences (ITD) and interaural level differences (ILD). To overcome this problem, the multiple input-single output structure of conventional MCLP is modified to a binaural input-output structure for suppressing reverberation and preserving binaural cues simultaneously. First, by employing a binaural coherence model with head shadowing effects, the variance of desired signal can be estimated the same to both ears, which can ensure no modification of ILD. Then, the variance is utilized to calculate the prediction coefficients in a maximum-likelihood (ML) sense. Finally, the desired signals can be obtained as the prediction errors in MCLP. And since the algorithm does not disturb the phase of input signal, the ITD cue is kept. Evaluations with measured binaural room impulse responses (BRIRs) show that the proposed method yields a good performance on both speech dereverberation and binaural cues preservation.

Single-Channel Speech Enhancement Using Double Spectrum

Martin Blass[1], Pejman Mowlaee[1], W. Bastiaan Kleijn[2]; [1]Technische Universität Graz, Austria; [2]Victoria University of Wellington, New Zealand

Sat-O-5-6-3, Time: 14:10

Single-channel speech enhancement is often formulated in the *Short-Time Fourier Transform* (STFT) domain. As an alternative, several previous studies have reported advantages of speech processing using pitch-synchronous analysis and filtering in the modulation transform domain. We propose to use the *Double Spectrum* (DS) obtained by combining pitch-synchronous transform followed by modulation transform. The linearity and sparseness properties of DS domain are beneficial for single-channel speech enhancement. The effectiveness of the proposed DS-based speech enhancement is demonstrated by comparing it with STFT-based and modulation-based benchmarks. In contrast to the benchmark methods, the proposed method does not exploit any statistical information nor does it use temporal smoothing. The proposed method leads to an improvement of 0.3 PESQ on average for babble noise.

On the Appropriateness of Complex-Valued Neural Networks for Speech Enhancement

Lukas Drude[1], Bhiksha Raj[2], Reinhold Haeb-Umbach[1]; [1]Universität Paderborn, Germany; [2]Carnegie Mellon University, USA

Sat-O-5-6-4, Time: 14:30

Although complex-valued neural networks (CVNNs) — networks which can operate with complex arithmetic — have been around for a while, they have not been given reconsideration since the breakthrough of deep network architectures. This paper presents a critical assessment whether the novel tool set of deep neural networks (DNNs) should be extended to complex-valued arithmetic. Indeed, with DNNs making inroads in speech enhancement tasks, the use of complex-valued input data, specifically the short-time Fourier transform coefficients, is an obvious consideration. In particular when it comes to performing tasks that heavily rely on phase information, such as acoustic beamforming, complex-valued algorithms are omnipresent. In this contribution we recapitulate backpropagation in CVNNs, develop complex-valued network elements, such as the split-rectified non-linearity, and compare real- and complex-valued networks on a beamforming task. We find that

NOTES

CVNNs hardly provide a performance gain and conclude that the effort of developing the complex-valued counterparts of the building blocks of modern deep or recurrent neural networks can hardly be justified.

Introducing the Turbo-Twin-HMM for Audio-Visual Speech Enhancement

Steffen Zeiler [1], Hendrik Meutzner [1], Ahmed Hussen Abdelaziz [2], Dorothea Kolossa [1]; [1] Ruhr-Universität Bochum, Germany; [2] ICSI, USA

Sat-O-5-6-5, Time: 14:50

Models for automatic speech recognition (ASR) hold detailed information about spectral and spectro-temporal characteristics of clean speech signals. Using these models for speech enhancement is desirable and has been the target of past research efforts. In such model-based speech enhancement systems, a powerful ASR is imperative. To increase the recognition rates especially in low-SNR conditions, we suggest the use of the additional visual modality, which is mostly unaffected by degradations in the acoustic channel. An optimal integration of acoustic and visual information is achievable by joint inference in both modalities within the turbo-decoding framework. Thus combining turbo-decoding with Twin-HMMs for speech enhancement, notable improvements can be achieved, not only in terms of instrumental estimates of speech quality, but also in actual speech intelligibility. This is verified through listening tests, which show that in highly challenging noise conditions, average human recognition accuracy can be improved from 64% without signal processing to 80% when using the presented architecture.

Assessing Speech Quality in Speech-Aware Hearing Aids Based on Phoneme Posteriorgrams

Constantin Spille [1], Hendrik Kayser [1], Hynek Hermansky [2], Bernd T. Meyer [2]; [1] Carl von Ossietzky Universität Oldenburg, Germany; [2] Johns Hopkins University, USA

Sat-O-5-6-6, Time: 15:10

Current behind-the-ear hearing aids (HA) allow to perform spatial filtering to enhance localized sound sources; however, they often lack processing strategies that are tailored to spoken language. Hence, without a feedback about speech quality achieved by the system, spatial filtering potentially remains unused, in case of a conservative enhancement strategy, or can even be detrimental to the speech intelligibility of the output signal. In this paper we apply phoneme posteriorgrams obtained from HA signals processed with deep neural networks to measure the quality of speech representations in spatial scenes. Inverse entropy of phoneme probabilities is proposed as a measure that allows to evaluate if current hearing aid parameters are optimal for the given acoustic condition. We investigate how varying noise levels and wrong estimates of the to-be-enhanced direction affect this measure in anechoic and reverberant conditions and show our measure to provide a high reliability when varying each parameter. Experiments show that entropy as a function of the beam angle has a distinct minimum at the speaker's true position and its immediate vicinity. Thus, it can be used to determine the beam angle which optimizes the speech representation. Further, variations of the SNR cause a consistent offset of the entropy.

Sat-P-5-1 : Speech Analysis

Pacific Concourse – Poster A, 13:30–15:30, Saturday, 10 Sept. 2016
Chair: Asterios Toutios

Time-Varying Quasi-Closed-Phase Weighted Linear Prediction Analysis of Speech for Accurate Formant Detection and Tracking

Dhananjaya Gowda, Paavo Alku; Aalto University, Finland

Sat-P-5-1-1, Time: 13:30

In this paper, we propose a new method for accurate detection, estimation and tracking of formants in speech signals using time-varying quasi-closed phase analysis (TVQCP). The proposed method combines two different methods of analysis namely, the time-varying linear prediction (TVLP) and quasi-closed phase (QCP) analysis. TVLP helps in better tracking of formant frequencies by imposing a time-continuity constraint on the linear prediction (LP) coefficients. QCP analysis, a type of weighted LP (WLP), improves the estimation accuracies of the formant frequencies by using a carefully designed weight function on the error signal that is minimized. The QCP weight function emphasizes the closed-phase region of the glottal cycle, and also weights down the regions around the main excitations. This results in reduced coupling of the subglottal cavity and the excitation source. Experimental results on natural speech signals show that the proposed method performs considerably better than the detect-and-track approach used in popular tools like Wavesurfer or Praat.

Improved Depiction of Tissue Boundaries in Vocal Tract Real-Time MRI Using Automatic Off-Resonance Correction

Yongwan Lim, Sajan Goud Lingala, Asterios Toutios, Shrikanth S. Narayanan, Krishna S. Nayak; University of Southern California, USA

Sat-P-5-1-2, Time: 13:30

Real-time magnetic resonance imaging (RT-MRI) is a powerful tool to study the dynamics of vocal tract shaping during speech production. The dynamic articulators of interest include the surfaces of the lips, tongue, hard palate, soft palate, and pharyngeal airway. All of these are located at air-tissue interfaces and are vulnerable to MRI off-resonance effect due to magnetic susceptibility. In RT-MRI using spiral or radial scanning, this appears as a signal loss or blurring in images and may impair the analysis of dynamic speech data. We apply an automatic off-resonance artifact correction method to speech RT-MRI data in order to enhance the sharpness of air-tissue boundaries. We demonstrate the improvement qualitatively and using an image sharpness metric offering an improved tool for speech science research.

Modeling and Transforming Speech Using Variational Autoencoders

Merlijn Blaauw, Jordi Bonada; Universitat Pompeu Fabra, Spain

Sat-P-5-1-3, Time: 13:30

Latent generative models can learn higher-level underlying factors from complex data in an unsupervised manner. Such models can be

Notes

used in a wide range of speech processing applications, including synthesis, transformation and classification. While there have been many advances in this field in recent years, the application of the resulting models to speech processing tasks is generally not explicitly considered. In this paper we apply the variational autoencoder (VAE) to the task of modeling frame-wise spectral envelopes. The VAE model has many attractive properties such as continuous latent variables, prior probability over these latent variables, a tractable lower bound on the marginal log likelihood, both generative and recognition models, and end-to-end training of deep models. We consider different aspects of training such models for speech data and compare them to more conventional models such as the Restricted Boltzmann Machine (RBM). While evaluating generative models is difficult, we try to obtain a balanced picture by considering both performance in terms of reconstruction error and when applying the model to a series of modeling and transformation tasks to get an idea of the quality of the learned features.

Phase-Encoded Speech Spectrograms

Chandra Sekhar Seelamantula; Indian Institute of Science, India

Sat-P-5-1-4, Time: 13:30

Spectrograms of speech and audio signals are time-frequency densities, and by construction, they are non-negative and do not have phase associated with them. Under certain conditions on the amount of overlap between consecutive frames and frequency sampling, it is possible to reconstruct the signal from the spectrogram. Deviating from this requirement, we develop a new technique to incorporate the phase of the signal in the spectrogram by satisfying what we call as the *delta dominance condition*, which in general is different from the well known minimum-phase condition. In fact, there are signals that are delta dominant but not minimum-phase and vice versa. The delta dominance condition can be satisfied in multiple ways, for example by placing a Kronecker impulse of the right amplitude or by choosing a suitable window function. A direct consequence of this novel way of constructing the spectrograms is that the phase of the signal is directly encoded or embedded in the spectrogram. We also develop a reconstruction methodology that takes such phase-encoded spectrograms and obtains the signal using the discrete Fourier transform (DFT). It is envisaged that the new class of phase-encoded spectrogram representations would find applications in various speech processing tasks such as analysis, synthesis, enhancement, and recognition.

Towards Minimally Invasive Velar State Detection in Normal and Silent Speech

Peter Birkholz[1], Petko Bakardjiev[1], Steffen Kürbis[1], Rico Petrick[2]; [1]Technische Universität Dresden, Germany; [2]Linguwerk, Germany

Sat-P-5-1-5, Time: 13:30

We present a portable minimally invasive system to determine the state of the velum (raised or lowered) at a sampling rate of 40 Hz that works both during normal and silent speech. The system consists of a small capsule containing a miniature loudspeaker and a miniature microphone. The capsule is inserted into one nostril by about 10 mm. The loudspeaker emits chirps with a power band from 12–24 kHz into the nostril and the microphone records the signal reflected from the nasal cavity. The chirp response differs between raised and lowered velar positions, because the velar position determines the shape of the nasal cavity in the posterior part and hence its acoustic behaviour. Reference chirp responses for raised and lowered velar positions in combination with a spectral distance measure are used to infer the state of the velum. Here we discuss critical design aspects of the system and outline future improvements. Possible applications of the device include the detection of the velar state during silent speech recognition, medical assessment of velar mobility and speech production research.

RNN-BLSTM Based Multi-Pitch Estimation

Jianshu Zhang, Jian Tang, Li-Rong Dai; USTC, China

Sat-P-5-1-6, Time: 13:30

Multi-pitch estimation is critical in many applications, including computational auditory scene analysis (CASA), speech enhancement/separation and mixed speech analysis; however, despite much effort, it remains a challenging problem. This paper uses the PEFAC algorithm to extract features and proposes the use of recurrent neural networks with bidirectional Long Short-Term Memory (RNN-BLSTM) to model the two pitch contours of a mixture of two speech signals. Compared with feed-forward deep neural networks (DNN), which are trained on static frame-level acoustic features, RNN-BLSTM is trained on sequential frame-level features and has more power to learn pitch contour temporal dynamics. The results of evaluations using a speech dataset containing mixtures of two speech signals demonstrate that RNN-BLSTM can substantially outperform DNN in multi-pitch estimation of mixed speech signals.

TUSK: A Framework for Overviewing the Performance of F0 Estimators

Masanori Morise[1], Hideki Kawahara[2]; [1]University of Yamanashi, Japan; [2]Wakayama University, Japan

Sat-P-5-1-7, Time: 13:30

This article presents a framework for overviewing the performance of fundamental frequency (F0) estimators and evaluates its effectiveness. Over the past few decades, many F0 estimators and evaluation indices have been proposed and have been evaluated using various speech databases. In speech analysis/ synthesis research, modern estimators are used as the algorithm to fulfill the demand for high-quality speech synthesis, but at the same time, they are competing with one another on minor issues. Specifically, while all of them meet the demands for high-quality speech synthesis, the result depends on the speech database used in the evaluation. Since there are various types of speech, it is inadvisable to discuss the effectiveness of each estimator on the basis of minor differences. It would be better to select the appropriate F0 estimator in accordance with the speech characteristics. The framework we propose, TUSK, does not rank the estimators but rather attempts to overview them. In TUSK, six parameters are introduced to observe the trends in the characteristics in each F0 estimator. The signal is artificially generated so that six parameters can be controllable independently. In this article, we introduce the concept of TUSK and determine its effectiveness using several modern F0 estimators.

NOTES

A Robust Non-Parametric and Filtering Based Approach for Glottal Closure Instant Detection

Pradeep Rengaswamy, Gurunath Reddy M., K. Sreenivasa Rao, Pallab Dasgupta; IIT Kharagpur, India

Sat-P-5-1-8, Time: 13:30

In this paper, a novel non-parametric based glottal closure instant (GCI) detection method after filtering the speech signal through a pulse shaping filter is proposed. The pulse shaping filter essentially de-emphasises the vocal tract resonances by emphasising the frequency components containing the pitch information. The filtered signal is subjected to non-linear processing to emphasise the GCI locations. The GCI locations are finally obtained by a non-parametric histograms based approach in the detected voiced regions from the filtered speech signal. The proposed method is compared with the two state-of-the-art epoch extraction methods : Zero frequency filtering (ZFF) and SEDREAMS (both of which requires upfront knowledge of average pitch period). The performance of the method is evaluated on the complete CMU-ARCTIC dataset consisting of both speech and Electroglottograph (EGG) signals. The robustness of the proposed method to the additive white noise is evaluated with several degradation levels. The experimental results showed that the proposed method is indeed immune to noise and the obtained results are comparably better than the two state-of-the-art methods.

Sat-P-5-2 : Speaker Recognition

Pacific Concourse – Poster B, 13:30–15:30, Saturday, 10 Sept. 2016
Chairs: Eliathamby Ambikairajah, Rahim Saeidi

Analysis of Face Mask Effect on Speaker Recognition

Rahim Saeidi, Ilkka Huhtakallio, Paavo Alku; Aalto University, Finland

Sat-P-5-2-1, Time: 13:30

Wearing a face mask affects the speech production. On top of that, the frequency response and radiation characteristics of the face mask — depending on the material and shape of the mask — adds to the complexity of analyzing speech under face mask. Our target is to separate the effect of muscle constriction and increased vocal effort in speech produced under face mask from sound transmission and radiation properties of face mask. In this paper, we measure up the far-field effects of wearing four different face masks; motorcycle helmet, rubber mask, surgical mask and scarf inside anechoic chamber. The measurement setup follows the recording configuration of a speech corpus used for speaker recognition experiments. In matching speech under face mask with speech under no mask, the frequency response of the respective face mask is accounted for and compensated for before acoustic feature extraction. The speaker recognition performance is reported using the state-of-the-art i-vector method for mismatched and compensated conditions in order to demonstrate the significance of knowing the type of mask and accounting for its sound transmission properties.

Data Selection for Within-Class Covariance Estimation

Elliot Singer[1], Tyler Campbell[2], Douglas Reynolds[1]; [1]MIT Lincoln Laboratory, USA; [2]Rensselaer Polytechnic Institute, USA

Sat-P-5-2-2, Time: 13:30

Methods for performing channel and session compensation in conjunction with subspace techniques have been a focus of considerable study recently and have led to significant gains in speaker recognition performance. While developers have typically exploited the vast archive of speaker labeled data available from earlier NIST evaluations to train the within-class and across-class covariance matrices required by these techniques, little attention has been paid to the characteristics of the data required to perform the training efficiently. This paper focuses on within-class covariance normalization (WCCN) and shows that a reduction in training data requirements can be achieved by proper data selection. In particular, it is shown that the key variables are the total amount of data and the degree of handset variability, with total calls per handset playing a smaller role. The study offers insight into efficient WCCN training data collection in real world applications.

Inter-Task System Fusion for Speaker Recognition

M. Ferras, Srikanth Madikeri, S. Dey, Petr Motlicek, Hervé Bourlard; Idiap Research Institute, Switzerland

Sat-P-5-2-3, Time: 13:30

Fusion is a common approach to improving the performance of speaker recognition systems. Multiple systems using different data, features or algorithms tend to bring complementary contributions to the final decisions being made. It is known that factors such as native language or accent contribute to speaker identity. In this paper, we explore inter-task fusion approaches to incorporating side information from accent and language identification systems to improve the performance of a speaker verification system. We explore both score level and model level approaches, linear logistic regression and linear discriminant analysis respectively, reporting significant gains on accented and multi-lingual data sets of the NIST Speaker Recognition Evaluation 2008 data. Equal error rate and expected rank metrics are reported for speaker verification and speaker identification tasks.

Mahalanobis Metric Scoring Learned from Weighted Pairwise Constraints in I-Vector Speaker Recognition System

Zhenchun Lei, Yanhong Wan, Jian Luo, Yingen Yang; Jiangxi Normal University, China

Sat-P-5-2-4, Time: 13:30

The i-vector model is widely used by the state-of-the-art speaker recognition system. We proposed a new Mahalanobis metric scoring learned from weighted pairwise constraints (WPCML), which use the different weights for the empirical error of the similar and dissimilar pairs. In the new i-vector space described by the metric, the distance between the same speaker's i-vectors is small, while that of the different speakers' is large. In forming the training set, we use the traditional way in random fashion and develop a new nearest distance based way. The results on the NIST 2008 telephone data shown that our model can get better performance than the classical cosine similarity scoring. When using the nearest distance based way to form the training set, our model is better than the state-of-the-art PLDA. And the results on the NIST 2014 i-vector challenge show that our model is also better than the PLDA.

NOTES

Novel Subband Autoencoder Features for Detection of Spoofed Speech

Meet H. Soni, Tanvina B. Patel, Hemant A. Patil; DA-IICT, India

Sat-P-5-2-5, Time: 13:30

Deep Neural Network (DNN) have been extensively used in Automatic Speech Recognition (ASR) applications. Very recently, DNNs have also found application in detecting natural vs. spoofed speech at ASV spoof challenge held at INTERSPEECH 2015. Along the similar lines, in this work, we propose a new feature extraction architecture of DNN called the subband autoencoder (SBAE) for spoof detection task. The SBAE is inspired by the human auditory system and extracts features from the speech spectrum in an unsupervised manner. The features derived from SBAE are compared with state-of-the-art Mel Frequency Cepstral Coefficient (MFCC) features. The experiments were performed on ASV spoof challenge database and the performance was evaluated using Equal Error Rate (EER). It was observed that on the evaluation set, MFCC features with 36-dimensional (static+Δ+$\Delta\Delta$) features gave 4.32% EER which reduced to 2.9% when 36-dimensional SBAE features were used. Further on fusing SBAE features at score-level with MFCC, a further reduction till 1.93% EER was observed. This improvement in EER was due to the fact that the dynamics of SBAE features captured significant spoof specific characteristics leading to detect significantly even vocoder-independent speech, which is not the case for MFCC.

On the Issue of Calibration in DNN-Based Speaker Recognition Systems

Mitchell McLaren[1], Diego Castan[1], Luciana Ferrer[2], Aaron Lawson[1]; [1]SRI International, USA; [2]Universidad de Buenos Aires, Argentina

Sat-P-5-2-6, Time: 13:30

This article is concerned with the issue of calibration in the context of Deep Neural Network (DNN) based approaches to speaker recognition. DNNs have provided a new standard in technology when used in place of the traditional universal background model (UBM) for feature alignment, or to augment traditional features with those extracted from a bottleneck layer of the DNN. These techniques provide extremely good performance for constrained trial conditions that are well matched to development conditions. However, when applied to unseen conditions or a wide variety of conditions, some DNN-based techniques offer poor calibration performance. Through analysis on both PRISM and the recently released Speakers in the Wild (SITW) corpora, we illustrate that bottleneck features hinder calibration if used in the calculation of first-order Baum Welch statistics during i-vector extraction. We propose a hybrid alignment framework, which stems from our previous work in DNN senone alignment, that uses the bottleneck features only for the alignment of features during statistics calculation. This framework not only addresses the issue of calibration, but provides a more computationally efficient system based on bottleneck features with improved discriminative power.

Probabilistic Approach Using Joint Long and Short Session i-Vectors Modeling to Deal with Short Utterances for Speaker Recognition

Waad Ben Kheder, Driss Matrouf, Moez Ajili, Jean-François Bonastre; LIA, France

Sat-P-5-2-7, Time: 13:30

Speaker recognition with short utterance is highly challenging. The use of i-vectors in SR systems became a standard in the last years and many algorithms were developed to deal with the short utterances problem. We present in this paper a new technique based on modeling jointly the i-vectors corresponding to short utterances and those of long utterances. The joint distribution is estimated using a large number of i-vectors pairs (coming from short and long utterances) corresponding to the same session. The obtained distribution is then integrated in an MMSE estimator in the test phase to compute an "improved" version of short utterance i-vectors. We show that this technique can be used to deal with duration mismatch and that it achieves up to 40% of relative improvement in EER(%) when used on NIST data. We also apply this technique on the recently published SITW database and show that it yields 25% of EER(%) improvement compared to a regular PLDA scoring.

Short Utterance Variance Modelling and Utterance Partitioning for PLDA Speaker Verification

Ahilan Kanagasundaram, David Dean, Sridha Sridharan, Clinton Fookes, Ivan Himawan; Queensland University of Technology, Australia

Sat-P-5-2-8, Time: 13:30

This paper analyses the short utterance probabilistic linear discriminant analysis (PLDA) speaker verification with utterance partitioning and short utterance variance (SUV) modelling approaches. Experimental studies have found that instead of using single long-utterance as enrolment data, if long enrolled-utterance is partitioned into multiple short utterances and average of short utterance i-vectors is used as enrolled data, that improves the Gaussian PLDA (GPLDA) speaker verification. This is because short utterance i-vectors have speaker, session and utterance variations, and utterance-partitioning approach compensates the utterance variation. Subsequently, SUV-PLDA is also studied with utterance partitioning approach, and utterance-partitioning-based SUV-GPLDA system shows relative improvement of 9% and 16% in EER for NIST 2008 and NIST 2010 truncated 10sec-10sec evaluation condition as utterance-partitioning approach compensates the utterance variation and SUV modelling approach compensates the mismatch between full-length development data and short-length evaluation data.

Speaker-Dependent Dictionary-Based Speech Enhancement for Text-Dependent Speaker Verification

Nicolai Bæk Thomsen, Dennis Alexander Lehmann Thomsen, Zheng-Hua Tan, Børge Lindberg, Søren Holdt Jensen; Aalborg University, Denmark

Sat-P-5-2-9, Time: 13:30

The problem of text-dependent speaker verification under noisy conditions is becoming ever more relevant, due to increased usage for authentication in real-world applications. Classical methods for

NOTES

noise reduction such as spectral subtraction and Wiener filtering introduce distortion and do not perform well in this setting. In this work we compare the performance of different noise reduction methods under different noise conditions in terms of speaker verification when the text is known and the system is trained on clean data (mis-matched conditions). We furthermore propose a new approach based on dictionary-based noise reduction and compare it to the baseline methods.

Text-Available Speaker Recognition System for Forensic Applications

Chengzhu Yu, Chunlei Zhang, Finnian Kelly, Abhijeet Sangwan, John H.L. Hansen; University of Texas at Dallas, USA

Sat-P-5-2-10, Time: 13:30

This paper examines a text-available speaker recognition approach targeting scenarios where the transcripts of test utterances are either available or obtainable through manual transcription. Forensic speaker recognition is one of such applications where the human supervision can be expected. In our study, we extend an existing Deep Neural Network (DNN) i-vector-based speaker recognition system to effectively incorporate text information associated with test utterances. We first show experimentally that speaker recognition performance drops significantly if the DNN output posteriors are directly replaced with their target *senone*, obtained from force alignment. The cause of such performance drops can be attributed to the fact that forced alignment selects only the single most probable *senone* as their output, which is not desirable in a current speaker recognition framework. To resolve this problem, we propose a posterior mapping approach where the relationship between forced aligned *senones* and its corresponding DNN posteriors are modeled. By replacing DNN output posteriors with *senone* mapped posteriors, a robust text-available speaker recognition system can be obtained in mismatched environments. Experiments using the proposed approach are performed on the Aurora-4 dataset.

Transfer Learning for Speaker Verification on Short Utterances

Qingyang Hong, Lin Li, Lihong Wan, Jun Zhang, Feng Tong; Xiamen University, China

Sat-P-5-2-11, Time: 13:30

Short utterance lacks enough discriminative information and its duration variation will propagate uncertainty into a probability linear discriminant analysis (PLDA) classifier. For speaker verification on short utterances, it can be considered as a domain with limited amount of long utterances. Therefore, transfer learning of PLDA can be adopted to learn discriminative information from other domain with a large amount of long utterances. In this paper, we explore the effectiveness of transfer learning based PLDA (TL-PLDA) on the NIST SRE and Switchboard (SWB) corpus. Experimental results showed that it could produce the largest gain of performance compared with the traditional PLDA, especially for short utterances with the duration of 5s and 10s.

Twin Model G-PLDA for Duration Mismatch Compensation in Text-Independent Speaker Verification

*Jianbo Ma[1], Vidhyasaharan Sethu[1], Eliathamby Ambikairajah[1], Kong Aik Lee[2]; [1]University of New South Wales, Australia; [2]A*STAR, Singapore*

Sat-P-5-2-12, Time: 13:30

Short duration speaker verification is a challenging problem partly due to utterance duration mismatch. This paper proposes a novel method that modifies the standard Gaussian probabilistic linear discriminant analysis (G-PLDA) to use two separate generative models for i-vectors from long and short utterances which are jointly trained. The proposed twin model G-PLDA employs distinct models for i-vectors corresponding to different durations from the same speaker but shares the same latent variables. Unlike the standard G-PLDA, this twin model G-PLDA takes the differences between utterances of varying durations into account. Hyper-parameter estimation and scoring formulae for the twin model G-PLDA are presented. Experimental results obtained using NIST 2010 data show that the proposed technique leads to relative improvements of 8.5% and 15.6% when tested on utterances of 5 second and 3 second durations respectively.

Universal Background Sparse Coding and Multilayer Bootstrap Network for Speaker Clustering

Xiao-Lei Zhang; Northwestern Polytechnical University, China

Sat-P-5-2-13, Time: 13:30

We apply multilayer bootstrap network (MBN) to speaker clustering. The proposed method first extracts supervectors by a universal background model, then reduces the dimension of the high-dimensional supervectors by MBN, and finally conducts speaker clustering by clustering the low-dimensional data. We also propose an MBN-based universal background model, named universal background sparse coding. The comparison results demonstrate the effectiveness and robustness of the proposed method.

Improving Deep Neural Networks Based Speaker Verification Using Unlabeled Data

Yao Tian[1], Meng Cai[2], Liang He[1], Wei-Qiang Zhang[1], Jia Liu[1]; [1]Tsinghua University, China; [2]Microsoft, China

Sat-P-5-2-14, Time: 13:30

Recently, deep neural networks (DNNs) trained to predict senones have been incorporated into the conventional i-vector based speaker verification systems to provide soft frame alignments and show promising results. However, the data mismatch problem may degrade the performance since the DNN requires transcribed data (out-domain data) while the data sets (in-domain data) used for i-vector training and extraction are mostly untranscribed. In this paper, we try to address this problem by exploiting the unlabeled in-domain data during the training of the DNN, hoping the DNN can provide a more robust basis for the in-domain data. In this work, we first explore the impact of using in-domain data during the unsupervised DNN pre-training process. In addition, we decode the in-domain data using a hybrid DNN-HMM system to get its transcription, and then we retrain the DNN model with the “labeled” in-domain data. Experimental results on the NIST SRE 2008 and

NOTES

the NIST SRE 2010 databases demonstrate the effectiveness of the proposed methods.

Sat-P-5-3 : Decoding, System Combination

Pacific Concourse – Poster C, 13:30–15:30, Saturday, 10 Sept. 2016
Chair: Alexey Karpov

Maximum a posteriori Based Decoding for CTC Acoustic Models

Naoyuki Kanda, Xugang Lu, Hisashi Kawai; NICT, Japan

Sat-P-5-3-1, Time: 13:30

This paper presents a novel decoding framework for connectionist temporal classification (CTC)-based acoustic models (AM). Although CTC-based AM inherently has the property of a language model (LM) in itself, an external LM trained with a large text corpus is still essential to obtain the best results. In the previous literatures, a naive interpolation of the CTC-based AM score and the external LM score was used, although there is no theoretical justification for it. In this paper, we propose a theoretically more sound decoding framework derived from a maximization of the posterior probability of a word sequence given an observation. In our decoding framework, a subword LM (SLM) is newly introduced to coordinate the CTC-based AM score and the word-level LM score. In experiments with the Wall Street Journal (WSJ) corpus and Corpus of Spontaneous Japanese (CSJ), our proposed framework consistently achieved improvements of 7.4–15.3% over the conventional interpolation-based framework. In the CSJ experiment, given 586 hours of training data, the CTC-based AM finally achieved a 6.7% better word error rate than the baseline method with deep neural networks and hidden Markov models.

Phonetic and Phonological Posterior Search Space Hashing Exploiting Class-Specific Sparsity Structures

Afsaneh Asaei, Gil Luyet, Milos Cernak, Hervé Bourlard; Idiap Research Institute, Switzerland

Sat-P-5-3-2, Time: 13:30

This paper shows that exemplar-based speech processing using class-conditional posterior probabilities admits a highly effective search strategy relying on posteriors' intrinsic sparsity structures. The posterior probabilities are estimated for phonetic and phonological classes using deep neural network (DNN) computational framework. Exploiting the class-specific sparsity leads to a simple quantized posterior hashing procedure to reduce the search space of posterior exemplars. To that end, small number of quantized posteriors are regarded as representatives of the posterior space and used as hash keys to index subsets of neighboring exemplars. The k nearest neighbor (kNN) method is applied for posterior based classification problems. The phonetic posterior probabilities are used as exemplars for phonetic classification whereas the phonological posteriors are used as exemplars for automatic prosodic event detection. Experimental results demonstrate that posterior hashing improves the efficiency of kNN classification drastically. This work encourages the use of posteriors as discriminative exemplars appropriate for large scale speech classification tasks.

Model Compression Applied to Small-Footprint Keyword Spotting

George Tucker[1], Minhua Wu[2], Ming Sun[2], Sankaran Panchapagesan[2], Gengshen Fu[2], Shiv Vitaladevuni[2]; [1]Google, USA; [2]Amazon.com, USA

Sat-P-5-3-3, Time: 13:30

Several consumer speech devices feature voice interfaces that perform on-device keyword spotting to initiate user interactions. Accurate on-device keyword spotting within a tight CPU budget is crucial for such devices. Motivated by this, we investigated two ways to improve deep neural network (DNN) acoustic models for keyword spotting without increasing CPU usage. First, we used low-rank weight matrices throughout the DNN. This allowed us to increase representational power by increasing the number of hidden nodes per layer without changing the total number of multiplications. Second, we used knowledge distilled from an ensemble of much larger DNNs used only during training. We systematically evaluated these two approaches on a massive corpus of far-field utterances. Alone both techniques improve performance and together they combine to give significant reductions in false alarms and misses without increasing CPU or memory usage.

Why do ASR Systems Despite Neural Nets Still Depend on Robust Features

Angel Mario Castro Martinez, Marc René Schädler; Carl von Ossietzky Universität Oldenburg, Germany

Sat-P-5-3-4, Time: 13:30

To which extent can neural nets learn traditional signal processing stages of current robust ASR front-ends? Will neural nets replace the classical, often auditory-inspired feature extraction in the near future? To answer these questions, a DNN-based ASR system was trained and tested on the Aurora4 robust ASR task using various (intermediate) processing stages. Additionally, the training set was divided into several fractions to reveal the amount of data needed to account for a missing processing step on the input signal or prior knowledge about the auditory system. The DNN system was able to learn from ordinary spectrograms representations outperforming MFCC using 75% of the training set and almost as good as log-Mel-spectrograms with the full set; on the other hand, it was unable to compensate the robustness of auditory-based Gabor features, which even using 40% of the training data outperformed every other representation. The study concludes that even with deep learning approaches, current ASR systems still benefit from a suitable feature extraction.

An Adaptive Multi-Band System for Low Power Voice Command Recognition

Qing He[1], Gregory W. Wornell[1], Wei Ma[2]; [1]MIT, USA; [2]Texas Instruments, USA

Sat-P-5-3-5, Time: 13:30

A complete voice-driven experience in applications such as wearable electronics requires always-on keyword monitoring, which is prohibitively power consuming using current speech recognition methods. In this work, we propose an ultra-low power voice command recognition system that is designed to recognize short commands such as 'Hi Galaxy'. To achieve power-efficient designs, the system uses adaptive feature pre-selection such that only a subset of all available features are selected and extracted based on

Notes

the noise spectrum. The back-end classifier, supporting adaptive feature selection, is enabled by a novel multi-band deep neural networks (DNNs) model that processes only the selected features at each decision. In experiments, our adaptive scheme achieves comparable accuracy and improved efficiency using an average of 5 spectral feature bands, than a generic fully-connected DNNs model using the full speech spectrum. The system makes a recognition decision every 40ms on 1.2s of buffered speech and consumes ~230μW of power, thus promising low-power, low-complexity and robust application-specific voice recognition.

Memory-Efficient Modeling and Search Techniques for Hardware ASR Decoders

Michael Price, Anantha Chandrakasan, James Glass; MIT, USA

Sat-P-5-3-6, Time: 13:30

This paper gives an overview of acoustic modeling and search techniques for low-power embedded ASR decoders. Our design decisions prioritize memory bandwidth, which is the main driver in system power consumption. We evaluate three acoustic modeling approaches — Gaussian mixture model (GMM), subspace GMM (SGMM) and deep neural network (DNN) — and identify trade-offs between memory bandwidth and recognition accuracy. We also present an HMM search scheme with WFST compression and caching, predictive beam width control, and a word lattice. Our results apply to embedded system implementations using microcontrollers, DSPs, FPGAs, or ASICs.

Log-Linear System Combination Using Structured Support Vector Machines

J. Yang, Anton Ragni, Mark J.F. Gales, Kate M. Knill; University of Cambridge, UK

Sat-P-5-3-7, Time: 13:30

Building high accuracy speech recognition systems with limited language resources is a highly challenging task. Although the use of multi-language data for acoustic models yields improvements, performance is often unsatisfactory with highly limited acoustic training data. In these situations, it is possible to consider using multiple well trained acoustic models and combine the system outputs together. Unfortunately, the computational cost associated with these approaches is high as multiple decoding runs are required. To address this problem, this paper examines schemes based on log-linear score combination. This has a number of advantages over standard combination schemes. Even with limited acoustic training data, it is possible to train, for example, phone-specific combination weights, allowing detailed relationships between the available well trained models to be obtained. To ensure robust parameter estimation, this paper casts log-linear score combination into a structured support vector machine (SSVM) learning task. This yields a method to train model parameters with good generalisation properties. Here the SSVM feature space is a set of scores from well-trained individual systems. The SSVM approach is compared to lattice rescoring and confusion network combination using language packs released within the IARPA Babel program.

Efficient Segmental Cascades for Speech Recognition

Hao Tang, Weiran Wang, Kevin Gimpel, Karen Livescu; TTIC, USA

Sat-P-5-3-8, Time: 13:30

Discriminative segmental models offer a way to incorporate flexible feature functions into speech recognition. However, their appeal has been limited by their computational requirements, due to the large number of possible segments to consider. Multi-pass cascades of segmental models introduce features of increasing complexity in different passes, where in each pass a segmental model rescores lattices produced by a previous (simpler) segmental model. In this paper, we explore several ways of making segmental cascades efficient and practical: reducing the feature set in the first pass, frame subsampling, and various pruning approaches. In experiments on phonetic recognition, we find that with a combination of such techniques, it is possible to maintain competitive performance while greatly reducing decoding, pruning, and training time.

A WFST Framework for Single-Pass Multi-Stream Decoding

Sirui Xu, Eric Fosler-Lussier; Ohio State University, USA

Sat-P-5-3-9, Time: 13:30

Combining disparate automatic speech recognition systems has long been an important strategy to improve recognition accuracy. Typically, each system requires a separate decoder; final results are derived by combining hypotheses from multiple lattices, necessitating multiple passes of decoding. We propose a novel Weighted Finite State Transducer (WFST) framework for integrating disparate systems. Our framework is different from the current popular system combination techniques in that the combination is done in one-pass decoding and allows the flexibility to combine systems at different levels of the decoding pipeline. Initial experiments with the framework achieved comparable performance as MBR-based combination which is reported to outperform ROVER and Confusion Network Combination (CNC). In this paper, we describe our methodology and present pilot study results for combining systems that use different sets of acoustic models, 1) gender-dependent GMM models, 2) MFCC and PLP features with GMM models, 3) MFCC, PLP and Filter Bank features with DNN models, and 4) SNR-specific DNN acoustic models. For each experiment, we also compared the computation time of the combined systems with their corresponding baseline systems. Our results show encouraging benefits of using the proposed framework to improve recognition performance while reducing computation time.

Comparison of Multiple System Combination Techniques for Keyword Spotting

William Hartmann, Le Zhang, Kerri Barnes, Roger Hsiao, Stavros Tsakalidis, Richard Schwartz; Raytheon BBN Technologies, USA

Sat-P-5-3-10, Time: 13:30

System combination is a common approach to improving results for both speech transcription and keyword spotting — especially in the context of low-resourced languages where building multiple complementary models requires less computational effort. Using state-of-the-art CNN and DNN acoustic models, we analyze the performance, cost, and trade-offs of four system combination approaches: feature combination, joint decoding, hitlist combination,

Notes

and a novel lattice combination method. Previous work has focused solely on accuracy comparisons. We show that joint decoding, lattice combination, and hitlist combination perform comparably, significantly better than feature combination. However, for practical systems, earlier combination reduces computational cost and storage requirements. Results are reported on four languages from the IARPA Babel dataset.

Rescoring by Combination of Posteriorgram Score and Subword-Matching Score for Use in Query-by-Example

Masato Obara[1], Kazunori Kojima[1], Kazuyo Tanaka[2], Shi-wook Lee[3], Yoshiaki Itoh[1]; [1]Iwate Prefectural University, Japan; [2]University of Tsukuba, Japan; [3]AIST, Japan

Sat-P-5-3-11, Time: 13:30

There has been much discussion recently regarding spoken term detection (STD) in speech processing communities. Query-by-Example (QbE) has also been an important topic in spoken-term detection (STD), where a query is issued using a speech signal. This paper proposes a rescoring method using a posteriorgram, which is a sequence of posterior probabilities obtained by a deep neural network (DNN) to be matched against both a speech signal of a query and spoken documents. Because direct matching between two posteriorgrams requires significant computation time, we first apply a conventional STD method that performs matching at a subword or state level, where the subword denotes an acoustic model, and the state composes a hidden Markov model of the acoustic model. Both the spoken query and the spoken documents are converted to subword sequences, using an automatic speech recognizer. After obtaining scores of candidates by subword/state matching, matching at the frame level using the posteriorgram is performed with continuous dynamic programming (CDP) verification for the top *N* candidates acquired by the subword/state matching. The score of the subword/state matching and the score of the posteriorgram matching are integrated and rescored, using a weighting coefficient. To reduce computation time, the proposed method is restricted to only top candidates for rescoring. Experiments for evaluation have been carried out using open test collections (Spoken-Doc tasks of NTCIR-10 workshops), and the results have demonstrated the effectiveness of the proposed method.

Phone Synchronous Decoding with CTC Lattice

Zhehuai Chen[1], Wei Deng[2], Tao Xu[2], Kai Yu[1]; [1]Shanghai Jiao Tong University, China; [2]AISpeech, China

Sat-P-5-3-12, Time: 13:30

Connectionist Temporal Classification (CTC) has recently shown improved efficiency in LVCSR decoding. One popular implementation is to use a CTC model to predict the phone posteriors at each frame which are then used for Viterbi beam search on a modified WFST network. This is still within the traditional frame synchronous decoding framework. In this paper, the peaky posterior property of a CTC model is carefully investigated and it is found that ignoring blank frames will not introduce additional search errors. Based on this phenomenon, a novel *phone synchronous* decoding framework is proposed. Here, a phone-level CTC lattice is constructed purely using the CTC acoustic model. The resultant CTC lattice is highly compact and removes tremendous search redundancy due to blank frames. Then, the CTC lattice can be composed with the standard WFST to yield the final decoding result. The proposed approach effectively separates the acoustic evidence calculation and the search operation. This not only significantly improves online search efficiency, but also allows flexible acoustic/linguistic resources to be used. Experiments on LVCSR tasks show that phone synchronous decoding can yield an extra 2–3 times speed up compared to the traditional frame synchronous CTC decoding implementation.

Sat-P-5-4 : Special Session: Clinical and Neuroscience-Inspired Vocal Biomarkers of Neurological and Psychiatric Disorders

Pacific Concourse – Poster D, 13:30–15:30, Saturday, 10 Sept. 2016

Chairs: Nicholas Cummins, Julien Epps, Emily Mower Provost, Thomas Quatieri, Stefan Scherer

Speech Features for Depression Detection

Saurabh Sahu, Carol Espy-Wilson; University of Maryland, USA

Sat-P-5-4-1, Time: 13:30

In this paper we discuss speech features that are useful in the detection of depression. Neuro-physiological changes associated with depression affect motor coordination and can disrupt articulatory precision in speech. We use the Mundt database and focus on six speakers in the database that transitioned between being depressed and not depressed based on their Hamilton depression scores. We quantify the degree of breathiness, jitter and shimmer computed from an AMDF based parameter. Measures from sustained vowels spoken in isolation show that all of these attributes can increase when a person is depressed. In this study, we focused on using features from free-flowing speech to classify the depressed state of an individual. To do so we looked at vowel regions that look the most like sustained vowels. We train an SVM for each speaker and do a speaker dependent classification of the test speech frames. Using the AMDF based feature we got a better accuracy (62–87% frame-wise accuracy for 5 out of 6 speakers) for most speakers than 13 dimensional MFCC along with its velocity and acceleration coefficients. Using the AMDF based feature, we also trained a speaker independent SVM which gave an average accuracy of 77.8% for utterance based classification.

Parkinson's Disease Progression Assessment from Speech Using GMM-UBM

T. Arias-Vergara[1], J.C. Vasquez-Correa[1], Juan Rafael Orozco-Arroyave[1], J.F. Vargas-Bonilla[1], Elmar Nöth[2]; [1]Universidad de Antioquia, Colombia; [2]FAU Erlangen-Nürnberg, Germany

Sat-P-5-4-2, Time: 13:30

The Gaussian Mixture Model Universal Background Model (GMM-UBM) approach is used to assess the Parkinson's disease (PD) progression per speaker. The disease progression is assessed individually per patient following a user modeling-approach. Voiced and unvoiced segments are extracted and grouped separately to train the models. Additionally, the Bhattacharyya distance is used to estimate the difference between the UBM and the user model. Speech recordings from 62 PD patients (34 male and 28 female)

Notes

were captured from 2012 to 2015 in four recording sessions. The validation of the models is performed with recordings of 7 patients. All of the patients were diagnosed by a neurologist expert according to the MDS-UPDRS-III scale. The features used to model the speech of the patients are validated by doing a regression based on a Support Vector Regressor (SVR). According to the results, it is possible to track the disease progression with a Pearson's correlation of up to 0.60 with respect to the MDS-UPDRS-III labels.

Speech-Based Detection of Alzheimer's Disease in Conversational German

Jochen Weiner, Christian Herff, Tanja Schultz; Universität Bremen, Germany

Sat-P-5-4-3, Time: 13:30

The worldwide population is aging. With a larger population of elderly people, the numbers of people affected by cognitive impairment such as Alzheimer's disease are growing. Unfortunately, there is no known cure for Alzheimer's disease. The only way to alleviate it's serious effects is to start therapy very early before the disease has wrought too much irreversible damage. Current diagnostic procedures are neither cost nor time efficient and therefore do not meet the demands for frequent mass screening required to mitigate the consequences of cognitive impairments on the global scale.

We present an experiment to detect Alzheimer's disease using spontaneous conversational speech. The speech data was recorded during biographic interviews in the Interdisciplinary Longitudinal Study on Adult Development and Aging (ILSE), a large data resource on healthy and satisfying aging in middle adulthood and later life in Germany. From these recordings we extract ten speech-based features using voice activity detection and transcriptions. In an experimental setup with 98 data samples we train a linear discriminant analysis classifier to distinguish subjects with Alzheimer's disease from the control group. This setup results in an F-score of 0.8 for the detection of Alzheimer's disease, clearly showing our approach detects dementia well.

Cross-Cultural Depression Recognition from Vocal Biomarkers

Sharifa Alghowinem[1], Roland Goecke[2], Julien Epps[3], Michael Wagner[2], Jeffrey Cohn[4]; [1]Australian National University, Australia; [2]University of Canberra, Australia; [3]University of New South Wales, Australia; [4]University of Pittsburgh, USA

Sat-P-5-4-4, Time: 13:30

No studies have investigated cross-cultural and cross-language characteristics of depressed speech. We investigated the generalisability of a vocal biomarker-based approach to depression detection in clinical interviews recorded in three countries (Australia, the USA and Germany), two languages (German and English) and different accents (Australian and American). Several approaches to training and testing within and between datasets were evaluated. Using the same experimental protocol separately within each dataset, (cross-classification) accuracy was high. Combining datasets, high accuracy was high again and consistent across language, recording environment, and culture. Training and testing between datasets, however, attenuated accuracy. These finding emphasize the importance of heterogeneous training sets for robust depression detection.

Speech Recognition in Alzheimer's Disease and in its Assessment

Luke Zhou[1], Kathleen C. Fraser[1], Frank Rudzicz[2]; [1]University of Toronto, Canada; [2]Toronto Rehabilitation Institute, Canada

Sat-P-5-4-5, Time: 13:30

Narrative, spontaneous speech can provide a valuable source of information about an individual's cognitive state. Unfortunately, clinical transcription of this type of data is typically done by hand, which is prohibitively time-consuming. In order to automate the entire process, we optimize automatic speech recognition (ASR) for participants with Alzheimer's disease (AD) in a relatively large clinical database. We extract text features from the resulting transcripts and use these features to identify AD with an SVM classifier. While the accuracy of automatic assessment decreases with increased WER, this is weakly correlated (-0.31). This relative robustness to ASR error is aided by selecting features that are resilient to ASR error.

Does She Speak RTT? Towards an Earlier Identification of Rett Syndrome Through Intelligent Pre-Linguistic Vocalisation Analysis

Florian B. Pokorny[1], Peter B. Marschik[1], Christa Einspieler[1], Björn Schuller[2]; [1]Medizinische Universität Graz, Austria; [2]Universität Passau, Germany

Sat-P-5-4-6, Time: 13:30

For many years, an apparently normal early development has been regarded as a main characteristic of Rett syndrome (RTT), a severe progressive neurodevelopmental disorder almost exclusively affecting girls/females. The speech-language domain represents a key domain for the clinical diagnosis of RTT, which usually happens around three years of age. Recent studies have built upon the assumption that this domain is already affected in the prodromal period. Aiming to find RTT-specific speech-language atypicalities on signal level as early acoustic markers, we analysed more than 16 hours of home video recordings of 4 girls later diagnosed with RTT and 4 typically developing girls aged 6 to 12 months. We segmented a total of 4 678 pre-linguistic vocalisations. A comprehensive set of acoustic features was extracted from the vocalisations as basis for the classification paradigm RTT versus typical development. A promising mean unweighted recognition accuracy of 76.5% was achieved using linear kernel support vector machines and 4-fold leave-one-speaker-pair-out cross-validation. To the best of our knowledge, this is the first approach to automatically identify infants later diagnosed with RTT based on acoustic characteristics of pre-linguistic vocalisations. Our findings may build the basis for facilitating earlier identification and thus an avenue for an earlier entry into intervention.

Speech Rhythm in Parkinson's Disease: A Study on Italian

Massimo Pettorino[1], Maria Grazia Busà[2], Elisa Pellegrino[1]; [1]Università di Napoli "L'Orientale", Italy; [2]Università di Padova, Italy

Sat-P-5-4-7, Time: 13:30

Experimental studies on different languages have shown that neurogenetic disorders connected with Parkinson's disease (PD) determine a series of variations in the speech rhythm. This study aims at

NOTES

verifying whether the speech of PD patients presents rhythmic abnormalities compared to healthy speakers also in Italian. The read speech of 15 healthy speakers and of 11 patients with mild PD was segmented in consonantal and vocalic portions. After extracting the durations of all segments, the vowel percentage (%V) and the interval between two consecutive vowel onset points (VtoV) were calculated. The results show that %V has significantly different values in mildly affected patients as compared to controls. For Italian, %V spans between 44% and 50% for healthy subjects and between 51% and 58% for PD subjects. A positive correlation was found between %V and the number of years of PD since its insurgence. The correlation with the age at which the disease insurges is weak. With regard to VtoV, PD subjects do not speak at a significantly slower rate than healthy controls, though a trend in this direction was found. The data suggest that %V could be used as a more reliable parameter for the early diagnosis of PD than speech rate.

Sat-S&T-5 : Show & Tell Session 5

Market Street Foyer, 13:30–15:30, Saturday, 10 Sept. 2016
Chairs: Shiva Sundaram, Nicolas Scheffer

English Language Speech Assistant

Xavier Anguera [1], Vu Van [2]; [1] ELSA, Portugal; [2] ELSA, USA

Sat-S&T-5-1, Time: 13:30

This show&tell demo presentation showcases ELSA Speak, an app for English Language pronunciation and intonation improvement that uses speech technology to assess the users speech and to offer consistent feedback on the errors the students make.

Remeeting — Deep Insights to Conversations

Allen Guo, Arlo Faria, Korbinian Riedhammer; Remeeting, USA

Sat-S&T-5-2, Time: 13:30

Remeeting is a cloud service that helps you get insights to (spoken) conversations. Audio and video data such as recorded meetings, online conferences, sales or customer success calls are processed using speaker separation and identification, speech recognition and indexing, and an automated keyword analysis. The resulting annotated “documents” can be shared with others and reviewed using a web app that acts as a visual index to the meeting. Furthermore, the extracted metadata is index by a search engine to allow for efficient cross-document search. A powerful query DSL allows the user to make sophisticated queries such as “what did X say about topic Y in the first quarter of this year” or “show me the keywords for all meetings where X and Y attended”. Similar to retrieval, a watchdog can be set to deliver real-time insights to operations. Use cases include productivity in meetings, compliance and policies, real time callcenter analytics and better accessibility of large archives.

Remeeting is leveraging, promoting and contributing to open source projects including Kaldi, Elasticsearch and Docker.

SERAPHIM *Live!* — Singing Synthesis for the Performer, the Composer, and the 3D Game Developer

*Paul Yaozhu Chan [1], Minghui Dong [1], Grace Xue Hui Ho [2], Haizhou Li [1]; [1] A*STAR, Singapore; [2] NTU, Singapore*

Sat-S&T-5-3, Time: 13:30

The human singing voice is highly expressive instrument capable of producing a variety of complex timbres. Singing synthesis today is popular amongst composers and studio musicians accessing the technology by means of offline sequencing platforms. Only a couple of singing synthesizers are known to be equipped with both the real-time capability and the user interface to successfully target live performances. These are LIMSI’s Cantor Digitalis and Yamaha’s VOCALOID Keyboard. However, both systems have their own shortcomings. The former is limited to vowels and does not synthesize complete words or syllables. The latter is only real-time to the syllable level and thus requires specifications of the entire syllable before it commences in the performance. A demand remains for a singing synthesis system that truly solves the problem of real-time synthesis — a system capable of synthesizing both vowels and consonants to form entire words while being capable of synthesizing in real-time to the sub-frame level. Such a system has to be versatile enough to exhaustively present all acoustic options possible to the user for maximal control while being intelligent enough to fill in acoustic details that are too fine for human reflexes to control.

SERAPHIM is a real-time singing synthesizer developed in answer to this demand. This paper presents the implementation of SERAPHIM for performing musicians and studio musicians, together with how 3D game developers may use Seraphim to deploy singing in their games.

My-Own-Voice: A Web Service That Allows You to Create a Text-to-Speech Voice From Your Own Voice

Fabrice Malfrere, Olivier Deroo, Emmanuelle Franques, Jonathan Hourez, Nicolas Mazars, Vincent Pagel, Geoffrey Wilfart; Acapela Group, Belgium

Sat-S&T-5-4, Time: 13:30

My-Own-Voice is a service that provides a tool to end-users who want to have their voices synthesized by a high-quality commercial-grade Text-to-Speech system without the need to install, configure or manage speech-processing software and equipment. The system records and validates users’ utterances with Automatic Speech Recognition (ASR), to build an HMM or a Unit Selection synthetic voice. All the procedures are automated to avoid human intervention. We describe here the system for particular end-users about to lose the ability to speak with their own voice, who can now synthetically recreate it with the help of their speech therapist, enabling them to preserve this essential part of their identity.

NOTES

Keynote 3: Anne Fernald

Grand Ballroom ABC, 08:30–09:30, Sunday, 11 Sept. 2016
Chair: Shri Narayanan

Talking with Kids Really Matters: Early Language Experience Shapes Later Life Chances

Anne Fernald; Stanford University, USA

Sun-Keynote-3, Time: 08:30

The foundation for lifelong literacy is built through a child's experience with language in the first five years. Integrating research from biological, psycholinguistic, and sociocultural perspectives, I will examine why millions of children fail to reach their developmental potential in the early years and enter school without a strong foundation for learning, resulting in enormous loss of human potential.

Sun-O-6-1 : Far-Field Speech Processing

Grand Ballroom A, 10:00–12:00, Sunday, 11 Sept. 2016
Chairs: Jinyu Li, Tomohiro Nakatani

Reducing the Computational Complexity of Multimicrophone Acoustic Models with Integrated Feature Extraction

Tara N. Sainath, Arun Narayanan, Ron J. Weiss, Ehsan Variani, Kevin W. Wilson, Michiel Bacchiani, Izhak Shafran; Google, USA

Sun-O-6-1-1, Time: 10:00

Recently, we presented a multichannel neural network model trained to perform speech enhancement jointly with acoustic modeling [1], directly from raw waveform input signals. While this model achieved over a 10% relative improvement compared to a single channel model, it came at a large cost in computational complexity, particularly in the convolutions used to implement a time-domain filterbank. In this paper we present several different approaches to reduce the complexity of this model by reducing the stride of the convolution operation and by implementing filters in the frequency domain. These optimizations reduce the computational complexity of the model by a factor of 3 with no loss in accuracy on a 2,000 hour Voice Search task.

Neural Network Adaptive Beamforming for Robust Multichannel Speech Recognition

Bo Li, Tara N. Sainath, Ron J. Weiss, Kevin W. Wilson, Michiel Bacchiani; Google, USA

Sun-O-6-1-2, Time: 10:20

Joint multichannel enhancement and acoustic modeling using neural networks has shown promise over the past few years. However, one shortcoming of previous work [1, 2, 3] is that the filters learned during training are fixed for decoding, potentially limiting the ability of these models to adapt to previously unseen or changing conditions. In this paper we explore a neural network adaptive beamforming (NAB) technique to address this issue. Specifically, we use LSTM layers to predict time domain beamforming filter coefficients at each input frame. These filters are convolved with the framed time domain input signal and summed across channels, essentially performing FIR filter-and-sum beamforming using the dynamically adapted filter. The beamformer output is passed into a waveform CLDNN acoustic model [4] which is trained jointly with the filter prediction LSTM layers. We find that the proposed NAB model achieves a 12.7% relative improvement in WER over a single channel model [4] and reaches similar performance to a "factored" model architecture which utilizes several fixed spatial filters [3] on a 2,000-hour Voice Search task, with a 17.9% decrease in computational cost.

Improved MVDR Beamforming Using Single-Channel Mask Prediction Networks

Hakan Erdogan [1], John R. Hershey [2], Shinji Watanabe [2], Michael I. Mandel [3], Jonathan Le Roux [2]; [1] Sabancı Üniversitesi, Turkey; [2] MERL, USA; [3] CUNY Brooklyn College, USA

Sun-O-6-1-3, Time: 10:40

Recent studies on multi-microphone speech databases indicate that it is beneficial to perform beamforming to improve speech recognition accuracies, especially when there is a high level of background noise. Minimum variance distortionless response (MVDR) beamforming is an important beamforming method that performs quite well for speech recognition purposes especially if the steering vector is known. However, steering the beamformer to focus on speech in unknown acoustic conditions remains a challenging problem. In this study, we use single-channel speech enhancement deep networks to form masks that can be used for noise spatial covariance estimation, which steers the MVDR beamforming toward the speech. We analyze how mask prediction affects performance and also discuss various ways to use masks to obtain the speech and noise spatial covariance estimates in a reliable way. We show that using a single mask across microphones for covariance prediction with minima-limited post-masking yields the best result in terms of signal-level quality measures and speech recognition word error rates in a mismatched training condition.

Channel Selection for Distant Speech Recognition Exploiting Cepstral Distance

Cristina Guerrero [1], Georgina Tryfou [1], Maurizio Omologo [2]; [1] Università di Trento, Italy; [2] FBK, Italy

Sun-O-6-1-4, Time: 11:00

In a multi-microphone distant speech recognition task, the redundancy of information that results from the availability of multiple instances of the same source signal can be exploited through channel selection. In this work, we propose the use of cepstral distance as a means of assessment of the available channels, in an informed and a blind fashion. In the informed approach the distances between the close-talk and all of the channels are calculated. In the blind method, the cepstral distances are computed using an estimated reference signal, assumed to represent the average distortion among the available channels. Furthermore, we propose a new evaluation methodology that better illustrates the strengths and weaknesses of a channel selection method, in comparison to the sole use of word error rate. The experimental results suggest that the proposed blind method successfully selects the least distorted channel, when sufficient room coverage is provided by the microphone network. As a result, improved recognition rates are obtained in a distant speech recognition task, both in a simulated and a real context.

Notes

Multichannel Spatial Clustering for Robust Far-Field Automatic Speech Recognition in Mismatched Conditions

Michael I. Mandel[1], Jon Barker[2]; [1]CUNY Brooklyn College, USA; [2]University of Sheffield, UK

Sun-O-6-1-5, Time: 11:20

Recent automatic speech recognition (ASR) results are quite good when the training data is matched to the test data, but much worse when they differ in some important regard, like the number and arrangement of microphones or differences in reverberation and noise conditions. This paper proposes an unsupervised spatial clustering approach to microphone array processing that can overcome such train-test mismatches. This approach, known as Model-based EM Source Separation and Localization (MESSL), clusters spectrogram points based on the relative differences in phase and level between pairs of microphones. Here it is used for the first time to drive minimum variance distortionless response (MVDR) beamforming in several ways. We compare it to a standard delay-and-sum beamformer on the CHiME-3 noisy test set (real recordings), using each system as a pre-processor for the same recognizer trained on the AMI meeting corpus. We find that the spatial clustering front end reduces word error rates by between 9.9 and 17.1% relative to the baseline.

Far-Field ASR Without Parallel Data

Vijayaditya Peddinti, Vimal Manohar, Yiming Wang, Daniel Povey, Sanjeev Khudanpur; Johns Hopkins University, USA

Sun-O-6-1-6, Time: 11:40

In far-field speech recognition systems, training acoustic models with alignments generated from parallel close-talk microphone data provides significant improvements. However it is not practical to assume the availability of large corpora of parallel close-talk microphone data, for training. In this paper we explore methods to reduce the performance gap between far-field ASR systems trained with alignments from distant microphone data and those trained with alignments from parallel close-talk microphone data. These methods include the use of a lattice-free sequence objective function which tolerates minor mis-alignment errors; and the use of data selection techniques to discard badly aligned data. We present results on single distant microphone and multiple distant microphone scenarios of the AMI LVCSR task. We identify prominent causes of alignment errors in AMI data.

Sun-O-6-2 : Special Session: Interspeech 2016 Computational Paralinguistics Challenge (ComParE): Deception, Sincerity & Native Language

Grand Ballroom BC, 10:00–12:00, Sunday, 11 Sept. 2016

Chairs: Björn Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee Burgoon, Eduardo Coutinho

The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language

Björn Schuller[1], Stefan Steidl[2], Anton Batliner[3], Julia Hirschberg[4], Judee K. Burgoon[5], Alice Baird[4], Aaron Elkins[5], Yue Zhang[1], Eduardo Coutinho[1], Keelan Evanini[6]; [1]Imperial College London, UK; [2]FAU Erlangen-Nürnberg, Germany; [3]Universität Passau, Germany; [4]Columbia University, USA; [5]University of Arizona, USA; [6]Educational Testing Service, USA

Sun-O-6-2-1, Time: 10:00

The INTERSPEECH 2016 Computational Paralinguistics Challenge addresses three different problems for the first time in research competition under well-defined conditions: classification of deceptive vs. non-deceptive speech, the estimation of the degree of sincerity, and the identification of the native language out of eleven L1 classes of English L2 speakers. In this paper, we describe these sub-challenges, their conditions, the baseline feature extraction and classifiers, and the resulting baselines, as provided to the participants.

The Deception Sub-Challenge: The Data

Björn Schuller[1], Stefan Steidl[2], Anton Batliner[3], Julia Hirschberg[4], Judee K. Burgoon[5], Alice Baird[4], Aaron Elkins[5], Yue Zhang[1], Eduardo Coutinho[1], Keelan Evanini[6]; [1]Imperial College London, UK; [2]FAU Erlangen-Nürnberg, Germany; [3]Universität Passau, Germany; [4]Columbia University, USA; [5]University of Arizona, USA; [6]Educational Testing Service, USA

Sun-O-6-2-2, Time: 10:10

(No abstract available at the time of publication)

Combining Acoustic-Prosodic, Lexical, and Phonotactic Features for Automatic Deception Detection

Sarah Ita Levitan[1], Guozhen An[2], Min Ma[2], Rivka Levitan[3], Andrew Rosenberg[4], Julia Hirschberg[2]; [1]Columbia University, USA; [2]CUNY Graduate Center, USA; [3]CUNY Brooklyn College, USA; [4]CUNY Queens College, USA

Sun-O-6-2-3, Time: 10:20

Improving methods of automatic deception detection is an important goal of many researchers from a variety of disciplines, including psychology, computational linguistics, and criminology.

NOTES

We present a system to automatically identify deceptive utterances using acoustic-prosodic, lexical, syntactic, and phonotactic features. We train and test our system on the Interspeech 2016 ComParE challenge corpus, and find that our combined features result in performance well above the challenge baseline on the development data. We also perform feature ranking experiments to evaluate the usefulness of each of our feature sets. Finally, we conduct a cross-corpus evaluation by training on another deception corpus and testing on the ComParE corpus.

Is Deception Emotional? An Emotion-Driven Predictive Approach

Shahin Amiriparian, Jouni Pohjalainen, Erik Marchi, Sergey Pugachevskiy, Björn Schuller; Universität Passau, Germany

Sun-O-6-2-4, Time: 10:30

In this paper, we propose a method for automatically detecting deceptive speech by relying on predicted scores derived from emotion dimensions such as arousal, valence, regulation, and emotion categories. The scores are derived from task-dependent models trained on the GEMEP emotional speech database. Inputs from the INTERSPEECH 2016 Computational Paralinguistics Deception sub-challenge are processed to obtain predictions of emotion attributes and associated scores that are then used as features in detecting deception. We show that using the new emotion-related features, it is possible to improve upon the challenge baseline.

Prosodic Cues and Answer Type Detection for the Deception Sub-Challenge

Claude Montacié, Marie-José Caraty; STIH (EA 4509), France

Sun-O-6-2-5, Time: 10:40

Deception is a deliberate act to deceive interlocutor by transmitting a message containing false or misleading information. Detection of deception consists in the search for reliable differences between liars and truth-tellers. In this paper, we used the Deceptive Speech Database (DSD) provided for the Deception sub-challenge. DSD consists of deceptive and non-deceptive answers to a set of unknown questions. We have investigated linguistic cues: prosodic cues (pauses and phone duration, speech segmentation) and answer types (e.g., opinion, self-report, offense denial). These cues were automatically detected using the CMU-Sphinx toolkit for speech recognition (acoustic-phonetic decoding, isolated word recognition and keyword spotting). Two kinds of prosodic features were computed from the speech transcriptions (phoneme, silent pause, filled pause, and breathing): the usual speech rate measures and the audio feature based on the multi-resolution paradigm. The answer type features were introduced. A set of answer types was chosen from the transcription of the Training set and each answer type was modeled by a bag-of-words. Experiments have shown improvements of 13.0% and 3.8% on the Development and Test sets respectively, compared to the official baseline Unweighted Average Recall.

The Sincerity Sub-Challenge: The Data

Björn Schuller[1], Stefan Steidl[2], Anton Batliner[3], Julia Hirschberg[4], Judee K. Burgoon[5], Alice Baird[4], Aaron Elkins[5], Yue Zhang[1], Eduardo Coutinho[1], Keelan Evanini[6]; [1]Imperial College London, UK; [2]FAU Erlangen-Nürnberg, Germany; [3]Universität Passau, Germany; [4]Columbia University, USA; [5]University of Arizona, USA; [6]Educational Testing Service, USA

Sun-O-6-2-6, Time: 10:50

(No abstract available at the time of publication)

Automatic Estimation of Perceived Sincerity from Spoken Language

Brandon M. Booth, Rahul Gupta, Pavlos Papadopoulos, Ruchir Travadi, Shrikanth S. Narayanan; University of Southern California, USA

Sun-O-6-2-7, Time: 11:00

Sincerity is important in everyday human communication and perception of genuineness can greatly affect emotions and outcomes in social interactions. In this paper, submitted for the INTERSPEECH 2016 Sincerity Challenge, we examine a corpus of six different types of apologetic utterances from a variety of English speakers articulated in different prosodic styles, and we rate the sincerity of each remark. Since the utterances and semantic meaning in the examined database are controlled, we focus on tone of voice by exploring a plethora of acoustic and paralinguistic features not present in the baseline model and how well they contribute to human assessment of sincerity. We show that these additional features improve the performance using the baseline model, and furthermore that conditioning learning models on the prosody of utterances boosts the prediction accuracy. Our best system outperforms the challenge baseline and in principle can generalize well to other corpora.

Estimating the Sincerity of Apologies in Speech by DNN Rank Learning and Prosodic Analysis

Gábor Gosztolya[1], Tamás Grósz[1], György Szaszák[2], László Tóth[3]; [1]University of Szeged, Hungary; [2]BME, Hungary; [3]MTA-SZTE RGAI, Hungary

Sun-O-6-2-8, Time: 11:10

In the Sincerity Sub-Challenge of the Interspeech ComParE 2016 Challenge, the task is to estimate user-annotated sincerity scores for speech samples. We interpret this challenge as a rank-learning regression task, since the evaluation metric (Spearman's correlation) is calculated from the rank of the instances. As a first approach, Deep Neural Networks are used by introducing a novel error criterion which maximizes the correlation metric directly. We obtained the best performance by combining the proposed error function with the conventional MSE error. This approach yielded results that outperform the baseline on the Challenge test set. Furthermore, we introduce a compact prosodic feature set based on a dynamic representation of F0, energy and sound duration. We extract syllable-based prosodic features which are used as the basis of another machine learning step. We show that a small set of prosodic features is capable of yielding a result very close to the baseline one and that by combining the predictions yielded by DNN

NOTES

and the prosodic feature set, further improvement can be reached, significantly outperforming the baseline SVR on the Challenge test set.

Minimization of Regression and Ranking Losses with Shallow Neural Networks on Automatic Sincerity Evaluation

Hung-Shin Lee [1], Yu Tsao [2], Chi-Chun Lee [3], Hsin-Min Wang [2], Wei-Cheng Lin [3], Wei-Chen Chen [3], Shan-Wen Hsiao [3], Shyh-Kang Jeng [1]; [1]National Taiwan University, Taiwan; [2]Academia Sinica, Taiwan; [3]National Tsing Hua University, Taiwan

Sun-O-6-2-9, Time: 11:20

To estimate the degree of sincerity conveyed by a speech utterance and received by listeners, we propose an instance-based learning framework with shallow neural networks. The framework plays as not only a regressor that intends to fit the predicted value to the actual value but also a ranker that preserves the relative target magnitude between each pair of utterances, in an attempt to derive a higher Spearman's rank correlation coefficient. In addition to describing how to simultaneously minimize regression and ranking losses, the issue of how utterance pairs work in the training and evaluation phases is also addressed by two kinds of realizations. The intuitive one is related to random sampling while the other seeks for representative utterances, named anchors, to form non-stochastic pairs. Our system outperforms the baseline by more than 25% relative improvement in the development set.

Prediction of Deception and Sincerity from Speech Using Automatic Phone Recognition-Based Features

Robert Herms; Technische Universität Chemnitz, Germany

Sun-O-6-2-10, Time: 11:30

As part of the Interspeech 2016 COMPARE challenge, the two different sub-challenges Deception and Sincerity are addressed. The former refers to the identification of deceptive speech whereas the degree of perceived sincerity of speakers has to be estimated in the latter. In this paper, we investigate the potential of automatic phone recognition-based features for these use case scenarios. The speech transcriptions were used to process the appearing tokens (phoneme, silent pause, filled pause) and the corresponding durations. We designed a high-level feature set including the four groups: vowels, phones, pseudo syllables, and pauses. Additionally, we selected suitable predefined acoustic feature sets and fused them with our introduced features showing a positive effect on the prediction. Moreover, the performance is further boosted by refining these fused features using the ReliefF feature selection method. Experiments show that the final systems outperform the baseline results of both sub-challenges.

Sincerity and Deception in Speech: Two Sides of the Same Coin? A Transfer- and Multi-Task Learning Perspective

Yue Zhang [1], Felix Weninger [2], Zhao Ren [3], Björn Schuller [1]; [1]Imperial College London, UK; [2]Nuance Communications, Germany; [3]Northwestern Polytechnical University, China

Sun-O-6-2-11, Time: 11:40

In this work, we investigate the coherence between inferable deception and perceived sincerity in speech, as featured in the Deception and Sincerity tasks of the INTERSPEECH 2016 Computational Paralinguistics ChallengE (ComParE). We demonstrate an effective approach that combines the corpora of both Challenge tasks to achieve higher classification accuracy. We show that the naïve label mapping method based on the assumption that sincerity and deception are just 'two sides of the same coin', i. e., taking deceptive speech as equivalent to non-sincere speech and vice versa, does not yield satisfactory results. However, we can exploit the interplay and synergies between these characteristics. To achieve this, we combine our previously introduced approach for data aggregation by semi-supervised cross-task label completion with multi-task learning, and knowledge-based instance selection. In the result, our approach achieves significant error rate reductions compared to the official Challenge baseline.

Fusing Acoustic Feature Representations for Computational Paralinguistics Tasks

Heysem Kaya [1], Alexey A. Karpov [2]; [1]Namık Kemal Üniversitesi, Turkey; [2]Russian Academy of Sciences, Russia

Sun-O-6-2-12, Time: 11:50

The field of Computational Paralinguistics is rapidly growing and is of interest in various application domains ranging from biomedical engineering to forensics. The INTERSPEECH ComParE challenge series has a field-leading role, introducing novel problems with a common benchmark protocol for comparability. In this work, we tackle all three ComParE 2016 Challenge corpora (Native Language, Sincerity and Deception) benefiting from multi-level normalization on features followed by fast and robust kernel learning methods. Moreover, we employ computer vision inspired low level descriptor representation methods such as the Fisher vector encoding. After non-linear preprocessing, obtained Fisher vectors are kernelized and mapped to target variables by classifiers based on Kernel Extreme Learning Machines and Partial Least Squares regression. We finally combine predictions of models trained on popularly used functional based descriptor encoding (openSMILE features) with those obtained from the Fisher vector encoding. In the preliminary experiments, our approach has significantly outperformed the baseline systems for Native Language and Sincerity sub-challenges both in the development and test sets.

Notes

Sun-O-6-3 : Special Session: Speech, Audio, and Language Processing Techniques Applied to Bird and Animal Vocalizations

Bayview A, 10:00–12:00, Sunday, 11 Sept. 2016
Chairs: Naomi Harte, Peter Jančovič, Karl-L. Schuchmann

Introduction

Naomi Harte[1], Peter Jančovič[2], Karl-L. Schuchmann[3]; [1]Trinity College Dublin, Ireland; [2]University of Birmingham, UK; [3]ZFMK, Germany

Sun-O-6-3-1, Time: 10:00

(No abstract available at the time of publication)

Poster Overview Presentations

Naomi Harte[1], Peter Jančovič[2], Karl-L. Schuchmann[3]; [1]Trinity College Dublin, Ireland; [2]University of Birmingham, UK; [3]ZFMK, Germany

Sun-O-6-3-2, Time: 10:05

(No abstract available at the time of publication)

Discussion

Naomi Harte[1], Peter Jančovič[2], Karl-L. Schuchmann[3]; [1]Trinity College Dublin, Ireland; [2]University of Birmingham, UK; [3]ZFMK, Germany

Sun-O-6-3-3, Time: 11:15

(No abstract available at the time of publication)

Closing Remarks

Naomi Harte[1], Peter Jančovič[2], Karl-L. Schuchmann[3]; [1]Trinity College Dublin, Ireland; [2]University of Birmingham, UK; [3]ZFMK, Germany

Sun-O-6-3-4, Time: 11:55

(No abstract available at the time of publication)

Sun-O-6-4 : Dialogue Systems and Analysis of Dialogue

Bayview B, 10:00–12:00, Sunday, 11 Sept. 2016
Chairs: Jens Edlund, Alexandros Potamianos

A Stochastic Model for Computer-Aided Human-Human Dialogue

Merwan Barlier[1], Romain Laroche[1], Olivier Pietquin[2]; [1]Orange Labs, France; [2]CRIStAL, France

Sun-O-6-4-1, Time: 10:00

In this paper we introduce a novel model for computer-aided human-human dialogue. In this context, the computer aims at improving the outcome of a human-human task-oriented dialogue by intervening during the course of the interaction. While dialogue state and topic tracking in human-human dialogue have already been studied, few work has been devoted to the sequential part of the problem, where the impact of the system's actions on the future of the conversation is taken into account. This paper addresses this issue by first modelling human-human dialogue as a Markov Reward Process. The task of purposely taking part into the conversation is then optimised within the Linearly Solvable Markov Decision Process framework. Utterances of the Conversational Agent are seen as perturbations in this process, which aim at satisfying the user's long-term goals while keeping the conversation natural. Finally, results obtained by simulation suggest that such an approach is suitable for computer-aided human-human dialogue and is a first step towards three-party dialogue.

Highlighting Psychological Features for Predicting Child Interjections During Story Telling

Gaël Lejeune, François Rioult, Bruno Crémilleux; GREYC, France

Sun-O-6-4-2, Time: 10:20

Conversational agents are more and more investigated by the community but their ability to keep the user committed in the interaction is limited. Predicting the behavior of children in a human-machine interaction setting is a key issue for the success of narrative conversational agents. In this paper, we investigate solutions to evaluate the child's commitment in the story and to detect when the child is likely to react during the story. We show that the conversational agent cannot solely count on questions and requests for attention to stimulate the child. We assess how (1) psychological features allow to improve the prediction of children interjections and how (2) exploiting these features with Pattern Mining techniques offers better results. Experiments show that psychological features improves the predictions and furthermore help to produce robust dialog models.

Hybrid Dialogue State Tracking for Real World Human-to-Human Dialogues

Kai Sun, Su Zhu, Lu Chen, Siqiu Yao, Xueyang Wu, Kai Yu; Shanghai Jiao Tong University, China

Sun-O-6-4-3, Time: 10:40

Dialogue state tracking is a key sub-task of dialogue management. The fourth Dialog State Tracking Challenge (DSTC-4) focuses on dialogue state tracking for real world human-to-human dialogues. The task is more challenging than previous challenges because of more complex domain and coreferences, more synonyms and abbreviations, sub-dialogue level labelled utterances, and no spoken language understanding output provided. To deal with these challenges, this paper proposes a novel hybrid dialogue state tracking method, which can take advantage of the strength of both rule-based and statistical methods. Thousands of rules are first automatically generated using a template-based rule generation approach and then combined together with several manually designed rules to yield the output of the rule-based method. In parallel, a statistical method is applied to track the state. The tracker finally takes the union of the outputs of the two methods. In DSTC-4 evaluation, the proposed hybrid tracker obtained state-of-the-art results. It ranked the second and significantly outperformed the baseline system and most submissions.

Notes

Automatic Recognition of Social Roles Using Long Term Role Transitions in Small Group Interactions

Gaurav Fotedar [1], Aditya Gaonkar P. [1], Saikat Chatterjee [2], Prasanta Kumar Ghosh [1]; [1] Indian Institute of Science, India; [2] KTH, Sweden

Sun-O-6-4-4, Time: 11:00

Recognition of social roles in small group interactions is challenging because of the presence of disfluency in speech, frequent overlaps between speakers, short speaker turns and the need for reliable data annotation. In this work, we consider the problem of recognizing four roles, namely Gatekeeper, Protagonist, Neutral, and Supporter in small group interactions in AMI corpus. In general, Gatekeeper and Protagonist roles occur less frequently compared to Neutral, and Supporter. In this work, we exploit role transitions across segments in a meeting by incorporating role transition probabilities and formulating the role recognition as a decoding problem over the sequence of segments in an interaction. Experiments are performed in a five fold cross validation setup using acoustic, lexical and structural features with precision, recall and F-score as the performance metrics. The results reveal that precision averaged across all folds and different feature combinations improves in the case of Gatekeeper and Protagonist by 13.64% and 12.75% when the role transition information is used which in turn improves the F-score for Gatekeeper by 6.58% while the F-scores for the rest of the roles do not change significantly.

On the Influence of Gender on Interruptions in Multiparty Dialogue

Paul Van Eecke [1], Raquel Fernández [2]; [1] Sony, France; [2] Universiteit van Amsterdam, The Netherlands

Sun-O-6-4-5, Time: 11:20

During conversations, participants do not always alternate turns smoothly. One cause of disturbance particularly prominent in multiparty dialogue is the presence of interruptions: interventions that prevent current speakers from finishing their turns. Previous work, mostly within the field of sociolinguistics, has suggested that the gender of the dialogue participants plays an important role in their interruptive behaviour. We investigate existing hypotheses in this respect by systematically analysing interruptions in a corpus of spoken multiparty meetings that include a minimum of two male and two female participants. We find a number of significant differences, including the fact that women are more often interrupted overall and that men interrupt more often women than other men, in particular using speech overlap to grab the floor. We do not find evidence for the hypothesis that women interrupt other women more frequently than they interrupt men.

Detection of User Escalation in Human-Computer Interactions

Ian Beaver, Cynthia Freeman; NextIT, USA

Sun-O-6-4-6, Time: 11:40

Detection of virtual agent conversations where a user requests an alternative channel for the completion of a task, known as an escalation request, is necessary for the improvement of language models and for a better user experience. Although methods exist for proactive escalation, we instead wish to explicitly detect escalation requests. In addition, these proactive methods depend on features that do not correlate highly with open-ended chats found in many modern virtual agents. We propose a strategy that can apply to both bounded and open-ended systems since our method has no assumptions on the implementation of the underlying language model. By combining classifiers with several conversation features, we successfully detect escalation requests in real world data.

Sun-O-6-5 : Interaction between Speech Production and Perception

Seacliff BCD, 10:00–12:00, Sunday, 11 Sept. 2016
Chairs: Outi Tuomainen, Christopher Davis

Assessing Idiosyncrasies in a Bayesian Model of Speech Communication

Marie-Lou Barnaud [1], Julien Diard [2], Pierre Bessière [3], Jean-Luc Schwartz [1]; [1] GIPSA, France; [2] LPNC, France; [3] ISIR, France

Sun-O-6-5-1, Time: 10:00

Although speakers of one specific language share the same phoneme representations, their productions can differ. We propose to investigate the development of these differences in production, called idiosyncrasies, by using a Bayesian model of communication. Supposing that idiosyncrasies appear during the development of the motor system, we present two versions of the motor learning phase, both based on the guidance of an agent master: “a repetition model” where agents try to imitate the *sounds* produced by the master and “a communication model” where agents try to replicate the *phonemes* produced by the master. Our experimental results show that only the “communication model” provides production idiosyncrasies, suggesting that idiosyncrasies are a natural output of a motor learning process based on a communicative goal.

Prosodic and Linguistic Analysis of Semantic Fluency Data: A Window into Speech Production and Cognition

Maria K. Wolters [1], Najoung Kim [2], Jung-Ho Kim [2], Sarah E. MacPherson [1], Jong C. Park [2]; [1] University of Edinburgh, UK; [2] KAIST, Korea

Sun-O-6-5-2, Time: 10:20

Semantic fluency is a commonly used task in psychology that provides data about executive function and semantic memory. Performance on the task is affected by conditions ranging from depression to dementia. The task involves participants naming as many members of a given category (e.g. animals) as possible in sixty seconds. Most of the analyses reported in the literature only rely on word counts and transcribed data, and do not take into account the evidence of utterance planning present in the speech signal. Using data from Korean, we show how prosodic analyses can be combined with computational linguistic analyses of the words produced to provide further insights into the processes involved in producing fluency data. We compare our analyses to an established analysis method for semantic fluency data, manual determination of lexically coherent clusters of words.

NOTES

Sensorimotor Response to Visual Imagery of Tongue Displacement

William F. Katz [1], Divya Prabhakaran [2]; [1] University of Texas at Dallas, USA; [2] Plano East Senior High School, USA

Sun-O-6-5-3, Time: 10:40

To better understand audiovisual speech processing, we investigated the effects of viewing time-synchronized videos of a 3D tongue avatar on vowel production by healthy individuals. A group of 15 American English-speaking subjects heard pink noise over headphones and produced the word *head* under four viewing conditions: First, while viewing repetitions of the same vowel, /ɛ/ (baseline phase), then during a series of "morphed" videos shifting gradually from /ɛ/ to /æ/ (ramp phase), followed by repetitions of /æ/ (maximum hold phase), and finally repetitions of /ɛ/ (after effects phase). Results of a formant frequency (F1) analysis indicated that the visual mismatch phases (ramp and maximum hold) caused all subjects to align their productions to the visually-presented vowel, /æ/. No subjects reported being aware that their vowel quality had changed. We conclude that the visual moving tongue stimuli produced entrainment to the viewed vowel category, rather than adaptation in the opposite direction of the perturbation. Further experimentation is needed to determine whether these effects are due to inherent imitation behaviors or subjects' lack of agency with the tongue avatar.

Does Auditory-Motor Learning of Speech Transfer from the CV Syllable to the CVCV Word?

Tiphaine Caudrelier, Pascal Perrier, Jean-Luc Schwartz, Amélie Rochet-Capellan; GIPSA, France

Sun-O-6-5-4, Time: 11:00

Speech is often described as a sequence of units associating linguistic, sensory and motor representations. Is the connection between these representations preferentially maintained at a specific level in terms of a linguistic unit? In the present study, we contrasted the possibility of a link at the level of the syllable (CV) and the word (CVCV). We modified the production of the syllable /be/ in French speakers using an auditory-motor adaptation paradigm that consists of altering the speakers' auditory feedback. After stopping the perturbation, we studied to what extent this modification would transfer to the production of the disyllabic word /bebe/ and compared it to the after-effect on /be/.

The results show that changes in /be/ transfer partially to /bebe/. The partial influence of the somatosensory and motor representations associated with the syllable on the production of the disyllabic word suggests that both units may contribute to the specification of the motor goals in speech sequences. In addition, the transfer occurs to a larger extent in the first syllable of /bebe/ than in the second one. It raises new questions about a possible interaction between the transfer of auditory-motor learning and serial control processes.

Exemplar Dynamics in Phonetic Convergence of Speech Rate

Antje Schweitzer, Michael Walsh; Universität Stuttgart, Germany

Sun-O-6-5-5, Time: 11:20

We motivate and test an exemplar-theoretic view of phonetic convergence, in which convergence effects arise because exemplars just perceived in a conversation are stored in a speaker's memory, and used subsequently in speech production. Most exemplar models assume that production targets are established using stored exemplars, taking into account their frequency- and recency-influenced level of activation. Thus, convergence effects are expected to arise because the exemplars just perceived from a partner have a comparably high activation. However, in the case of frequent exemplars, this effect should be countered by the high frequency of already stored, older exemplars. We test this assumption by examining speech rate convergence in spontaneous speech by female German speakers. We fit two linear mixed models, calculating speech rate on the basis of either infrequent, or frequent, syllables, and predict a speaker's speech rate in a phrase by the partner's speech rate in the preceding phrase. As anticipated, we find a significant main effect indicating convergence only for the infrequent syllables. We also find an unexpected significant interaction of the partner's speech rate and the speaker's assessment of the partner in terms of likeability, indicating divergence, but again, only for the infrequent case.

Articulation Rate in Adverse Listening Conditions in Younger and Older Adults

Outi Tuomainen, Valerie Hazan; University College London, UK

Sun-O-6-5-6, Time: 11:40

Speech communication becomes increasingly difficult with age, especially in adverse listening conditions. We compared speech adaptations made by 'older adult' (65–84 years) and 'younger adult' (19–26 years) talkers when speech is produced with communicative intent. The aim was to investigate how articulation rate is affected by the type of adverse listening condition and by the change in task demands. Articulation rate was recorded in 35 older and 18 younger adult talkers when they were reading and repeating BKB-sentences and when they were doing an interactive 'spot-the-difference' game in a good and three adverse listening conditions (Hearing Loss Simulation, one speaker in noise, both speakers in noise). Similar to younger adults, older adults reduced their articulation rate in the cognitively simpler sentence repetition task in response to adverse conditions. However, in spontaneous speech, only older adult women decreased their articulation rate to counter the effect of the adverse conditions to the same degree as the younger adult talkers. Older men did not reduce their articulation rate in any of the three adverse conditions. These sex differences were not due to differences in the task difficulty experienced by men and women nor were they associated with sensory or cognitive factors.

Sun-O-6-6 : Multimodal Processing

Seacliff A, 10:00–12:00, Sunday, 11 Sept. 2016
Chairs: Florian Metze, Angeliki Metallinou

Error Correction in Lightly Supervised Alignment of Broadcast Subtitles

Julia Olcoz [1], Oscar Saz [2], Thomas Hain [2]; [1] Universidad de Zaragoza, Spain; [2] University of Sheffield, UK

Sun-O-6-6-1, Time: 10:00

This paper presents a range of error correction techniques aimed at improving the accuracy of a lightly supervised alignment task for broadcast subtitles. Lightly supervised approaches are frequently

Notes

used in the multimedia domain, either for subtitling purposes or for providing a more reliable source for training speech-based systems. The proposed methods focus on directly correcting of the alignment output using different techniques to infer word insertions and words with inaccurate time boundaries. The features used by the classification models are the outputs from the alignment system, such as confidence measures, and word or segment duration. Experiments in this paper are based on broadcast material provided by the BBC to the Multi-Genre Broadcast (MGB) challenge participants. Results, show that the order alignment F-measure improves up to 2.6% absolute (15.8% relative) when combining insertion and word-boundary correction.

Automatic Genre and Show Identification of Broadcast Media

Mortaza Doulaty, Oscar Saz, Raymond W.M. Ng, Thomas Hain; University of Sheffield, UK

Sun-O-6-6-2, Time: 10:20

Huge amounts of digital videos are being produced and broadcast every day, leading to giant media archives. Effective techniques are needed to make such data accessible further. Automatic meta-data labelling of broadcast media is an essential task for multimedia indexing, where it is standard to use multi-modal input for such purposes. This paper describes a novel method for automatic detection of media genre and show identities using acoustic features, textual features or a combination thereof. Furthermore the inclusion of available meta-data, such as time of broadcast, is shown to lead to very high performance. Latent Dirichlet Allocation is used to model both acoustics and text, yielding fixed dimensional representations of media recordings that can then be used in Support Vector Machines based classification. Experiments are conducted on more than 1200 hours of TV broadcasts from the British Broadcasting Corporation (BBC), where the task is to categorise the broadcasts into 8 genres or 133 show identities. On a 200-hour test set, accuracies of 98.6% and 85.7% were achieved for genre and show identification respectively, using a combination of acoustic and textual features with meta-data.

Speaker-Targeted Audio-Visual Models for Speech Recognition in Cocktail-Party Environments

Guan-Lin Chao, William Chan, Ian Lane; Carnegie Mellon University, USA

Sun-O-6-6-3, Time: 10:40

Speech recognition in cocktail-party environments remains a significant challenge for state-of-the-art speech recognition systems, as it is extremely difficult to extract an acoustic signal of an individual speaker from a background of overlapping speech with similar frequency and temporal characteristics. We propose the use of speaker-targeted acoustic and audio-visual models for this task. We complement the acoustic features in a hybrid DNN-HMM model with information of the target speaker's identity as well as visual features from the mouth region of the target speaker. Experimentation was performed using simulated cocktail-party data generated from the GRID audio-visual corpus by overlapping two speakers's speech on a single acoustic channel. Our audio-only baseline achieved a WER of 26.3%. The audio-visual model improved the WER to 4.4%. Introducing speaker identity information had an even more pronounced effect, improving the WER to 3.6%. Combining both approaches, however, did not significantly improve performance further. Our work demonstrates that speaker-targeted models can significantly improve the speech recognition in cocktail-party environments.

Text-Dependent Audiovisual Synchrony Detection for Spoofing Detection in Mobile Person Recognition

Amit Aides, Hagai Aronowitz; IBM, Israel

Sun-O-6-6-4, Time: 11:00

Liveness detection is an important countermeasure against spoofing attacks on biometric authentication systems. In the context of audiovisual biometrics, synchrony detection is a proposed method for liveness confirmation. This paper presents a novel, text-dependent scheme for checking audiovisual synchronization in a video sequence. We present custom visual features learned using a unique deep learning framework and show that they outperform other commonly used visual features. We tested our system on two testing sets representing realistic spoofing attack approaches. On our mobile dataset of short video clips of people talking, we obtained equal error rates of 0.8% and 2.7% for liveness detection of photos and video attacks, respectively.

Improving Boundary Estimation in Audiovisual Speech Activity Detection Using Bayesian Information Criterion

Fei Tao, John H.L. Hansen, Carlos Busso; University of Texas at Dallas, USA

Sun-O-6-6-5, Time: 11:20

A key preprocessing step in multimodal interfaces is to detect when a user is speaking to the system. While push-to-talk approaches are effective, its use limits the flexibility of the system. Solutions based on *speech activity detection* (SAD) offer more intuitive and user-friendly alternatives. A limitation in current SAD solutions is the drop in performance observed in noisy environments or when the speech mode differs from neutral speech (e.g., whisper speech). Emerging audiovisual solutions provide a principled framework to improve detection of speech boundaries by incorporating lip activity detection. In our previous work, we proposed an unsupervised *visual speech activity detection* (V-SAD) system that combines temporal and dynamic facial features. The key limitation of the system was the precise detection of boundaries between speech and non-speech regions due to anticipatory facial movements and low video resolution (29.97fps). This study builds upon this system by (a) combining speech and facial features creating an unsupervised *audiovisual speech activity detection* (AV-SAD) system, (b) refining the decision boundary with the *Bayesian information criterion* (BIC) algorithm, resulting in improved speech boundary detection. The evaluation considers the challenging case of whisper speech, where the proposed AV-SAD achieves a 10% absolute improvement over a state-of-the-art audio SAD.

Dynamic Stream Weighting for Turbo-Decoding-Based Audiovisual ASR

Sebastian Gergen[1], Steffen Zeiler[1], Ahmed Hussen Abdelaziz[2], Robert Nickel[3], Dorothea Kolossa[1]; [1]Ruhr-Universität Bochum, Germany; [2]ICSI, USA; [3]Bucknell University, USA

Sun-O-6-6-6, Time: 11:40

Automatic speech recognition (ASR) enables very intuitive human-machine interaction. However, signal degradations due to rever-

NOTES

beration or noise reduce the accuracy of audio-based recognition. The introduction of a second signal stream that is not affected by degradations in the audio domain (e.g., a video stream) increases the robustness of ASR against degradations in the original domain. Here, depending on the signal quality of audio and video at each point in time, a dynamic weighting of both streams can optimize the recognition performance. In this work, we introduce a strategy for estimating optimal weights for the audio and video streams in turbo-decoding-based ASR using a discriminative cost function. The results show that turbo decoding with this maximally discriminative dynamic weighting of information yields higher recognition accuracy than turbo-decoding-based recognition with fixed stream weights or optimally dynamically weighted audiovisual decoding using coupled hidden Markov models.

Sun-P-6-1 : Pitch, Tone, and Music

Pacific Concourse – Poster A, 10:00–12:00, Sunday, 11 Sept. 2016
Chair: Vikramjit Mitra

Retrieval of Textual Song Lyrics from Sung Inputs

Anna M. Kruspe; Fraunhofer IDMT, Germany

Sun-P-6-1-1, Time: 10:00

Retrieving the lyrics of a sung recording from a database of text documents is a research topic that has not received attention so far. Such a retrieval system has many practical applications, e.g. for karaoke applications or for indexing large song databases by their lyric content.

In this paper, we present such a lyrics retrieval system. In a first step, phoneme posteriorgrams are extracted from sung recordings using various acoustic models trained on *TIMIT* and a variation thereof, and on subsets of a large database of recordings of unaccompanied singing (*DAMP*). On the other side, we generate binary templates from the available textual lyrics. Since these lyrics do not have any temporal information, we then employ an approach based on Dynamic Time Warping to retrieve the most likely lyrics document for each recording.

The approach is tested on a different subset of the unaccompanied singing database which includes 601 recordings of 301 different songs (12000 lines of lyrics). The approach is evaluated both on a song-wise and on a line-wise scale.

The results are highly encouraging and could be used further to perform automatic lyrics alignment and keyword spotting for large databases of songs.

Phoneme, Phone Boundary, and Tone in Automatic Scoring of Mandarin Proficiency

Jiahong Yuan, Mark Liberman; University of Pennsylvania, USA

Sun-P-6-1-2, Time: 10:00

Not every phone, word, or sentence is equally good for assessing language proficiency. We investigated three phonetic factors that may affect automatic scoring of Mandarin proficiency — phoneme, phone boundary, and tone. Results showed that phone boundaries performed the best, and within-syllable boundaries were better than cross-syllable boundaries. The retroflex consonants as well as the vowel following these consonants outperformed the other phonemes. Tone0 and Tone3 outperformed the other tones, and ditone models significantly improved the performance of Tone0. These results suggest that phone boundary models and phoneme- and tone- dependent scoring algorithms should be employed in automatic assessment of Mandarin proficiency. It may also be helpful to separate phoneme and tone scoring prior to the combination of individual scores, as we found that the worst phoneme and the best tone, with respect to automatic scoring of Mandarin proficiency, appeared in the same word.

Tone Classification in Mandarin Chinese Using Convolutional Neural Networks

Charles Chen, Razvan Bunescu, Li Xu, Chang Liu; Ohio University, USA

Sun-P-6-1-3, Time: 10:00

In tone languages, different tone patterns of the same syllable may convey different meanings. Tone perception is important for sentence recognition in noise conditions, especially for children with cochlear implants (CI). We propose a method that fully automates tone classification of syllables in Mandarin Chinese. Our model takes as input the raw tone data and uses convolutional neural networks to classify syllables into one of the four tones in Mandarin. When evaluated on syllables recorded from normal-hearing children, our method achieves substantially higher accuracy compared with previous tone classification techniques based on manually edited F_0. The new approach is also more efficient, as it does not require manual checking of F_0. The new tone classification system could have significant clinical applications in the speech evaluation of the hearing impaired population.

Robust Estimation of Fundamental Frequency Using Single Frequency Filtering Approach

Vishala Pannala, G. Aneeja, Sudarsana Reddy Kadiri, B. Yegnanarayana; IIIT Hyderabad, India

Sun-P-6-1-4, Time: 10:00

A new method for robust estimation of fundamental frequency (F_0) from speech signal is proposed in this paper. The method exploits the high SNR regions of speech in time and frequency domains in the outputs of single frequency filtering (SFF) of speech signal. The high resolution in the frequency domain brings out the harmonic characteristics of speech clearly. The harmonic spacing in the high SNR regions of spectrum determine the F_0. The concept of root cepstrum is used to reduce the effects of vocal tract resonances in the F_0 estimation. The proposed method is evaluated for clean speech and noisy speech simulated for 15 different degradations at different noise levels. Performance of the proposed method is compared with four other standard methods of F_0 extraction. From the results it is evident that the proposed method is robust for most types of degradations.

A Fast and Accurate Fundamental Frequency Estimator Using Recursive Moving Average Filters

Ryunosuke Daido, Yuji Hisaminato; Yamaha, Japan

Sun-P-6-1-5, Time: 10:00

We propose a fundamental frequency (F0) estimation method which is fast, accurate and suitable for real-time use. While the proposed method is based on the same framework as DIO [1, 2], it has two clear differences: it uses RMA (Recursive Moving Average) filters for attenuating high order harmonics, and the period detector is

NOTES

designed to work well even for signals which contain some higher harmonics. Effect of trace-back duration of post-processing was also examined. Evaluation experiments using natural speech databases showed that the accuracy of the proposed method was better than DIO, SWIPE'[3] and YIN [4] and computation speed was the fastest compared to those existing methods.

Frequency Estimation from Waveforms Using Multi-Layered Neural Networks

Prateek Verma, Ronald W. Schafer; Stanford University, USA

Sun-P-6-1-6, Time: 10:00

For frequency estimation in noisy speech or music signals, time domain methods based on signal processing techniques such as autocorrelation or average magnitude difference, often do not perform well. As deep neural networks (DNNs) have become feasible, some researchers have attempted with some success to improve the performance of signal processing based methods by learning on autocorrelation, Fourier transform or constant-Q filter bank based representations. In our approach, blocks of signal samples are input *directly* to a neural network to perform end to end learning. The emergence of sub-harmonic structure in the posterior vector of the output layer, along with analysis of the filter-like structures emerging in the DNN shows strong correlations with some signal processing based approaches. These NNs appear to learn a nonlinearly-spaced frequency representation in the first layer followed by comb-like filters. We find that learning representations from raw time-domain signals can achieve performance on par with the current state of the art algorithms for frequency estimation in noisy and polyphonic settings. The emergence of sub-harmonic structure in the posterior vector suggests that existing post-processing techniques such as harmonic product spectra and salience mapping may further improve the performance.

Sun-P-6-2 : Speaker Diarization and Recognition

Pacific Concourse – Poster B, 10:00–12:00, Sunday, 11 Sept. 2016
Chairs: Thomas Hain, Itshak Lapidot

Speaker Linking and Applications Using Non-Parametric Hashing Methods

Douglas E. Sturim, William M. Campbell; MIT Lincoln Laboratory, USA

Sun-P-6-2-1, Time: 10:00

Large unstructured audio data sets have become ubiquitous and present a challenge for organization and search. One logical approach for structuring data is to find common speakers and link occurrences across different recordings. Prior approaches to this problem have focused on basic methodology for the linking task. In this paper, we introduce a novel trainable non-parametric hashing method for indexing large speaker recording data sets. This approach leads to tunable computational complexity methods for speaker linking. We focus on a scalable clustering method based on hashing — canopy-clustering. We apply this method to a large corpus of speaker recordings, demonstrate performance tradeoffs, and compare to other hashing methods.

Iterative PLDA Adaptation for Speaker Diarization

Gaël Le Lan [1], Delphine Charlet [1], Anthony Larcher [2], Sylvain Meignier [2]; [1] Orange Labs, France; [2] LIUM, France

Sun-P-6-2-2, Time: 10:00

This paper investigates iterative PLDA adaptation for cross-show speaker diarization applied to small collections of French TV archives based on an i-vector framework. Using the target collection itself for unsupervised adaptation, PLDA parameters are iteratively tuned while score normalization is applied for convergence. Performances are compared, using combinations of target and external data for training and adaptation. The experiments on two distinct target corpora show that the proposed framework can gradually improve an existing system trained on external annotated data. Such results indicate that performing speaker diarization on small collections of unlabeled audio archives should only rely on the availability of a sufficient bootstrap system, which can be incrementally adapted to every target collection. The proposed framework also widens the range of acceptable speaker clustering thresholds for a given performance objective.

A Speaker Diarization System for Studying Peer-Led Team Learning Groups

Harishchandra Dubey, Lakshmish Kaushik, Abhijeet Sangwan, John H.L. Hansen; University of Texas at Dallas, USA

Sun-P-6-2-3, Time: 10:00

Peer-led team learning (PLTL) is a model for teaching STEM courses where small student groups meet periodically to collaboratively discuss coursework. Automatic analysis of PLTL sessions would help education researchers to get insight into how learning outcomes are impacted by individual participation, group behavior, team dynamics, *etc.*. Towards this, speech and language technology can help, and speaker diarization technology will lay the foundation for analysis. In this study, a new corpus is established called CRSS-PLTL, that contains speech data from 5 PLTL teams over a semester (10 sessions per team with 5-to-8 participants in each team). In CRSS-PLTL, every participant wears a LENA device (portable audio recorder) that provides multiple audio recordings of the event. Our proposed solution is unsupervised and contains a new online speaker change detection algorithm, termed G^3 algorithm in conjunction with Hausdorff-distance based clustering to provide improved detection accuracy. Additionally, we also exploit cross channel information to refine our diarization hypothesis. The proposed system provides good improvements in diarization error rate (DER) over the baseline LIUM system. We also present higher level analysis such as the number of conversational turns taken in a session, and speaking-time duration (participation) for each speaker.

DNN-Based Speaker Clustering for Speaker Diarisation

Rosanna Milner, Thomas Hain; University of Sheffield, UK

Sun-P-6-2-4, Time: 10:00

Speaker diarisation, the task of answering “who spoke when?”, is often considered to consist of three independent stages: speech activity detection, speaker segmentation and speaker clustering. These represent the separation of speech and non-speech, the splitting into speaker homogeneous speech segments, followed by grouping together those which belong to the same speaker. This paper is concerned with speaker clustering, which is typically performed by bottom-up clustering using the Bayesian information criterion (BIC). We present a novel semi-supervised method of speaker clustering based on a deep neural network (DNN) model. A speaker separation DNN trained on independent data is used to iteratively relabel the test data set. This is achieved by reconfiguration of the output layer, combined with fine tuning in each iteration. A stopping criterion involving posteriors as confidence scores is investigated. Results are shown on a meeting task (RT07) for single distant microphones and compared with standard diarisation approaches. The new method achieves a diarisation error rate (DER) of 14.8%, compared to a baseline of 19.9%.

On the Importance of Efficient Transition Modeling for Speaker Diarization

Itshak Lapidot[1], Jean-François Bonastre[2]; [1]Afeka Tel Aviv Academic College of Engineering, Israel; [2]LIA, France

Sun-P-6-2-5, Time: 10:00

In recent years speaker diarization becomes an important issue. In previous works, we presented the Hidden Distortion Model (HDM) approach, in order to overcome the limitations of traditional HMMs in terms of emission and transition modeling. In this work, we show that HDM allows to build more efficient speaker diarization systems both in terms of diarization error rated and in terms of memory footprint. The best diarization performance is obtained using smaller than usual emission models which constitutes potentially a key advantage for embedded applications with limited memory resources and computational power. A significant memory size reduction was observed using LDC CALLHOME (American) for both SOM- and GMM-based emission probability models.

Priors for Speaker Counting and Diarization with AHC

Gregory Sell, Alan McCree, Daniel Garcia-Romero; Johns Hopkins University, USA

Sun-P-6-2-6, Time: 10:00

Estimating the number of speakers in an audio segment is a necessary step in the process of speaker diarization, but current diarization algorithms do not explicitly define a prior probability on this estimation. This work proposes a process for including priors in speaker diarization with agglomerative hierarchical clustering (AHC). It is also shown that the exclusion of a prior with AHC is itself implicitly a prior, which is found to be geometric growth in the number of speakers. By using more sensible priors, we are able to demonstrate significantly improved robustness to calibration error for speaker counting and speaker diarization.

Two-Pass IB Based Speaker Diarization System Using Meeting-Specific ANN Based Features

Nauman Dawalatabad[1], Srikanth Madikeri[2], Chandra Sekhar C.[1], Hema A. Murthy[1]; [1]IIT Madras, India; [2]Idiap Research Institute, Switzerland

Sun-P-6-2-7, Time: 10:00

In this paper, we present a two-pass Information Bottleneck (IB) based system for speaker diarization which uses meeting-specific artificial neural network (ANN) based features. We first use IB based speaker diarization system to get the labelled speaker segments. These segments are re-segmented using Kullback-Leibler Hidden Markov Model (KL-HMM) based re-segmentation. The multi-layer ANN is then trained to discriminate these speakers using the re-segmented output labels and the spectral features. We then extract the bottleneck features from the trained ANN and perform principal component analysis (PCA) on these features. After performing PCA, these bottleneck features are used along with the different spectral features in the second pass using the same IB based system with KL-HMM re-segmentation. Our experiments on NIST RT and AMI datasets show that the proposed system performs better than the baseline IB system in terms of speaker error rate (SER) with a best case relative improvement of 28.6% amongst AMI datasets and 27.1% on NIST RT04eval dataset.

DNN-Based Amplitude and Phase Feature Enhancement for Noise Robust Speaker Identification

Zeyan Oo[1], Yuta Kawakami[1], Longbiao Wang[1], Seiichi Nakagawa[2], Xiong Xiao[3], Masahiro Iwahashi[1]; [1]Nagaoka University of Technology, Japan; [2]Toyohashi University of Technology, Japan; [3]NTU, Singapore

Sun-P-6-2-8, Time: 10:00

The importance of the phase information of speech signal is gathering attention. Many researches indicate system combination of the amplitude and phase features is effective for improving speaker recognition performance under noisy environments. On the other hand, speech enhancement approach is taken usually to reduce the influence of noises. However, this approach only enhances the amplitude spectrum, therefore noisy phase spectrum is used for reconstructing the estimated signal. Recent years, DNN based feature enhancement is studied intensively for robust speech processing. This approach is expected to be effective also for phase-based feature. In this paper, we propose feature space enhancement of amplitude and phase features using deep neural network (DNN) for speaker identification. We used mel-frequency cepstral coefficients as an amplitude feature, and modified group delay cepstral coefficients as a phase feature. Simultaneous enhancement of amplitude and phase based feature was effective, and it achieved about 24% relative error reduction comparing with individual feature enhancement.

Unit-Selection Attack Detection Based on Unfiltered Frequency-Domain Features

Ulrich Scherhag, Andreas Nautsch, Christian Rathgeb, Christoph Busch; Hochschule Darmstadt, Germany

Sun-P-6-2-9, Time: 10:00

Modern text-to-speech algorithms pose a vital threat to the security of speaker identification and verification (SIV) systems, in terms of subversive usage, i.e. generating presentation attacks. In order to distinguish between presentation attacks and bona fide authentication attempts, presentation attack detection (PAD) subsystems are of utmost importance. Until now, the vast majority of introduced spoofing countermeasures rely on speech production and perception based features. In this paper, we utilize the complete frequency band without further filter-bank processing in order to detect non-smooth transitions in the full and high frequency domain caused by unit-selection attacks. For the purpose of especially detecting unit selection attacks, the applicability of Fast Fourier Transformation (FFT) and Discrete Wavelet Transformation (DWT)

Notes

is examined regarding non-smooth transitions in the full and high frequency domain, excluding filter-bank analyses. Gaussian Mixture Model (GMM) and Support Vector Machine (SVM) classifiers are trained on the German Speech Data Corpus (GSDC) and validated on the standard ASVspoof 2015 corpus resulting in EERs of 7.1% and 11.7%, respectively. Despite language and data shifts, the proposed unit-selection PAD scheme achieves promising biometric performance and hence, introduces a new direction to voice PAD.

Investigating the Impact of Dialect Prestige on Lexical Decision

Mairym Lloréns Monteserín, Jason Zevin; University of Southern California, USA

Sun-P-6-2-10, Time: 10:00

The speech signal encodes both a talker's message and indexical information about a talker's identity. Dialectal variation is one way in which non-linguistic information about a talker is conveyed through her speech. A talker's dialect tends to correlate strongly with her demographic background, and listeners are known to form beliefs about speakers based on their dialect alone: talkers of lower-status dialects are consistently downgraded on positively-valued attributes relative to talkers of canonical dialects. Hypothesizing that pre-formed beliefs about a low-status talker might impact optimal perception of her speech, this study investigated the influence of the relative prestige of talker dialect on listeners' behavior in three lexical decision experiments. The finding of significantly increased propensity to incorrectly reject words uttered in an arguably low-prestige variety of American English relative to both normative General American English and British English suggests that talker status may play a role in the success with which talker messages are perceived by listeners. These results as well as unexpected interactions of dialect and word frequency in some but not all experiments are discussed in the context of signal detection theory.

Speaker Verification Using Short Utterances with DNN-Based Estimation of Subglottal Acoustic Features

Jinxi Guo[1], Gary Yeung[1], Deepak Muralidharan[1], Harish Arsikere[2], Amber Afshan[1], Abeer Alwan[1]; [1]University of California at Los Angeles, USA; [2]Xerox Research Center India, India

Sun-P-6-2-11, Time: 10:00

Speaker verification in real-world applications sometimes deals with limited duration of enrollment and/or test data. MFCC-based i-vector systems have defined the state-of-the-art for speaker verification, but it is well known that they are less effective with short utterances. To address this issue, we propose a method to leverage the speaker specificity and stationarity of subglottal acoustics. First, we present a deep neural network (DNN) based approach to estimate subglottal features from speech signals. The approach involves training a DNN-regression model that maps the log filter-bank coefficients of a given speech signal to those of its corresponding subglottal signal. Cross-validation experiments on the WashU-UCLA corpus (which contains parallel recordings of speech and subglottal acoustics) show the effectiveness of our DNN-based estimation algorithm. The average correlation coefficient between the actual and estimated subglottal filter-bank coefficients is 0.9. A score-level fusion of MFCC and subglottal-feature systems in the i-vector PLDA framework yields statistically-significant improvements over the MFCC-only baseline. On the NIST SRE 08 truncated 10sec–10sec and 5sec–5sec core evaluation tasks, the relative reduction in equal error rate ranges between 6 and 14% for the conditions tested with both microphone and telephone speech.

Factor Analysis Based Speaker Verification Using ASR

Hang Su, Steven Wegmann; ICSI, USA

Sun-P-6-2-12, Time: 10:00

In this paper, we propose to improve speaker verification performance by importing better posterior statistics from acoustic models trained for Automatic Speech Recognition (ASR). This approach aims to introduce state-of-the-art techniques in ASR to speaker verification task. We compare statistics collected from several ASR systems, and show that those collected from deep neural networks (DNN) trained with fMLLR features can effectively reduce equal error rate (EER) by more than 30% on NIST SRE 2010 task, compared with those DNN trained without feature transformations. We also present derivation of factor analysis using variational Bayes inference, and illustrate implementation details of factor analysis and probabilistic linear discriminant analysis (PLDA) in Kaldi recognition toolkit.

Joint Sound Source Separation and Speaker Recognition

Jeroen Zegers, Hugo Van hamme; Katholieke Universiteit Leuven, Belgium

Sun-P-6-2-13, Time: 10:00

Non-negative Matrix Factorization (NMF) has already been applied to learn speaker characterizations from single or non-simultaneous speech for speaker recognition applications. It is also known for its good performance in (blind) source separation for simultaneous speech. This paper explains how NMF can be used to jointly solve the two problems in a multichannel speaker recognizer for simultaneous speech. It is shown how state-of-the-art multichannel NMF for blind source separation can be easily extended to incorporate speaker recognition. Experiments on the CHiME corpus show that this method outperforms the sequential approach of first applying source separation, followed by speaker recognition that uses state-of-the-art i-vector techniques.

Robust Multichannel Gender Classification from Speech in Movie Audio

Naveen Kumar, Md. Nasir, Panayiotis Georgiou, Shrikanth S. Narayanan; University of Southern California, USA

Sun-P-6-2-14, Time: 10:00

Speech in the form of scripted dialogues forms an important part of the audio signal in movies. However, it is often masked by background audio signals such as music, ambient noise or background chatter. These background sounds make even otherwise simple tasks, such as gender classification, challenging. Additionally, the variability in this noise across movies renders standard approaches to source separation or enhancement inadequate. Instead, we exploit multichannel information present in different language channels (English, Spanish, French) for each movie to improve the robustness of our gender classification system. We exploit the fact that the speaker labels of interest in this case co-occur in each language

Notes

channel. We fuse the predictions obtained for each channel using Recognition Output Voting Error Reduction (ROVER) and show that this approach improves the gender accuracy by 7% absolute (11% relative) compared to the best independent prediction on any single channel. In the case of surround movies, we further investigate fusion of mono audio and front center channels which shows 5% and 3% absolute (8% and 4% relative) increase in accuracy compared to only using mono and front center channel, respectively.

Sun-P-6-3 : Speech Synthesis Poster

Pacific Concourse – Poster C, 10:00–12:00, Sunday, 11 Sept. 2016
Chair: Keiichi Tokuda

Recent Advances in Google Real-Time HMM-Driven Unit Selection Synthesizer

Xavi Gonzalvo, Siamak Tazari, Chun-an Chan, Markus Becker, Alexander Gutkin, Hanna Silen; Google, USA

Sun-P-6-3-1, Time: 10:00

This paper presents advances in Google's hidden Markov model (HMM)-driven unit selection speech synthesis system. We describe several improvements to the run-time system; these include minimal latency, high-quality and fast refresh cycle for new voices. Traditionally unit selection synthesizers are limited in terms of the amount of data they can handle and the real applications they are built for. That is even more critical for real-life large-scale applications where high-quality is expected and low latency is required given the available computational resources. In this paper we present an optimized engine to handle a large database at runtime, a composite unit search approach for combining diphones and phrase-based units. In addition a new voice building strategy for handling big databases and keeping the building times low is presented.

First Step Towards End-to-End Parametric TTS Synthesis: Generating Spectral Parameters with Neural Attention

Wenfu Wang, Shuang Xu, Bo Xu; Chinese Academy of Sciences, China

Sun-P-6-3-2, Time: 10:00

In conventional neural networks (NN) based parametric text-to-speech (TTS) synthesis frameworks, text analysis and acoustic modeling are typically processed separately, leading to some limitations. On one hand, much significant human expertise is normally required in text analysis, which presents a laborious task for researchers; on the other hand, training of the NN-based acoustic models still relies on the hidden Markov model (HMM) to obtain frame-level alignments. This acquisition process normally goes through multiple complicated stages. The complex pipeline makes constructing a NN-based parametric TTS system a challenging task. This paper attempts to bypass these limitations using a novel end-to-end parametric TTS synthesis framework, i.e. the text analysis and acoustic modeling are integrated together employing an attention-based recurrent neural network. Thus the alignments can be learned automatically. Preliminary experimental results show that the proposed system can generate moderately smooth spectral parameters and synthesize fairly intelligible speech on short utterances (less than 8 Chinese characters).

The Parameterized Phoneme Identity Feature as a Continuous Real-Valued Vector for Neural Network Based Speech Synthesis

Zhengqi Wen, Ya Li, Jianhua Tao; Chinese Academy of Sciences, China

Sun-P-6-3-3, Time: 10:00

In the speech synthesis systems, the phoneme identity feature indicated as the pronunciation unit is influenced by external contexts like the neighboring words and phonemes. This paper proposes to encode such relatedness and parameterize the pronunciation of the phoneme identity feature as a continuous real-valued vector. The vector, composed by a phoneme embedded vector (PEV) and a word embedded vector (WEV), is applied to substitute the binary vector whose representation is one-hot. It is realized in the word embedding model with the joint training structure where the PEV and WEV are learned together. The effectiveness of the proposed technique was evaluated by comparing it with the binary vector in the bidirectional long short term memory recurrent neural network (BLSTM-RNN) based speech synthesis systems. Improvement on the quality of the synthesized speech has been achieved from the proposed system, which proves the effectiveness of replacing the binary vector with the continuous real-valued vector in describing the phoneme identity feature.

Improved Time-Frequency Trajectory Excitation Vocoder for DNN-Based Speech Synthesis

Eunwoo Song [1], Frank K. Soong [1], Hong-Goo Kang [2]; [1] Microsoft, China; [2] Yonsei University, Korea

Sun-P-6-3-4, Time: 10:00

We investigate an improved time-frequency trajectory excitation (ITFTE) vocoder for deep neural network (DNN)-based statistical parametric speech synthesis (SPSS) systems. The ITFTE is a linear predictive coding-based vocoder, where a pitch-dependent excitation signal is represented by a periodicity distribution in a time-frequency domain. The proposed method significantly improves the parameterization efficiency of ITFTE vocoder for the DNN-based SPSS system, even if its dimension changes due to the inherent nature of pitch variation. By utilizing an orthogonality property of discrete cosine transform, we not only accurately reconstruct the ITFTE parameters but also improve the perceptual quality of synthesized speech. Objective and subjective test results confirm that the proposed method provides superior synthesized speech compared to the previous system.

Voice Quality Control Using Perceptual Expressions for Statistical Parametric Speech Synthesis Based on Cluster Adaptive Training

Yamato Ohtani, Koichiro Mori, Masahiro Morita; Toshiba, Japan

Sun-P-6-3-5, Time: 10:00

This paper describes novel voice quality control of synthetic speech using cluster adaptive training (CAT). In this method, we model voice quality factors labeled with perceptual expressions such as "Gender," "Age" and "Brightness." In advance, we obtain the intensity scores of the perceptual expressions by conducting a listening test, which evaluates differences of voice qualities between synthetic speech of average voice and that of the target. Then we build perceptual

Notes

expression (PE) clusters that we call PE models (PEM) under the conditions that the average voice model is used as the bias cluster and the PE intensity scores are employed as the CAT weights. In synthesis, we can generate controlled synthetic speech by the linear combination of PEMs and the existing speaker's model. Subjective results demonstrate that the proposed method can control the voice qualities with PEs in many cases and the target synthetic speech modified by PEMs achieves comparatively good speech quality.

Waveform Generation Based on Signal Reshaping for Statistical Parametric Speech Synthesis

Felipe Espic, Cassia Valentini-Botinhao, Zhizheng Wu, Simon King; University of Edinburgh, UK

Sun-P-6-3-6, Time: 10:00

We propose a new paradigm of waveform generation for Statistical Parametric Speech Synthesis that is based on neither source-filter separation nor sinusoidal modelling. We suggest that one of the main problems of current vocoding techniques is that they perform an extreme decomposition of the speech signal into source and filter, which is an underlying cause of "buzziness", "musical artifacts", or "muffled sound" in the synthetic speech. The proposed method avoids making unnecessary assumptions and decompositions as far as possible, and uses only the spectral envelope and F0 as parameters. Pre-recorded speech is used as a base signal, which is "reshaped" to match the acoustic specification predicted by the statistical model, without any source-filter decomposition. A detailed description of the method is presented, including implementation details and adjustments. Subjective listening test evaluations of complete DNN-based text-to-speech systems were conducted for two voices: one female and one male. The results show that the proposed method tends to outperform the state-of-the-art standard vocoder STRAIGHT, whilst using fewer acoustic parameters.

Speaker Representations for Speaker Adaptation in Multiple Speakers' BLSTM-RNN-Based Speech Synthesis

Yi Zhao, Daisuke Saito, Nobuaki Minematsu; University of Tokyo, Japan

Sun-P-6-3-7, Time: 10:00

Training a high quality acoustic model with a limited database and synthesizing a new speaker's voice with a few utterances have been hot topics in deep neural network (DNN) based statistical parametric speech synthesis (SPSS). To solve these problems, we built a unified framework for speaker adaptive training as well as speaker adaptation on Bidirectional Long Short-Term Memory with Recurrent Neural Network (BLSTM-RNN) acoustic model. In this paper, we mainly focus on speaker identity control at the input layer of our framework. We have investigated i-vector and speaker code as different speaker representations when used in an augmented input vector, and also propose two approaches to estimate a new speaker's code. Experimental results show that the speaker representations input to the first layer of acoustic model can effectively control speaker identity during speaker adaptive training, thus improving the synthesized speech quality of speakers included in training phase. For speaker adaptation, speaker code estimated from MFCCs can achieve higher preference than other speaker representations.

Fast, Compact, and High Quality LSTM-RNN Based Statistical Parametric Speech Synthesizers for Mobile Devices

Heiga Zen, Yannis Agiomyrgiannakis, Niels Egberts, Fergus Henderson, Przemysław Szczepaniak; Google, UK

Sun-P-6-3-8, Time: 10:00

Acoustic models based on long short-term memory recurrent neural networks (LSTM-RNNs) were applied to statistical parametric speech synthesis (SPSS) and showed significant improvements in naturalness and latency over those based on hidden Markov models (HMMs). This paper describes further optimizations of LSTM-RNN-based SPSS for deployment on mobile devices; weight quantization, multi-frame inference, and robust inference using an ϵ-contaminated Gaussian loss function. Experimental results in subjective listening tests show that these optimizations can make LSTM-RNN-based SPSS comparable to HMM-based SPSS in runtime speed while maintaining naturalness. Evaluations between LSTM-RNN-based SPSS and HMM-driven unit selection speech synthesis are also presented.

An Investigation of DNN-Based Speech Synthesis Using Speaker Codes

Nobukatsu Hojo [1], Yusuke Ijima [1], Hideyuki Mizuno [2]; [1]NTT, Japan; [2]Tokyo University of Science, Japan

Sun-P-6-3-9, Time: 10:00

Recent studies have shown that DNN-based speech synthesis can produce more natural synthesized speech than the conventional HMM-based speech synthesis. However, an open problem remains as to whether the synthesized speech quality can be improved by utilizing a multi-speaker speech corpus. To address this problem, this paper proposes DNN-based speech synthesis using speaker codes as a simple method to improve the performance of the conventional speaker dependent DNN-based method. In order to model speaker variation in the DNN, the augmented feature (speaker codes) is fed to the hidden layer(s) of the conventional DNN. The proposed method trains connection weights of the whole DNN using a multi-speaker speech corpus. When synthesizing a speech parameter sequence, a target speaker is chosen from the corpus and the speaker code corresponding to the selected target speaker is fed to the DNN to generate the speaker's voice. We investigated the relationship between the prediction performance and architecture of the DNNs by changing the input hidden layer for speaker codes. Experimental results showed that the proposed model outperformed the conventional speaker-dependent DNN when the model architecture was set at optimal for the amount of training data of the selected target speaker.

Using Text and Acoustic Features in Predicting Glottal Excitation Waveforms for Parametric Speech Synthesis with Recurrent Neural Networks

Lauri Juvela [1], Xin Wang [2], Shinji Takaki [2], Manu Airaksinen [1], Junichi Yamagishi [2], Paavo Alku [1]; [1]Aalto University, Finland; [2]NII, Japan

Sun-P-6-3-10, Time: 10:00

This work studies the use of deep learning methods to directly model glottal excitation waveforms from context dependent text features in a text-to-speech synthesis system. Glottal vocoding is integrated into a deep neural network-based text-to-speech framework where text

NOTES

and acoustic features can be flexibly used as both network inputs or outputs. Long short-term memory recurrent neural networks are utilised in two stages: first, in mapping text features to acoustic features and second, in predicting glottal waveforms from the text and/or acoustic features. Results show that using the text features directly yields similar quality to the prediction of the excitation from acoustic features, both outperforming a baseline system based on using a fixed glottal pulse for excitation generation.

Model Integration for HMM- and DNN-Based Speech Synthesis Using Product-of-Experts Framework

Kentaro Tachibana[1], *Tomoki Toda*[2], *Yoshinori Shiga*[1], *Hisashi Kawai*[1]; [1]*NICT, Japan;* [2]*Nagoya University, Japan*

Sun-P-6-3-11, Time: 10:00

In this paper, we propose a model integration method for hidden Markov model (HMM) and deep neural network (DNN) based acoustic models using a product-of-experts (PoE) framework in statistical parametric speech synthesis. In speech parameter generation, DNN predicts a mean vector of the probability density function of speech parameters frame by frame while keeping its covariance matrix constant over all frames. On the other hand, HMM predicts the covariance matrix as well as the mean vector but they are fixed within the same HMM state, i.e., they can actually vary state by state. To make it possible to predict a better probability density function by leveraging advantages of individual models, the proposed method integrates DNN and HMM as PoE, generating a new probability density function satisfying conditions of both DNN and HMM. Furthermore, we propose a joint optimization method of DNN and HMM within the PoE framework by effectively using additional latent variables. We conducted objective and subjective evaluations, demonstrating that the proposed method significantly outperforms the DNN-based speech synthesis as well as the HMM-based speech synthesis.

Idlak Tangle: An Open Source Kaldi Based Parametric Speech Synthesiser Based on DNN

Blaise Potard[1], *Matthew P. Aylett*[1], *David A. Baude*[1], *Petr Motlicek*[2]; [1]*CereProc, UK;* [2]*Idiap Research Institute, Switzerland*

Sun-P-6-3-12, Time: 10:00

This paper presents a text to speech (TTS) extension to Kaldi — a liberally licensed open source speech recognition system. The system, Idlak Tangle, uses recent deep neural network (DNN) methods for modelling speech, the Idlak XML based text processing system as the front end, and a newly released open source mixed excitation MLSA vocoder included in Idlak. The system has none of the licensing restrictions of current freely available HMM style systems, such as the HTS toolkit. To date no alternative open source DNN systems are available. Tangle combines the Idlak front-end and vocoder, with two DNNs modelling respectively the units duration and acoustic parameters, providing a fully functional end-to-end TTS system.

Experimental results using the freely available SLT speaker from CMU ARCTIC, reveal that the speech output is rated in a MUSHRA test as significantly more natural than the output of HTS-demo, the only other free to download HMM system available with no commercially restricted or proprietary IP. The tools, audio database and recipe required to reproduce the results presented in these paper are fully available online.

Probabilistic Amplitude Demodulation Features in Speech Synthesis for Improving Prosody

Alexandros Lazaridis, Milos Cernak, Philip N. Garner; Idiap Research Institute, Switzerland

Sun-P-6-3-13, Time: 10:00

Amplitude demodulation (AM) is a signal decomposition technique by which a signal can be decomposed to a product of two signals, i.e, a quickly varying carrier and a slowly varying modulator. In this work, the probabilistic amplitude demodulation (PAD) features are used to improve prosody in speech synthesis. The PAD is applied iteratively for generating syllable and stress amplitude modulations in a cascade manner. The PAD features are used as a secondary input scheme along with the standard text-based input features in statistical parametric speech synthesis. Specifically, deep neural network (DNN)-based speech synthesis is used to evaluate the importance of these features. Objective evaluation has shown that the proposed system using the PAD features has improved mainly prosody modelling; it outperforms the baseline system by approximately 5% in terms of relative reduction in root mean square error (RMSE) of the fundamental frequency (F0). The significance of this improvement is validated by subjective evaluation of the overall speech quality, achieving 38.6% over 19.5% preference score in respect to the baseline system, in an ABX test.

On Smoothing and Enhancing Dynamics of Pitch Contours Represented by Discrete Orthogonal Polynomials for Prosody Generation

Chen-Yu Chiang; National Taipei University, Taiwan

Sun-P-6-3-14, Time: 10:00

This paper presents a new pitch contour generation algorithm for statistical syllable-based logF0 generation models which represent logF0 contours of syllables by coefficients of discrete orthogonal polynomials, i.e. orthogonal expansion coefficients (OECs). The conventional statistical logF0 models can generate smooth pitch contour within a syllable because of the continuity property of polynomials. However, the models do not ensure to produce continuous and smooth logF0 contours in the proximity of syllable junctures. Besides, dynamic range of the generated logF0 contours is generally smaller than the one of real speech. The above two shortcomings would result in unnatural and monotonous prosody. To overcome these shortcomings, juncture-smooth and dynamics-enhancing OEC generation algorithms are hence proposed in this paper. Analysis on the generated logF0 contours by the proposed algorithm shows some improvements in logF0 smoothness at syllable junctures and enhanced logF0 dynamic range. In addition, a perceptual evaluation of the logF0 contour generated by the proposed algorithm shows an improvement in naturalness of the synthesized speech.

An Investigation of Recurrent Neural Network Architectures Using Word Embeddings for Phrase Break Prediction

Anandaswarup Vadapalli, Suryakanth V. Gangashetty; IIIT Hyderabad, India

Sun-P-6-3-15, Time: 10:00

This paper presents our investigations of recurrent neural networks (RNNs) for the phrase break prediction task. With the advent of deep learning, there have been attempts to apply deep neural networks

NOTES

(DNNs) to phrase break prediction. While deep neural networks are able to effectively capture dependencies across features, they lack the ability to capture long-term relations that are spread over time. On the other hand, RNNs are able to capture long-term temporal relations and thus are better suited for tasks where sequences have to be modeled. We model the phrase break prediction task as a sequence labeling task, and show by means of experimental results that RNNs perform better at phrase break prediction as compared to conventional DNN systems.

Model-Based Parametric Prosody Synthesis with Deep Neural Network

Hao Liu [1], Heng Lu [2], Xu Shao [2], Yi Xu [1]; [1] University College London, UK; [2] Nuance Communications, USA

Sun-P-6-3-16, Time: 10:00

Conventional statistical parametric speech synthesis (SPSS) captures only frame-wise acoustic observations and computes probability densities at HMM state level to obtain statistical acoustic models combined with decision trees, which is therefore a purely statistical data-driven approach without explicit integration of any articulatory mechanisms found in speech production research. The present study explores an alternative paradigm, namely, model-based parametric prosody synthesis (MPPS), which integrates dynamic mechanisms of human speech production as a core component of F0 generation. In this paradigm, contextual variations in prosody are processed in two separate yet integrated stages: linguistic to motor, and motor to acoustic. Here the motor model is target approximation (TA), which generates syllable-sized F0 contours with only three motor parameters that are associated to linguistic functions. In this study, we simulate this two-stage process by linking the TA model to a deep neural network (DNN), which learns the “linguistic-motor” mapping given the “motor-acoustic” mapping provided by TA-based syllable-wise F0 production. The proposed prosody modeling system outperforms the HMM-based baseline system in both objective and subjective evaluations.

Sun-P-6-4 : Language Model Adaptation

Pacific Concourse – Poster D, 10:00–12:00, Sunday, 11 Sept. 2016
Chair: Ilya Oparin

Active and Semi-Supervised Learning in ASR: Benefits on the Acoustic and Language Models

Thomas Drugman [1], Janne Pylkkönen [2], Reinhard Kneser [1]; [1] Amazon.com, Germany; [2] Amazon.com, Finland

Sun-P-6-4-1, Time: 10:00

The goal of this paper is to simulate the benefits of jointly applying active learning (AL) and semi-supervised training (SST) in a new speech recognition application. Our data selection approach relies on confidence filtering, and its impact on both the acoustic and language models (AM and LM) is studied. While AL is known to be beneficial to AM training, we show that it also carries out substantial improvements to the LM when combined with SST. Sophisticated confidence models, on the other hand, did not prove to yield any data selection gain. Our results indicate that, while SST is crucial at the beginning of the labeling process, its gains degrade rapidly as AL is set in place. The final simulation reports that AL allows a transcription cost reduction of about 70% over random selection. Alternatively, for a fixed transcription budget, the proposed approach improves the word error rate by about 12.5% relative.

Learning N-Gram Language Models from Uncertain Data

Vitaly Kuznetsov [1], Hank Liao [2], Mehryar Mohri [1], Michael Riley [2], Brian Roark [2]; [1] New York University, USA; [2] Google, USA

Sun-P-6-4-2, Time: 10:00

We present a new algorithm for efficiently training n-gram language models on uncertain data, and illustrate its use for semi-supervised language model adaptation. We compute the probability that an n-gram occurs k times in the sample of uncertain data, and use the resulting histograms to derive a generalized Katz back-off model. We compare three approaches to semi-supervised adaptation of language models for speech recognition of selected YouTube video categories: (1) using just the one-best output from the baseline speech recognizer or (2) using samples from lattices with standard algorithms versus (3) using full lattices with our new algorithm. Unlike the other methods, our new algorithm provides models that yield solid improvements over the baseline on the full test set, and, further, achieves these gains without hurting performance on any of the set of video categories. We show that categories with the most data yielded the largest gains. The algorithm has been released as part of the OpenGrm n-gram library [1].

Entropy Based Pruning for Non-Negative Matrix Based Language Models with Contextual Features

Barlas Oğuz, Issac Alphonso, Shuangyu Chang; Microsoft, USA

Sun-P-6-4-3, Time: 10:00

Non-negative matrix based language models have been recently introduced [1] as a computationally efficient alternative to other feature-based models such as maximum-entropy models. We present a new entropy based pruning algorithm for this class of language models, which is fast and scalable. We present perplexity and word error rate results and compare these against regular n-gram pruning. We also train models with location and personalization features and report results at various pruning thresholds. We demonstrate that contextual features are helpful over the vanilla model even after pruning to a similar size.

Unsupervised Adaptation of Recurrent Neural Network Language Models

Siva Reddy Gangireddy, Pawel Swietojanski, Peter Bell, Steve Renals; University of Edinburgh, UK

Sun-P-6-4-4, Time: 10:00

Recurrent neural network language models (RNNLMs) have been shown to consistently improve Word Error Rates (WERs) of large vocabulary speech recognition systems employing n-gram LMs. In this paper we investigate supervised and unsupervised discriminative adaptation of RNNLMs in a broadcast transcription task to target domains defined by either genre or show. We have explored two approaches based on (1) scaling forward-propagated hidden activations (Learning Hidden Unit Contributions (LHUC) technique) and (2) direct fine-tuning of the parameters of the whole RNNLM.

NOTES

To investigate the effectiveness of the proposed methods we carry out experiments on multi-genre broadcast (MGB) data following the MGB-2015 challenge protocol. We observe small but significant improvements in WER compared to a strong unadapted RNNLM model.

Contextual Prediction Models for Speech Recognition

Yoni Halpern[1], Keith Hall[1], Vlad Schogol[1], Michael Riley[1], Brian Roark[1], Gleb Skobeltsyn[2], Martin Bäuml[2]; [1]Google, USA; [2]Google, Switzerland

Sun-P-6-4-5, Time: 10:00

We introduce an approach to biasing language models towards known contexts without requiring separate language models or explicit contextually-dependent conditioning contexts. We do so by presenting an alternative ASR objective, where we predict the acoustics and words given the contextual cue, such as the geographic location of the speaker. A simple factoring of the model results in an additional *biasing* term, which effectively indicates how correlated a hypothesis is with the contextual cue (e.g., given the hypothesized transcript, how likely is the user's known location). We demonstrate that this factorization allows us to train relatively small contextual models which are effective in speech recognition. An experimental analysis shows a perplexity reduction of up to 35% and a relative reduction in word error rate of 1.6% on a targeted voice search dataset when using the user's coarse location as a contextual cue.

Combining Feature and Model-Based Adaptation of RNNLMs for Multi-Genre Broadcast Speech Recognition

Salil Deena, Madina Hasan, Mortaza Doulaty, Oscar Saz, Thomas Hain; University of Sheffield, UK

Sun-P-6-4-6, Time: 10:00

Recurrent neural network language models (RNNLMs) have consistently outperformed n-gram language models when used in automatic speech recognition (ASR). This is because RNNLMs provide robust parameter estimation through the use of a continuous-space representation of words, and can generally model longer context dependencies than n-grams. The adaptation of RNNLMs to new domains remains an active research area and the two main approaches are: feature-based adaptation, where the input to the RNNLM is augmented with auxiliary features; and model-based adaptation, which includes model fine-tuning and introduction of adaptation layer(s) in the network. This paper explores the properties of both types of adaptation on multi-genre broadcast speech recognition. Two hybrid adaptation techniques are proposed, namely the fine-tuning of feature-based RNNLMs and the use of a feature-based adaptation layer. A method for the semi-supervised adaptation of RNNLMs, using topic model-based genre classification, is also presented and investigated. The gains obtained with RNNLM adaptation on a system trained on 700h. of speech are consistent using both RNNLMs trained on a small (10Mwords) and large set (660M words), with 10% perplexity and 2% word error rate improvements on a 28.3h. test set.

Sun-S&T-6 : Show & Tell Session 6

Market Street Foyer, 10:00–12:00, Sunday, 11 Sept. 2016
Chairs: Nicolas Scheffer, Shiva Sundaram

A Low Cost Desktop Robot and Tele-Presence Device for Interactive Speech Research

Michael C. Brady; Tufts University, USA

Sun-S&T-6-1, Time: 10:00

In building robotic systems that interact with people through speech, many robotics engineers are obliged to treat artificial speech recognition and synthesis as a black-box problem best left to speech engineers to solve. Yet speech engineers today typically do not have access to the kinds of expensive robots needed for this development. Progress on the human-robot speech interface thus suffers from something of a diffusion of responsibility. In an attempt to remedy the situation, we have developed a low-cost interactive embodied speech device. The device is constructed from off-the-shelf components and from 3D-printed and laser-cut parts. We make the files for the 3D and laser-cut parts freely available for download. In addition to offering basic assembled devices and kits for self-assembly, we provide an assembly guide and a shopping list of components a user will need in order to build, maintain, and customize their own device. We supply a basic software framework (in both Matlab and in C/C++), and template code for a ROS node for interfacing with the device. The idea is to establish a standard and accessible hardware platform with an open-source foundation for the sharing of ideas and research.

Silent-Speech Command Word Recognition Using Electro-Optical Stomatography

Simon Stone, Peter Birkholz; Technische Universität Dresden, Germany

Sun-S&T-6-2, Time: 10:00

In this paper that accompanies a live Show & Tell demonstration at INTERSPEECH 2016, we present our current speaker-dependent silent-speech recognition system. Silent-speech recognition refers to the recognition of speech without any acoustic data. To that end, our system uses a novel technique called electro-optical stomatography to record the tongue and lip movements of a subject during the articulation of a set of isolated words in real-time. Based on these data, simple articulatory models are learned. The system then classifies unseen articulatory data of learned isolated words spoken by the same subject. This paper presents the system components and showcases the silent-speech recognition process with a set of the 30 most common German words. Since the system is language-independent and easy to train, the demonstration will also show both training and recognition of any other words on demand.

An Engine for Online Video Search in Large Archives of the Holocaust Testimonies

Petr Stanislav, Jan Švec, Pavel Ircing; University of West Bohemia, Czech Republic

Sun-S&T-6-3, Time: 10:00

In this paper we present an online system for cross-lingual lexical (full-text) searching in the large archive of the Holocaust testimonies.

Video interviews recorded in two languages (English and Czech) were automatically transcribed and indexed in order to provide

NOTES

efficient access to the lexical content of the recordings. The engine takes advantage of the state-of-the-art speech recognition system and performs fast spoken term detection (STD), providing direct access to the segments of interviews containing queried words or short phrases.

Sun-O-7-1 : Robustness in Speech Processing

Grand Ballroom A, 13:30–15:30, Sunday, 11 Sept. 2016
Chairs: Michael Seltzer, Ozlem Kalinli

Data Selection by Sequence Summarizing Neural Network in Mismatch Condition Training

Kateřina Žmolíková[1], Martin Karafiát[1], Karel Veselý[1], Marc Delcroix[2], Shinji Watanabe[3], Lukáš Burget[1], Jan Černocký[1]; [1]Brno University of Technology, Czech Republic; [2]NTT, Japan; [3]MERL, USA

Sun-O-7-1-1, Time: 13:30

Data augmentation is a simple and efficient technique to improve the robustness of a speech recognizer when deployed in mismatched training-test conditions. Our paper proposes a new approach for selecting data with respect to similarity of acoustic conditions. The similarity is computed based on a sequence summarizing neural network which extracts vectors containing acoustic summary (e.g. noise and reverberation characteristics) of an utterance. Several configurations of this network and different methods of selecting data using these “summary-vectors” were explored. The results are reported on a mismatched condition using AMI training set with the proposed data selection and CHiME3 test set.

Incorporating a Generative Front-End Layer to Deep Neural Network for Noise Robust Automatic Speech Recognition

Souvik Kundu[1], Khe Chai Sim[1], Mark J.F. Gales[2]; [1]NUS, Singapore; [2]University of Cambridge, UK

Sun-O-7-1-2, Time: 13:50

It is difficult to apply well-formulated model-based noise adaptation approaches to Deep Neural Network (DNN) due to the lack of interpretability of the model parameters. In this paper, we propose incorporating a generative front-end layer (GFL), which is parameterised by Gaussian Mixture Model (GMM), into the DNN. A GFL can be easily adapted to different noise conditions by applying the model-based Vector Taylor Series (VTS) to the underlying GMM. We show that incorporating a GFL to DNN yields 12.1% relative improvement over a baseline multi-condition DNN. We also show that the proposed system performs significantly better than the noise aware training method, where the per-utterance estimated noise parameters are appended to the acoustic features.

Robust Speech Recognition Using Generalized Distillation Framework

Konstantin Markov[1], Tomoko Matsui[2]; [1]University of Aizu, Japan; [2]ISM, Japan

Sun-O-7-1-3, Time: 14:10

In this paper, we propose a noise robust speech recognition system built using generalized distillation framework. It is assumed that during training, in addition to the training data, some kind of “privileged” information is available and can be used to guide the training process. This allows to obtain a system which at test time outperforms those built on regular training data alone. In the case of noisy speech recognition task, the privileged information is obtained from a model, called “teacher”, trained on clean speech only. The regular model, called “student”, is trained on noisy utterances and uses teacher’s output for the corresponding clean utterances. Thus, for this framework a parallel clean/noisy speech data are required. We experimented on the Aurora2 database which provides such kind of data. Our system uses hybrid DNN-HMM acoustic model where neural networks provide HMM state probabilities during decoding. The teacher DNN is trained on the clean data, while the student DNN is trained using multi-condition (various SNRs) data. The student DNN loss function combines the targets obtained from forced alignment of the training data and the outputs of the teacher DNN when fed with the corresponding clean features. Experimental results clearly show that distillation framework is effective and allows to achieve significant reduction in the word error rate.

Adversarial Multi-Task Learning of Deep Neural Networks for Robust Speech Recognition

Yusuke Shinohara; Toshiba, Japan

Sun-O-7-1-4, Time: 14:30

A method of learning deep neural networks (DNNs) for noise robust speech recognition is proposed. It is widely known that representations (activations) of well-trained DNNs are highly invariant to noise, especially in higher layers, and such invariance leads to the noise robustness of DNNs. However, little is known about how to enhance such invariance of representations, which is a key for improving robustness. In this paper, we propose adversarial multi-task learning of DNNs for explicitly enhancing the invariance of representations. Specifically, a primary task of senone classification and a secondary task of domain (noise condition) classification are jointly solved. What is different from the standard multi-task learning is that the representation is learned adversarially to the secondary task, so that representation with low domain-classification accuracy is induced. As a result, senone-discriminative and domain-invariant representation is obtained, which leads to an improved robustness of DNNs. Experimental results on a noise-corrupted Wall Street Journal data set show the effectiveness of the proposed method.

The Use of Locally Normalized Cepstral Coefficients (LNCC) to Improve Speaker Recognition Accuracy in Highly Reverberant Rooms

Víctor Poblete[1], Juan Pablo Escudero[1], Josué Fredes[2], José Novoa[2], Richard M. Stern[3], Simon King[4], Néstor Becerra Yoma[2]; [1]Universidad Austral de Chile, Chile; [2]Universidad de Chile, Chile; [3]Carnegie Mellon University, USA; [4]University of Edinburgh, UK

Sun-O-7-1-5, Time: 14:50

We describe the ability of LNCC features (Locally Normalized Cepstral Coefficients) to improve speaker recognition accuracy in highly reverberant environments. We used a realistic test environment, in which we changed the number and nature of reflective surfaces in the room, creating four increasingly reverberant times from approximately 1 to 9 seconds. In this room, we re-recorded reverberated versions of the Yoho speaker verification corpus. The recordings were made using four speaker-to-microphone distances, from 0.32m to 2.56m. Experimental results for a speaker verification task suggest that LNCC features are an attractive alternative to MFCC features under such reverberant conditions, as they were observed to improve verification accuracy compared to baseline MFCC features in all cases where the reverberation time exceeded 1 second or with a greater speaker-microphone distance (i.e. 2.56 m).

Two-Stage Data Augmentation for Low-Resourced Speech Recognition

William Hartmann, Tim Ng, Roger Hsiao, Stavros Tsakalidis, Richard Schwartz; Raytheon BBN Technologies, USA

Sun-O-7-1-6, Time: 15:10

Low resourced languages suffer from limited training data and resources. Data augmentation is a common approach to increasing the amount of training data. Additional data is synthesized by manipulating the original data with a variety of methods. Unlike most previous work that focuses on a single technique, we combine multiple, complementary augmentation approaches. The first stage adds noise and perturbs the speed of additional copies of the original audio. The data is further augmented in a second stage, where a novel fMLLR-based augmentation is applied to bottleneck features to further improve performance. A reduction in word error rate is demonstrated on four languages from the IARPA Babel program. We present an analysis exploring why these techniques are beneficial.

Sun-O-7-2 : Special Session: Interspeech 2016 Computational Paralinguistics Challenge (ComParE): Deception, Sincerity & Native Language

Grand Ballroom BC, 13:30–15:30, Sunday, 11 Sept. 2016

Chairs: Björn Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee Burgoon, Eduardo Coutinho

The Native Language Sub-Challenge: The Data

Björn Schuller[1], Stefan Steidl[2], Anton Batliner[3], Julia Hirschberg[4], Judee K. Burgoon[5], Alice Baird[4], Aaron Elkins[5], Yue Zhang[1], Eduardo Coutinho[1], Keelan Evanini[6]; [1]Imperial College London, UK; [2]FAU Erlangen-Nürnberg, Germany; [3]Universität Passau, Germany; [4]Columbia University, USA; [5]University of Arizona, USA; [6]Educational Testing Service, USA

Sun-O-7-2-1, Time: 13:30

(No abstract available at the time of publication)

Native Language Identification Using Spectral and Source-Based Features

Avni Rajpal[1], Tanvina B. Patel[1], Hardik B. Sailor[1], Maulik C. Madhavi[1], Hemant A. Patil[1], Hiroya Fujisaki[2]; [1]DA-IICT, India; [2]University of Tokyo, Japan

Sun-O-7-2-2, Time: 13:40

The task of native language (L1) identification from non-native language (L2) can be thought of as the task of identifying the common traits that each group of L1 speakers maintains while speaking L2 irrespective of the dialect or region. Under the assumption that speakers are L1 proficient, non-native cues in terms of segmental and prosodic aspects are investigated in our work. In this paper, we propose the use of longer duration cepstral features, namely, Mel frequency cepstral coefficients (MFCC) and auditory filterbank features learnt from the database using Convolutional Restricted Boltzmann Machine (ConvRBM) along with their delta and shifted delta features. MFCC and ConvRBM gave accuracy of 38.2% and 36.8%, respectively, on the development set provided for the ComParE 2016 Nativeness Task using Gaussian Mixture Model (GMM) classifier. To add complementary information about the prosodic and excitation source features, phrase information and its dynamics extracted from the $\log(F_0)$ contour of the speech was explored. The accuracy obtained using score-level fusion between system features (MFCC and ConvRBM) and phrase features were 39.6% and 38.3%, respectively, indicating that phrase information and MFCC capture complementary information than ConvRBM alone. Furthermore, score-level fusion of MFCC, ConvRBM and phrase improves the accuracy to 40.2%.

Accent Identification by Combining Deep Neural Networks and Recurrent Neural Networks Trained on Long and Short Term Features

Yishan Jiao, Ming Tu, Visar Berisha, Julie Liss; Arizona State University, USA

Sun-O-7-2-3, Time: 13:50

Automatic identification of foreign accents is valuable for many speech systems, such as speech recognition, speaker identification, voice conversion, etc. The INTERSPEECH 2016 Native Language Sub-Challenge is to identify the native languages of non-native English speakers from eleven countries. Since differences in accent are due to both prosodic and articulation characteristics, a combination of long-term and short-term training is proposed in this paper. Each speech sample is processed into multiple speech segments with equal length. For each segment, deep neural networks (DNNs) are used to train on long-term statistical features, while recurrent neural networks (RNNs) are used to train on short-term acoustic features. The result for each speech sample is calculated by linearly fusing the results from the two sets of networks on all segments. The performance of the proposed system greatly surpasses the provided baseline system. Moreover, by fusing the results with the baseline system, the performance can be further improved.

Convolutional Neural Networks with Data Augmentation for Classifying Speakers' Native Language

Gil Keren, Jun Deng, Jouni Pohjalainen, Björn Schuller; Universität Passau, Germany

Sun-O-7-2-4, Time: 14:00

We use a feedforward Convolutional Neural Network to classify speakers' native language for the INTERSPEECH 2016 Computational Paralinguistic Challenge Native Language Sub-Challenge, using no specialized features for computational paralinguistics tasks, but only MFCCs with their first and second order deltas. In addition, we augment the training data by replacing the original examples with shorter overlapping samples extracted from them, thus multiplying the number of training examples by almost 40. With the augmented training dataset and enhancements to neural network models such as Batch Normalization, Dropout, and Maxout activation function, we managed to improve upon the challenge baseline by a large margin, both for the development and the test set.

NOTES

Native Language Detection Using the I-Vector Framework

Mohammed Senoussaoui[1], Patrick Cardinal[1], Najim Dehak[2], Alessandro L. Koerich[1]; [1]École de Technologie Supérieure, Canada; [2]Johns Hopkins University, USA

Sun-O-7-2-5, Time: 14:10

Native-language identification is the task of determining a speaker's native language based only on their speeches in a second language. In this paper we propose the use of the well-known i-vector representation of the speech signal to detect the native language of an English speaker. The i-vector representation has shown an excellent performance on the quite similar task of distinguishing between different languages. We have evaluated different ways to extract i-vectors in order to adapt them to the specificities of the native language detection task. The experimental results on the 2016 ComParE Native language sub-challenge test set have shown that the proposed system based on a conventional i-vector extractor outperforms the baseline system with a 42% relative improvement.

Within-Speaker Features for Native Language Recognition in the Interspeech 2016 Computational Paralinguistics Challenge

Mark Huckvale; University College London, UK

Sun-O-7-2-6, Time: 14:20

The Interspeech 2016 Native Language recognition challenge was to identify the first language of 867 speakers from their spoken English. Effectively this was an L2 accent recognition task where the L1 was one of eleven languages. The lack of transcripts of the spontaneous speech recordings meant that the currently best performing accent recognition approach (ACCDIST) developed by the author could not be applied. Instead, the objectives of this study were to explore whether within-speaker features found to be effective in ACCDIST would also have value within a contemporary GMM-based accent recognition approach. We show that while Gaussian mean supervectors provide the best performance on this task, small gains may be had by fusing the mean supervector system with a system based on within-speaker Gaussian mixture distances.

Multimodal Fusion of Multirate Acoustic, Prosodic, and Lexical Speaker Characteristics for Native Language Identification

Prashanth Gurunath Shivakumar, Sandeep Nallan Chakravarthula, Panayiotis Georgiou; University of Southern California, USA

Sun-O-7-2-7, Time: 14:30

Native language identification from acoustic signals of L2 speakers can be useful in a range of applications such as informing automatic speech recognition (ASR), speaker recognition, and speech biometrics. In this paper we follow a multi-stream and multi-rate approach, for native language identification, in feature extraction, classification, and fusion. On the feature front we employ acoustic features such as MFCC and PLP features, at different time scales and different transformations; we evaluate speaker normalization as a feature and as a transform; investigate phonemic confusability and its interplay with paralinguistic cues at both the frame and phone-level temporal scales; and automatically extract lexical features; in addition to baseline features. On the classification side we employ SVM, i-Vector, DNN and bottleneck features, and maximum-likelihood models. Finally we employ fusion for system combination and analyze the complementarity of the individual systems. Our proposed system significantly outperforms the baseline system on both development and test sets.

Exploiting Phone Log-Likelihood Ratio Features for the Detection of the Native Language of Non-Native English Speakers

Alberto Abad, Eugénio Ribeiro, Fábio Kepler, Ramon Astudillo, Isabel Trancoso; INESC-ID Lisboa, Portugal

Sun-O-7-2-8, Time: 14:40

Detecting the native language (L1) of non-native English speakers may be of great relevance in some applications, such as computer assisted language learning or IVR services. In fact, the L1 detection problem closely resembles the problem of spoken language and dialect recognition. In particular, log-likelihood ratios of phone posterior probabilities, known as Phone LogLikelihood Ratios (PLLR), have been recently introduced as features for spoken language recognition systems. This representation has proven to be an effective way of retrieving acoustic-phonotactic information at frame-level, which allows for its use in state-of-the-art systems, that is, in i-vector systems. In this paper, we explore the use of PLLR-based i-vector systems for L1 native language detection. We also investigate several linear and non-linear L1 classification schemes on top of the PLLR i-vector front-ends. Moreover, we compare PLLR based systems with both conventional phonotactic systems based on n-gram modelling of phoneme sequences and acoustic-based i-vector systems. Finally, the potential complementarity of the different approaches is investigated based on a set of system fusion experiments.

Determining Native Language and Deception Using Phonetic Features and Classifier Combination

Gábor Gosztolya[1], Tamás Grósz[1], Róbert Busa-Fekete[2], László Tóth[3]; [1]University of Szeged, Hungary; [2]Universität Paderborn, Germany; [3]MTA-SZTE RGAI, Hungary

Sun-O-7-2-9, Time: 14:50

For several years, the Interspeech ComParE Challenge has focused on paralinguistic tasks of various kinds. In this paper we focus on the Native Language and the Deception sub-challenges of ComParE 2016, where the goal is to identify the native language of the speaker, and to recognize deceptive speech. As both tasks can be treated as classification ones, we experiment with several state-of-the-art machine learning methods (Support-Vector Machines, AdaBoost. MH and Deep Neural Networks), and also test a simple-yet-robust combination method. Furthermore, we will assume that the native language of the speaker affects the pronunciation of specific phonemes in the language he is currently using. To exploit this, we extract phonetic features for the Native Language task. Moreover, for the Deception Sub-Challenge we compensate for the highly unbalanced class distribution by instance re-sampling. With these techniques we are able to significantly outperform the baseline SVM on the unpublished test set.

Notes

The INTERSPEECH 2016 Computational Paralinguistics Challenge: A Summary of Results

Björn Schuller[1], Stefan Steidl[2], Anton Batliner[3], Julia Hirschberg[4], Judee K. Burgoon[5], Alice Baird[4], Aaron Elkins[5], Yue Zhang[1], Eduardo Coutinho[1], Keelan Evanini[6]; [1]Imperial College London, UK; [2]FAU Erlangen-Nürnberg, Germany; [3]Universität Passau, Germany; [4]Columbia University, USA; [5]University of Arizona, USA; [6]Educational Testing Service, USA

Sun-O-7-2-10, Time: 15:00

(No abstract available at the time of publication)

Discussion

Björn Schuller[1], Stefan Steidl[2], Anton Batliner[3], Julia Hirschberg[4], Judee K. Burgoon[5], Alice Baird[4], Aaron Elkins[5], Yue Zhang[1], Eduardo Coutinho[1], Keelan Evanini[6]; [1]Imperial College London, UK; [2]FAU Erlangen-Nürnberg, Germany; [3]Universität Passau, Germany; [4]Columbia University, USA; [5]University of Arizona, USA; [6]Educational Testing Service, USA

Sun-O-7-2-11, Time: 15:10

(No abstract available at the time of publication)

Sun-O-7-3 : Acoustic and Articulatory Phonetics

Bayview A, 13:30–15:30, Sunday, 11 Sept. 2016
Chairs: Francisco Lacerda, Mary Beckman

A Preliminary Ultrasound Study of Nasal and Lateral Coronals in Arrernte

Marija Tabain[1], Richard Beare[2]; [1]La Trobe University, Australia; [2]Monash University, Australia

Sun-O-7-3-1, Time: 13:30

Ultrasound tongue image data are presented for two female speakers of the Central Australian language Arrernte, focusing on the nasal and lateral coronal consonants. These coronal places of articulation are: dental, alveolar, retroflex and palatal. It is shown that the tongue back is particularly far forward for the palatal consonant, and to a lesser extent also for the retroflex consonant. There is a general flattening of the tongue for the dental consonants. In addition, the back of the tongue is consistently further forward for the nasal consonants than for the laterals — this is true for all places of articulation. Finally, a double-pivot pattern for the retroflex articulations is observed for one speaker, but not for the other.

Illustrating the Production of the International Phonetic Alphabet Sounds Using Fast Real-Time Magnetic Resonance Imaging

Asterios Toutios[1], Sajan Goud Lingala[1], Colin Vaz[1], Jangwon Kim[1], John Esling[2], Patricia Keating[3], Matthew Gordon[4], Dani Byrd[1], Louis Goldstein[1], Krishna S. Nayak[1], Shrikanth S. Narayanan[1]; [1]University of Southern California, USA; [2]University of Victoria, Canada; [3]University of California at Los Angeles, USA; [4]University of California at Santa Barbara, USA

Sun-O-7-3-2, Time: 13:50

Recent advances in real-time magnetic resonance imaging (rtMRI) of the upper airway for acquiring speech production data provide unparalleled views of the dynamics of a speaker's vocal tract at very high frame rates (83 frames per second and even higher). This paper introduces an effort to collect and make available on-line rtMRI data corresponding to a large subset of the sounds of the world's languages as encoded in the International Phonetic Alphabet, with supplementary English words and phonetically-balanced texts, produced by four prominent phoneticians, using the latest rtMRI technology. The technique images oral as well as laryngeal articulator movements in the production of each sound category. This resource is envisioned as a teaching tool in pronunciation training, second language acquisition, and speech therapy.

Marginal Contrast Among Romanian Vowels: Evidence from ASR and Functional Load

Margaret E.L. Renwick[1], Ioana Vasilescu[2], Camille Dutrey[3], Lori Lamel[2], Bianca Vieru[4]; [1]University of Georgia, USA; [2]LIMSI, France; [3]LNE, France; [4]Vocapia Research, France

Sun-O-7-3-3, Time: 14:10

This work quantifies the phonological contrast between the Romanian central vowels [ʌ] and [ɨ], which are considered separate phonemes, although they are historical allophones with few minimal pairs. We consider the vowels' functional load within the Romanian inventory and the usefulness of the contrast for automatic speech recognition (ASR). Using a 7 hour corpus of automatically aligned broadcast speech, the relative frequencies of vowels are compared across phonological contexts. Results indicate a near complementary distribution of [ʌ] and [ɨ]: the contrast scores lowest of all pairwise comparisons on measures of functional load, and shows the highest Kullback-Leibler divergence, suggesting that few lexical distinctions depend on the contrast. Thereafter, forced alignment is performed using an existing ASR system. The system selects among [ɨ], [ʌ], ø for lexical /ɨ/, testing for its reduction in continuous speech. The same data is transcribed using the ASR system where [ʌ]/[ɨ] are merged, testing the hypothesis that loss of a marginal contrast has little impact on ASR error rates. Both results are consistent with functional load calculations, indicating that the /ʌ/-/ɨ/ contrast is lexically and phonetically weak. These results show how automatic transcription tools can help test phonological predictions using continuous speech.

Notes

Effects of Subglottal-Coupling and Interdental-Space on Formant Trajectories During Front-to-Back Vowel Transitions in Chinese

Shuanglin Fan, Kiyoshi Honda, Jianwu Dang, Hui Feng; Tianjin University, China

Sun-O-7-3-4, Time: 14:30

Discontinuity of the second formant (F2 discontinuity) is often found during back-to-front vowel transitions, and it has been thought due to two possible effects: acoustic coupling with the subglottal tract (subglottal-coupling effect, SCE) and traveling anti-resonance of the interdental space (interdental-space effect, ISE). Although both are possible to appear together, either of the two is common to find in many spectrographic observations, and how to distinguish from one another is often puzzling. This study aims at exploring manifestations of the two effects in Chinese triphthongs through acoustic analysis on front-to-back vowel sequences. Test utterances were recorded from five Chinese speakers with simultaneous measurement of subglottal resonance via a vibration sensor adhered to their necks. Results revealed that F2 discontinuity occurs near the second subglottal formant (SgF2) but not always, and discontinuity of both F2 and F3 is more common that occurs with a short time lag, suggesting predominance of ISE rather than SCE in the data.

Perceptual Lateralization of Coda Rhotic Production in Puerto Rican Spanish

Mairym Lloréns Monteserín, Shrikanth S. Narayanan, Louis Goldstein; University of Southern California, USA

Sun-O-7-3-5, Time: 14:50

When speakers of Puerto Rican Spanish (PRS) produce phonemic coda taps, Spanish-speakers of other dialects often perceive these as laterals. We observed production of phonemic coda laterals and taps by a male PRS speaker in real-time MRI. Temporal and spatial characteristics of tongue tip movements during coda liquid production are inconsistent with accounts positing a categorical change from rhotic to lateral in coda for this speaker. Perceptual coding of coda tap production by naïve listeners suggests that both preceding vowel type and the relative strength of a proximal prosodic boundary may impact the proportion of the subject's phonemic taps that received a lateral percept. Results are discussed in the context of persistent difficulties in modeling the gestural representation of liquid consonants.

Interaction Between Lexical Tone and Intonation: An EMA Study

Hao Yi, Sam Tilsen; Cornell University, USA

Sun-O-7-3-6, Time: 15:10

This paper aims to examine the interaction of intonation and lexical tone within the framework of Articulatory Phonology, by investigating the timing relationship between oral articulatory gestures and tone-related/intonation-related F0 dynamics. Specifically, we compared the consonant-vowel-F0 (C-V-T) coordinative patterns at phrase-final position and at phrase-medial position. We found that the C-V-T coordination was altered by the presence of boundary tones, which is in line with the sequential model in which tone and intonation are conceptualized as events that interact at the phonological level before the phonetic implementation. However, the effect of boundary tone on the C-V-T coordination seemed to be tone-specific. Moreover, the presence of pitch accents also influenced the intra-syllabic C-V-T coordinative patterns. By presenting evidence from the coordinative patterns between articulatory gestures and F0 dynamics, the current study lent support to the sequential model of the interaction between intonation and lexical tone from a gestural perspective.

Sun-O-7-4 : Speech Synthesis Oral I: Neural Networks

Bayview B, 13:30–15:30, Sunday, 11 Sept. 2016
Chairs: Heiga Zen, Zhen-Hua Ling

Deep Bidirectional LSTM Modeling of Timbre and Prosody for Emotional Voice Conversion

*Huaiping Ming[1], Dongyan Huang[1], Lei Xie[2], Jie Wu[2], Minghui Dong[1], Haizhou Li[1]; [1]A*STAR, Singapore; [2]Northwestern Polytechnical University, China*

Sun-O-7-4-1, Time: 13:30

Emotional voice conversion aims at converting speech from one emotion state to another. This paper proposes to model timbre and prosody features using a deep bidirectional long short-term memory (DBLSTM) for emotional voice conversion. A continuous wavelet transform (CWT) representation of fundamental frequency (F0) and energy contour are used for prosody modeling. Specifically, we use CWT to decompose F0 into a five-scale representation, and decompose energy contour into a ten-scale representation, where each feature scale corresponds to a temporal scale. Both spectrum and prosody (F0 and energy contour) features are simultaneously converted by a sequence to sequence conversion method with DBLSTM model, which captures both frame-wise and long-range relationship between source and target voice. The converted speech signals are evaluated both objectively and subjectively, which confirms the effectiveness of the proposed method.

Visual Speech Synthesis Using Dynamic Visemes, Contextual Features and DNNs

Ausdang Thangthai, Ben Milner, Sarah Taylor; University of East Anglia, UK

Sun-O-7-4-2, Time: 13:50

This paper examines methods to improve visual speech synthesis from a text input using a deep neural network (DNN). Two representations of the input text are considered, namely into phoneme sequences or dynamic viseme sequences. From these sequences, contextual features are extracted that include information at varying linguistic levels, from frame level down to the utterance level. These are extracted from a broad sliding window that captures context and produces features that are input into the DNN to estimate visual features. Experiments first compare the accuracy of these visual features against an HMM baseline method which establishes that both the phoneme and dynamic viseme systems perform better with best performance obtained by a combined phoneme-dynamic viseme system. An investigation into the features then reveals the importance of the frame level information which is able to avoid discontinuities in the visual feature sequence and produces a smooth and realistic output.

NOTES

A Template-Based Approach for Speech Synthesis Intonation Generation Using LSTMs

Srikanth Ronanki, Gustav Eje Henter, Zhizheng Wu, Simon King; University of Edinburgh, UK

Sun-O-7-4-3, Time: 14:10

The absence of convincing intonation makes current parametric speech synthesis systems sound dull and lifeless, even when trained on expressive speech data. Typically, these systems use regression techniques to predict the fundamental frequency (F0) frame-by-frame. This approach leads to overly-smooth pitch contours and fails to construct an appropriate prosodic structure across the full utterance. In order to capture and reproduce larger-scale pitch patterns, this paper proposes a template-based approach for automatic F0 generation, where per-syllable pitch-contour templates (from a small, automatically learned set) are predicted by a recurrent neural network (RNN). The use of syllable templates mitigates the over-smoothing problem and is able to reproduce pitch patterns observed in the data. The use of an RNN, paired with connectionist temporal classification (CTC), enables the prediction of structure in the pitch contour spanning the entire utterance. This novel F0 prediction system is used alongside separate LSTMs for predicting phone durations and the other acoustic features, to construct a complete text-to-speech system. We report the results of objective and subjective tests on an expressive speech corpus of children's audiobooks, and include comparisons to a conventional baseline that predicts F0 directly at the frame level.

Multi-Language Multi-Speaker Acoustic Modeling for LSTM-RNN Based Statistical Parametric Speech Synthesis

Bo Li[1], Heiga Zen[2]; [1]Google, USA; [2]Google, UK

Sun-O-7-4-4, Time: 14:30

Building text-to-speech (TTS) systems requires large amounts of high quality speech recordings and annotations, which is a challenge to collect especially considering the variation in spoken languages around the world. Acoustic modeling techniques that could utilize inhomogeneous data are hence important as they allow us to pool more data for training. This paper presents a long short-term memory (LSTM) recurrent neural network (RNN) based statistical parametric speech synthesis system that uses data from multiple languages and speakers. It models language variation through cluster adaptive training and speaker variation with speaker dependent output layers. Experimental results have shown that the proposed multilingual TTS system can synthesize speech in multiple languages from a single model while maintaining naturalness. Furthermore, it can be adapted to new languages with only a small amount of data.

GlottDNN — A Full-Band Glottal Vocoder for Statistical Parametric Speech Synthesis

Manu Airaksinen[1], Bajibabu Bollepalli[1], Lauri Juvela[1], Zhizheng Wu[2], Simon King[2], Paavo Alku[1]; [1]Aalto University, Finland; [2]University of Edinburgh, UK

Sun-O-7-4-5, Time: 14:50

GlottHMM is a previously developed vocoder that has been successfully used in HMM-based synthesis by parameterizing speech into two parts (glottal flow, vocal tract) according to the functioning of the real human voice production mechanism. In this study, a new glottal vocoding method, GlottDNN, is proposed. The GlottDNN vocoder is built on the principles of its predecessor, GlottHMM, but the new vocoder introduces three main improvements: GlottDNN (1) takes advantage of a new, more accurate glottal inverse filtering method, (2) uses a new method of deep neural network (DNN) -based glottal excitation generation, and (3) proposes a new approach of band-wise processing of full-band speech.

The proposed GlottDNN vocoder was evaluated as part of a full-band state-of-the-art DNN-based text-to-speech (TTS) synthesis system, and compared against the release version of the original GlottHMM vocoder, and the well-known STRAIGHT vocoder. The results of the subjective listening test indicate that GlottDNN improves the TTS quality over the compared methods.

Singing Voice Synthesis Based on Deep Neural Networks

Masanari Nishimura, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, Keiichi Tokuda; Nagoya Institute of Technology, Japan

Sun-O-7-4-6, Time: 15:10

Singing voice synthesis techniques have been proposed based on a hidden Markov model (HMM). In these approaches, the spectrum, excitation, and duration of singing voices are simultaneously modeled with context-dependent HMMs and waveforms are generated from the HMMs themselves. However, the quality of the synthesized singing voices still has not reached that of natural singing voices. Deep neural networks (DNNs) have largely improved on conventional approaches in various research areas including speech recognition, image recognition, speech synthesis, etc. The DNN-based text-to-speech (TTS) synthesis can synthesize high quality speech. In the DNN-based TTS system, a DNN is trained to represent the mapping function from contextual features to acoustic features, which are modeled by decision tree-clustered context dependent HMMs in the HMM-based TTS system. In this paper, we propose singing voice synthesis based on a DNN and evaluate its effectiveness. The relationship between the musical score and its acoustic features is modeled in frames by a DNN. For the sparseness of pitch context in a database, a musical-note-level pitch normalization and linear-interpolation techniques are used to prepare the excitation features. Subjective experimental results show that the DNN-based system outperformed the HMM-based system in terms of naturalness.

Sun-O-7-5 : Speech Quality & Intelligibility

Seacliff BCD, 13:30–15:30, Sunday, 11 Sept. 2016

Chairs: Yannis Stylianou, Sebastian Möller

Blind Recovery of Perceptual Models in Distributed Speech and Audio Coding

Tom Bäckström, Florin Ghido, Johannes Fischer; FAU Erlangen-Nürnberg, Germany

Sun-O-7-5-1, Time: 13:30

A central part of speech and audio codecs are their perceptual models, which describe the relative perceptual importance of errors in different elements of the signal representation. In practice, the perceptual models consists of signal-dependent weighting factors which are used in quantization of each element. For optimal performance, we would like to use the same perceptual model at the

NOTES

decoder. While the perceptual model is signal-dependent, however, it is not known in advance at the decoder, whereby audio codecs generally transmit this model explicitly, at the cost of increased bit-consumption. In this work we present an alternative method which recovers the perceptual model at the decoder from the transmitted signal without any side-information. The approach will be especially useful in distributed sensor-networks and the Internet of things, where the added cost on bit-consumption from transmitting a perceptual model increases with the number of sensors.

Glimpse-Based Metrics for Predicting Speech Intelligibility in Additive Noise Conditions

Yan Tang [1], Martin Cooke [2]; [1] University of Salford, UK; [2] Ikerbasque, Spain

Sun-O-7-5-2, Time: 13:50

The glimpsing model of speech perception in noise operates by recognising those speech-dominant spectro-temporal regions, or glimpses, that survive energetic masking; hence, a speech recognition component is an integral part of the model. The current study evaluates whether a simpler family of metrics based solely on quantifying the amount of supra-threshold target speech available after energetic masking can account for subjective intelligibility. The predictive power of glimpse-based metrics is compared for natural, processed and synthetic speech in the presence of stationary and fluctuating maskers. These metrics are raw glimpse proportion, extended glimpse proportion, and two further refinements: one, FMGP, incorporates a component simulating the effect of forward masking; the other, HEGP, selects speech-dominant spectro-temporal regions with above-average energy on the noisy speech. The metrics are compared alongside a state-of-the-art non-glimpsing metric, using three large datasets of listener scores. Both FMGP and HEGP equal or improve upon the predictive power of the raw and extended metrics, with across-masker correlations ranging from 0.81–0.92; both metrics equal or exceed the state-of-the-art metric in all conditions. These outcomes suggests that easily-computed measures of unmasked, supra-threshold speech can serve as robust proxies for intelligibility across a range of speech styles and additive masking conditions.

Analyzing the Relation Between Overall Quality and the Quality of Individual Phases in a Telephone Conversation

Friedemann Köster, Sebastian Möller; T-Labs, Germany

Sun-O-7-5-3, Time: 14:10

Assessing and analyzing the quality of transmitted speech in a conversational situation is an important topic in current research. For this, a conversation has been separated into three individual conversational phases (listening, speaking, and interaction), and for each phase corresponding quality-relevant perceptual dimensions have been identified. The dimensions can be used to determine the quality of each phase, and the qualities of all phases, in turn, can be combined for overall conversational quality estimation. In this article we present the work that has been conducted to identify the weights of the individual phases for the overall quality. For this, we conducted an experiment that allows the participants to perceive each phase separately and to gather the overall quality as well as the quality ratings for each individual phase. The results enable to create a linear model to predict the overall quality on the basis of the three phases. This allows to draw first conclusions regarding the relation between the individual phases and the overall quality and provides a major landmark towards a diagnostic assessment of conversational quality.

Intelligibility Enhancement at the Receiving End of the Speech Transmission System — Effects of Far-End Noise Reduction

Emma Jokinen, Paavo Alku; Aalto University, Finland

Sun-O-7-5-4, Time: 14:30

Post-processing methods can be used in mobile communications to improve the intelligibility of speech in adverse near-end background noise conditions. Generally, it is assumed that the input of the post-processing contains quantization noise only, that is to say, no far-end noise is present. However, this assumption is not entirely realistic. Therefore, the effect of far-end noise with and without noise reduction on the performance of three post-processing methods is studied in this investigation. The performance evaluation is done using subjective intelligibility and quality tests in several far-end and near-end noise conditions. The results suggest that although the noise reduction generally improves performance in stationary far-end noise, the noise reduction does not improve intelligibility in unstationary far-end noise conditions but has a positive impact on perceptual quality for some of the post-processing methods.

Intelligibility of Disordered Speech: Global and Detailed Scores

Mario Ganzeboom, Marjoke Bakker, Catia Cucchiarini, Helmer Strik; Radboud Universiteit Nijmegen, The Netherlands

Sun-O-7-5-5, Time: 14:50

Measuring the intelligibility of disordered speech is a common practice in both clinical and research contexts. Over the years various methods have been proposed and studied, including methods relying on subjective ratings by human judges, and objective methods based on speech technology. Many of these methods measure speech intelligibility at the speaker or utterance level. While this may be satisfactory for some purposes, more detailed evaluations might be required in other cases such as diagnosis and measuring or comparing the outcomes of different types of therapy (by humans or computer programs). In the current paper we investigate intelligibility ratings at three different levels of granularity: utterance, word, and subword level. In a web experiment 50 speech fragments produced by seven dysarthric speakers were rated by 36 listeners in three ways: a score per utterance on a Visual Analogue and a Likert scale, and an orthographic transcription. The latter was used to obtain word and subword (grapheme and phoneme) level ratings using automatic alignment and conversion methods. The implemented phoneme scoring method proved feasible, reliable, and provided a more sensitive and informative measure of intelligibility. Possible implications for clinical practice and research are discussed.

Modulation Enhancement of Temporal Envelopes for Increasing Speech Intelligibility in Noise

Maria Koutsogiannaki, Yannis Stylianou; University of Crete, Greece

Sun-O-7-5-6, Time: 15:10

In this paper, speech intelligibility is enhanced by manipulating the

modulation spectrum of the signal. First, the signal is decomposed into Amplitude Modulation (AM) and Frequency Modulation (FM) components using a high resolution adaptive quasi-harmonic model of speech. Then, the AM part of midrange frequencies of speech spectrum is modified by applying a transforming function which follows the characteristics of the clear style of speaking. This results in increasing the modulation depth of the temporal envelopes of casual speech as in clear speech. The modified AM components of speech are then combined with the original FM parts to synthesize the final processed signal. Subjective listening tests evaluating the intelligibility of speech in noise showed that the suggested approach increases the intelligibility of speech by 40% on average, while it is comparable with recently suggested state-of-the-art algorithms of intelligibility boosters.

Sun-O-7-6 : Speech Translation and Metadata for Linguistic/Discourse Structure

Seacliff A, 13:30–15:30, Sunday, 11 Sept. 2016
Chairs: Tanja Schultz, Laurent Besacier

Dynamic Transcription for Low-Latency Speech Translation

Jan Niehues, Thai Son Nguyen, Eunah Cho, Thanh-Le Ha, Kevin Kilgour, Markus Müller, Matthias Sperber, Sebastian Stüker, Alex Waibel; KIT, Germany

Sun-O-7-6-1, Time: 13:30

Latency is one of the main challenges in the task of simultaneous spoken language translation. While significant improvements in recent years have led to high quality automatic translations, their usefulness in real-time settings is still severely limited due to the large delay between the input speech and the delivered translation.

In this paper, we present a novel scheme which reduces the latency of a large scale speech translation system drastically. Within this scheme, the transcribed text and its translation can be updated when more context is available, even after they are presented to the user. Thereby, this scheme allows us to display an initial transcript and its translation to the user with a very low latency. If necessary, both transcript and translation can later be updated to better, more accurate versions until eventually the final versions are displayed. Using this framework, we are able to reduce the latency of the source language transcript into half. For the translation, an average delay of 3.3s was achieved, which is more than twice as fast as our initial system.

Learning a Translation Model from Word Lattices

Oliver Adams[1], Graham Neubig[2], Trevor Cohn[1], Steven Bird[1]; [1]University of Melbourne, Australia; [2]NAIST, Japan

Sun-O-7-6-2, Time: 13:50

Translation models have been used to improve automatic speech recognition when speech input is paired with a written translation, primarily for the task of computer-aided translation. Existing approaches require large amounts of parallel text for training the translation models, but for many language pairs this data is not available. We propose a model for learning lexical translation parameters directly from the word lattices for which a transcription is sought. The model is expressed through composition of each lattice with a weighted finite-state transducer representing the translation model, where inference is performed by sampling paths through the composed finite-state transducer. We show consistent word error rate reductions in two datasets, using between just 20 minutes and 4 hours of speech input, additionally outperforming a translation model trained on the 1-best path.

Disfluency Detection Using a Bidirectional LSTM

Vicky Zayats, Mari Ostendorf, Hannaneh Hajishirzi; University of Washington, USA

Sun-O-7-6-3, Time: 14:10

We introduce a new approach for disfluency detection using a Bidirectional Long-Short Term Memory neural network (BLSTM). In addition to the word sequence, the model takes as input pattern match features that were developed to reduce sensitivity to vocabulary size in training, which lead to improved performance over the word sequence alone. The BLSTM takes advantage of explicit repair states in addition to the standard reparandum states. The final output leverages integer linear programming to incorporate constraints of disfluency structure. In experiments on the Switchboard corpus, the model achieves state-of-the-art performance for both the standard disfluency detection task and the correction detection task. Analysis shows that the model has better detection of non-repetition disfluencies, which tend to be much harder to detect.

Sentence Boundary Detection Based on Parallel Lexical and Acoustic Models

Xiaoyin Che, Sheng Luo, Haojin Yang, Christoph Meinel; Universität Potsdam, Germany

Sun-O-7-6-4, Time: 14:30

In this paper we propose a solution that detects sentence boundary from speech transcript. First we train a pure lexical model with deep neural network, which takes word vectors as the only input feature. Then a simple acoustic model is also prepared. Because the models work independently, they can be trained with different data. In next step, the posterior probabilities of both lexical and acoustic models will be involved in a heuristic 2-stage joint decision scheme to classify the sentence boundary positions. This approach ensures that the models can be updated or switched freely in actual use. Evaluation on TED Talks shows that the proposed lexical model can achieve good results: 75.5% accuracy on error-involved ASR transcripts and 82.4% on error-free manual references. The joint decision scheme can further improve the accuracy by 3∼10% when acoustic data is available.

Transferring Emphasis in Speech Translation Using Hard-Attentional Neural Network Models

Quoc Truong Do, Sakriani Sakti, Graham Neubig, Satoshi Nakamura; NAIST, Japan

Sun-O-7-6-5, Time: 14:50

While traditional speech translation systems are oblivious to paralinguistic information, there has been a recent focus on speech translation systems that transfer not only the linguistic content but also emphasis information across languages. A recent work has tried to tackle this task by developing a method for mapping emphasis between languages utilizing conditional random fields (CRFs). Although CRFs allow for consideration of rich features and

NOTES

local context, they have difficulty in handling continuous variables, and cannot capture long-distance dependencies easily. In this paper, we propose a new model for emphasis transfer in speech translation using an approach based on neural networks. The proposed model can handle long-distance dependencies by using long short-term memory (LSTM) neural networks, and is able to handle continuous emphasis values through a novel hard-attention mechanism, which uses word alignments to decide which emphasis values to map from the source to the target sentence. Our experiments on the emphasis translation task showed a significant improvement of the proposed model over the previous state-of-the-art model by 4% target-language emphasis prediction F-measure according to objective evaluation and 2% F-measure according to subjective evaluation.

Better Evaluation of ASR in Speech Translation Context Using Word Embeddings

Ngoc-Tien Le, Christophe Servan, Benjamin Lecouteux, Laurent Besacier; LIG (UMR 5217), France

Sun-O-7-6-6, Time: 15:10

This paper investigates the evaluation of ASR in spoken language translation context. More precisely, we propose a simple extension of WER metric in order to penalize differently substitution errors according to their context using word embeddings. For instance, the proposed metric should catch near matches (mainly morphological variants) and penalize less this kind of error which has a more limited impact on translation performance. Our experiments show that the correlation of the new proposed metric with SLT performance is better than the one of WER. Oracle experiments are also conducted and show the ability of our metric to find better hypotheses (to be translated) in the ASR N-best. Finally, a preliminary experiment where ASR tuning is based on our new metric shows encouraging results. For reproducible experiments, the code allowing to call our modified WER and the corpora used are made available to the research community.

Sun-P-7-1 : Speech Coding and Audio Processing for Noise Reduction

Pacific Concourse – Poster A, 13:30–15:30, Sunday, 11 Sept. 2016
Chairs: Tom Bäckström, Dayana Ribas

Entropy Coding of Spectral Envelopes for Speech and Audio Coding Using Distribution Quantization

Srikanth Korse[1], Tobias Jähnel[2], Tom Bäckström[1]; [1]Fraunhofer IIS, Germany; [2]FAU Erlangen-Nürnberg, Germany

Sun-P-7-1-1, Time: 13:30

Speech and audio codecs model the overall shape of the signal spectrum using envelope models. In speech coding the predominant approach is linear predictive coding, which offers high coding efficiency at the cost of computational complexity and a rigid systems design. Audio codecs are usually based on scale factor bands, whose calculation and coding is simple, but whose coding efficiency is lower than that of linear prediction. In the current work we propose an entropy coding approach for scale factor bands, with the objective of reaching the same coding efficiency as linear prediction, but simultaneously retaining a low computational complexity. The proposed method is based on quantizing the distribution of spectral mass using beta-distributions. Our experiments show that the perceptual quality achieved with the proposed method is similar to that of linear predictive models with the same bit rate, while the design simultaneously allows variable bit-rate coding and can easily be scaled to different sampling rates. The algorithmic complexity of the proposed method is less than one third of traditional multi-stage vector quantization of linear predictive envelopes.

An Objective Evaluation Methodology for Blind Bandwidth Extension

Stéphane Villette, Sen Li, Pravin Ramadas, Daniel J. Sinder; Qualcomm Technologies, USA

Sun-P-7-1-2, Time: 13:30

In this paper we introduce an objective evaluation methodology for Blind Bandwidth Extension (BBE) algorithms. The methodology combines an objective method, POLQA, with a bandwidth requirement, based on a frequency mask. We compare its results to subjective test data, and show that it gives consistent results across several bandwidth extension algorithms. Additionally, we show that our latest BBE algorithm achieves quality similar to AMR-WB at 8.85 kbps, using both subjective and objective evaluation methods.

EVS Channel Aware Mode Robustness to Frame Erasures

Anssi Rämö, Antti Kurittu, Henri Toukomaa; Nokia, Finland

Sun-P-7-1-3, Time: 13:30

This paper discusses the voice and audio quality characteristics of the EVS, the recently standardized 3GPP Enhanced Voice Services codec. Especially frame erasure conditions with and without channel aware mode were evaluated. The test consisted of two extended range MOS listening tests. The tests contained both clean and noisy speech in clean channel as well as with four frame erasure rates (5%, 10%, 20% and 30%) for selected codecs and bitrates. In addition to subjective test results some additional objective results are presented. The results show that EVS channel aware mode performs better than EVS native mode in high FER rates. For comparison also AMR, AMR-WB and Opus codecs were included to the listening tests.

An Interaural Magnification Algorithm for Enhancement of Naturally-Occurring Level Differences

Shadi Pirhosseinloo, Kostas Kokkinakis; University of Kansas, USA

Sun-P-7-1-4, Time: 13:30

In this work, we describe an interaural magnification algorithm for speech enhancement in noise and reverberation. The proposed algorithm operates by magnifying the interaural level differences corresponding to the interfering sound source. The enhanced signal outputs are estimated by processing the signal inputs with the interaurally-magnified head-related transfer functions. Experimental results with speech masked by a single interfering source in anechoic and reverberant scenarios indicate that the proposed algorithm yields an increased benefit due to spatial release from masking and a much higher perceived speech quality.

NOTES

Probabilistic Spatial Filter Estimation for Signal Enhancement in Multi-Channel Automatic Speech Recognition

Hendrik Kayser[1], *Niko Moritz*[2], *Jörn Anemüller*[1]; [1]*Carl von Ossietzky Universität Oldenburg, Germany;* [2]*Fraunhofer IDMT, Germany*

Sun-P-7-1-5, Time: 13:30

Speech recognition in multi-channel environments requires target speaker localization, multi-channel signal enhancement and robust speech recognition. We here propose a system that addresses these problems: Localization is performed with a recently introduced probabilistic localization method that is based on support-vector machine learning of GCC-PHAT weights and that estimates a spatial source probability map. The main contribution of the present work is the introduction of a probabilistic approach to (re-)estimation of location-specific steering vectors based on weighting of observed inter-channel phase differences with the spatial source probability map derived in the localization step. Subsequent speech recognition is carried out with a DNN-HMM system using amplitude modulation filter bank (AMFB) acoustic features which are robust to spectral distortions introduced during spatial filtering.

The system has been evaluated on the CHIME-3 multi-channel ASR dataset. Recognition was carried out with and without probabilistic steering vector re-estimation and with MVDR and delay-and-sum beamforming, respectively. Results indicate that the system attains on real-world evaluation data a relative improvement of 31.98% over the baseline and of 21.44% over a modified baseline. We note that this improvement is achieved without exploiting oracle knowledge about speech/non-speech intervals for noise covariance estimation (which is, however, assumed for baseline processing).

Improved *a priori* SAP Estimator in Complex Noisy Environment for Dual Channel Microphone System

Youna Ji, Young-cheol Park; Yonsei University, Korea

Sun-P-7-1-6, Time: 13:30

In this paper, *a priori* speech absence probability (SAP) estimator is proposed for accurately obtaining the speech presence probability (SPP) in a complex noise field. Unlike previous techniques, the proposed estimator considers a complex noise sound field where the target speech is corrupted by a coherent interference with diffuse noise around. The proposed algorithm estimates *a priori* SAP based on the normalized speech to interference plus diffuse noise ratio (SINR) being expressed in terms of the speech to interference ratio (SIR) and the directional to diffuse noise ratio (DDR). The SIR is obtained from a quadratic equation of the magnitude-squared coherence (MSC) between two microphone signals. A performance comparison with several advanced *a priori* SAP estimators was conducted in terms of the receiver operating characteristic (ROC) curve. The proposed algorithm attains a correct detection rate at a given false-alarm rate that is higher than those attained by conventional algorithms.

A Spectral Modulation Sensitivity Weighted Pre-Emphasis Filter for Active Noise Control System

Kah-Meng Cheong[1], *Yuh-Yuan Wang*[2], *Tai-Shih Chi*[1]; [1]*National Chiao Tung University, Taiwan;* [2]*TANGENT Microelectromechanics, Taiwan*

Sun-P-7-1-7, Time: 13:30

Psychoacoustic active noise control (ANC) systems by considering human hearing thresholds in different frequency bands were developed in the past. Besides the frequency sensitivity, human hearing also shows different sensitivity to spectral and temporal modulations of the sound. In this paper, we propose a new psychoacoustic active noise control system by further considering the spectral modulation sensitivity of human hearing. In addition to the sound pressure level (SPL), the loudness level is also objectively assessed to evaluate the noise reduction performance of the proposed ANC system. Simulation results demonstrate the proposed system outperforms two compared systems under test conditions of narrowband and broadband noise in terms of the loudness level. The proposed algorithm has been validated on TI C6713 DSP platform for real-time process.

Semi-Coupled Dictionary Based Automatic Bandwidth Extension Approach for Enhancing Children's ASR

Ganji Sreeram, Rohit Sinha; IIT Guwahati, India

Sun-P-7-1-8, Time: 13:30

The work presented in this paper is motivated by our earlier work exploring sparse representation based approach for automatic bandwidth extension (ABWE) of speech signals. In that work, two dictionaries one for voiced and the other for unvoiced speech frames are created using KSVD algorithm on wideband data. Each of the atoms of these dictionaries is then decimated and interpolated by a factor of 2 to generate narrowband interpolated (NBI) dictionaries whose atoms have one-to-one correspondence with those of the WB dictionaries. The given narrowband speech frames are also interpolated to generated NBI targets and those are sparse coded over the NBI dictionaries. The resulting sparse codes are then applied to the WB dictionaries to estimate the WB target data. In this work, we extend the said approach by making use of an existing semi-coupled dictionary learning (SCDL) algorithm. Unlike the direct dictionary learning, the SCDL algorithm also learns a set of bidirectional transforms coupling the dictionaries more flexibly. The bandwidth enhanced speech obtained employing the SCDL approach and a modified high/low band gain adjustment yields significant improvements in terms of speech quality measures as well as in the context of children's mismatched speech recognition.

NOTES

Sun-P-7-2 : Special Session: Speech, Audio, and Language Processing Techniques Applied to Bird and Animal Vocalizations

Pacific Concourse – Poster B, 13:30–15:30, Sunday, 11 Sept. 2016
Chairs: Naomi Harte, Peter Jančovič, Karl-L. Schuchmann

Bird Song Synthesis Based on Hidden Markov Models

Jordi Bonada[1], Robert Lachlan[2], Merlijn Blaauw[1]; [1]Universitat Pompeu Fabra, Spain; [2]Queen Mary University of London, UK

Sun-P-7-2-1, Time: 13:30

This paper focuses on the synthesis of bird songs using Hidden Markov Models (HMM). This technique has been widely used for speech modeling and synthesis. However, features and contextual factors typically used for human speech are not appropriate for modeling bird songs. Moreover, while for speech we can easily control the content of the recordings, this is not the case for bird songs, where we have to rely on the spontaneous singing of the animal. In this work we briefly overview the characteristics of bird songs, compare them to speech, and propose strategies for adapting the widely-used HTS (HMM-based Speech Synthesis System) framework to model and synthesize bird songs. In particular, we focus on Chaffinch species and a database of recordings of several song bouts of one male bird. At the end we discuss the synthesis results obtained.

Noise-Robust Hidden Markov Models for Limited Training Data for Within-Species Bird Phrase Classification

Kantapon Kaewtip, Charles Taylor, Abeer Alwan; University of California at Los Angeles, USA

Sun-P-7-2-2, Time: 13:30

Hidden Markov Models (HMMs) have been studied and used extensively in speech and birdsong recognition, but they are not robust to limited training data and noise. This paper presents two novel approaches to training continuous and discrete HMMs with extremely limited data. First, the algorithm learns the global Gaussian Mixture Models (GMMs) for all training phrases available. GMM parameters are then used to initialize state parameters of each individual model. For the GMM-HMM framework, the number of states and the mixture components for each state are determined by the acoustic variation of each phrase type. The (high-energy) time-frequency prominent regions are used to compute the state emitting probability to increase noise-robustness. For the discrete HMM framework, the probability distribution of each state is initialized by the global GMMs in training. In testing, the probability of each codebook is estimated using the prominent regions of each state to increase noise-robustness. In Cassins Vireo phrase classification using 75 phrase types, the new GMM-HMM approach achieves 79.5% and 87% classification accuracy using 1 and 2 phrases, respectively, while HTK's GMM-HMM framework makes guess predictions resulting in 1.33% accuracy. The performance of the other algorithm is presented in the paper.

A Framework for Automated Marmoset Vocalization Detection and Classification

Alan Wisler[1], Laura J. Brattain[2], Rogier Landman[3], Thomas F. Quatieri[2]; [1]Arizona State University, USA; [2]MIT Lincoln Laboratory, USA; [3]Broad Institute of MIT and Harvard, USA

Sun-P-7-2-3, Time: 13:30

This paper describes a novel framework for automated marmoset vocalization detection and classification from within long audio streams recorded in a noisy animal room, where multiple marmosets are housed. To overcome the challenge of limited manually annotated data, we implemented a data augmentation method using only a small number of labeled vocalizations. The feature sets chosen have the desirable property of capturing characteristics of the signals that are useful in both identifying and distinguishing marmoset vocalizations. Unlike many previous methods, feature extraction, call detection, and call classification in our system are completely automated. The system maintains a good performance of 20% equal error detection rate using data with high number of noise events and 15% of classification error. Performance can be further improved with additional labeled training data. Because this extensible system is capable of identifying both positive and negative welfare indicators, it provides a powerful framework for non-human primate welfare monitoring as well as behavior assessment.

Call Alternation Between Specific Pairs of Male Frogs Revealed by a Sound-Imaging Method in Their Natural Habitat

Ikkyu Aihara[1], Takeshi Mizumoto[2], Hiromitsu Awano[3], Hiroshi G. Okuno[4]; [1]University of Tsukuba, Japan; [2]Honda Research Institute Japan, Japan; [3]Kyoto University, Japan; [4]Waseda University, Japan

Sun-P-7-2-4, Time: 13:30

Male frogs vocalize calls to attract conspecific females as well as to announce their own territories to other male frogs. In the choruses, acoustic interaction allows the male frogs to alternate their calls with each other. Such call alternation is reported in various species of frogs including Japanese tree frogs (*Hyla japonica*). During call alternation, both male and female frogs are likely to discriminate calls of the male frogs because of small amount of call overlaps. Here, we show that call alternation is observed in natural choruses of male Japanese tree frogs especially between neighboring pairs. First, we demonstrate that caller positions and call timings can be estimated by a sound-imaging method. Second, the occurrence of call alternation is detected on the basis of statistical tests on phase differences of calls between respective pairs. Although our previous study revealed a global synchronization pattern in natural choruses of the male frogs, local chorus structures were not examined well. Through the observation of call alternation between specific pairs, this study suggests the existence of selective attention in the frog choruses.

Sinusoidal Modelling for Ecoacoustics

Patrice Guyot[1], Alice Eldridge[1], Ying Chen Eyre-Walker[1], Alison Johnston[2], Thomas Pellegrini[3], Mika Peck[1]; [1]University of Sussex, UK; [2]British Trust for Ornithology, UK; [3]IRIT, France

Sun-P-7-2-5, Time: 13:30

Biodiversity assessment is a central and urgent task, necessary to monitoring the changes to ecological systems and understanding the factors which drive these changes. Technological advances are providing new approaches to monitoring, which are particularly useful in remote regions. Situated within the framework of the emerging field of ecoacoustics, there is growing interest in the possibility of extracting ecological information from digital recordings of the acoustic environment. Rather than focusing on identification of individual species, an increasing number of automated indices attempt to summarise acoustic activity at the community level, in order to provide a proxy for biodiversity. Originally designed for speech processing, sinusoidal modelling has previously been used as a bioacoustic tool, for example to detect particular bird species. In this paper, we demonstrate the use of sinusoidal modelling as a proxy for bird abundance. Using data from acoustic surveys made during the breeding season in UK woodland, the number of extracted sinusoidal tracks is shown to correlate with estimates of bird abundance made by expert ornithologists listening to the recordings. We also report ongoing work exploring a new approach to investigate the composition of calls in spectro-temporal space that constitutes a promising new method for Ecoaoustic biodiversity assessment.

Individual Identity in Songbirds: Signal Representations and Metric Learning for Locating the Information in Complex Corvid Calls

Dan Stowell[1], Veronica Morfi[1], Lisa F. Gill[2]; [1]Queen Mary University of London, UK; [2]MPI for Ornithology, Germany

Sun-P-7-2-6, Time: 13:30

Bird calls range from simple tones to rich dynamic multi-harmonic structures. The more complex calls are very poorly understood at present, such as those of the scientifically important corvid family (jackdaws, crows, ravens, etc.). Individual birds can recognise familiar individuals from calls, but where in the signal is this identity encoded? We studied the question by applying a combination of feature representations to a dataset of jackdaw calls, including linear predictive coding (LPC) and adaptive discrete Fourier transform (aDFT). We demonstrate through a classification paradigm that we can strongly outperform a standard spectrogram representation for identifying individuals, and we apply metric learning to determine which time-frequency regions contribute most strongly to robust individual identification. Computational methods can help to direct our search for understanding of these complex biological signals.

Recognition of Multiple Bird Species Based on Penalised Maximum Likelihood and HMM-Based Modelling of Individual Vocalisation Elements

Peter Jančovič, Münevver Köküer; University of Birmingham, UK

Sun-P-7-2-7, Time: 13:30

This paper presents an extension of our recent work on recognition of multiple bird species from their vocalisations by incorporating an improved acoustic modelling. The acoustic scene is segmented into spectro-temporal isolated segments by employing a sinusoidal detection algorithm, which is able to handle multiple simultaneous bird vocalisations. Each segment is represented as a temporal sequence of frequencies of the detected sinusoid. Each bird species is represented by a set of hidden Markov models (HMMs), each HMM modelling a particular vocalisation element. A set of elements is discovered in an unsupervised manner using a partial dynamic time warping algorithm and agglomerative hierarchical clustering. Recognition of multiple bird species is performed based on maximising the likelihood of the set of detected segments on a subset of bird species models, with a penalisation applied for increasing the number of bird species. Experimental evaluations used audio field recordings containing 30 bird species. Detected segments from several bird species are joined to simulate the presence of multiple bird species. It is demonstrated that the use of improved acoustic modelling in conjunction with the maximum likelihood score combination method provides considerable improvements over previous results and the use of majority voting.

Cost Effective Acoustic Monitoring of Bird Species

Ciira wa Maina; DeKUT, Kenya

Sun-P-7-2-8, Time: 13:30

Climate change and human encroachment are some of the major threats facing several natural ecosystems around the world. To ensure the protection of ecosystems under threat, it is important to monitor the biodiversity within these ecosystems to determine when conservation efforts are necessary. For this to be achieved, technologies that allow large areas to be monitored in a cost effective manner are essential. In this work we investigate the use of acoustic recordings obtained using a low cost Raspberry Pi based recorder to monitor the Hartlaub's Turaco in central Kenya. This species is endemic to East Africa and faces habitat loss due to climate change. Using simple features derived from the spectrograms of the recordings, a Gaussian mixture model classifier is able to accurately screen large data sets for presence of the Hartlaub's Turaco call. In addition, we present a method based on musical note onset detection to determine the number of calls within a recording.

Feature Learning and Automatic Segmentation for Dolphin Communication Analysis

Daniel Kohlsdorf[1], Denise Herzing[2], Thad Starner[1]; [1]Georgia Institute of Technology, USA; [2]Wild Dolphin Project, USA

Sun-P-7-2-9, Time: 13:30

The study of dolphin cognition involves intensive research of animal vocalizations recorded in the field. We address the automated analysis of audible dolphin communication and propose a system that automatically discovers patterns in dolphin signals. These patterns are invariant to frequency shifts and time warping transformations. The discovery algorithm is based on feature learning and unsupervised time series segmentation using hidden Markov models. Researchers can inspect the patterns visually and interactively run comparative statistics between the distribution of dolphin signals in different behavioral contexts. Our results indicate that our system provides meaningful patterns to the marine biologist and that the comparative statistics are aligned with the biologists domain knowledge.

NOTES

Localizing Bird Songs Using an Open Source Robot Audition System with a Microphone Array

Reiji Suzuki[1], Shiho Matsubayashi[1], Kazuhiro Nakadai[2], Hiroshi G. Okuno[3]; [1]Nagoya University, Japan; [2]Honda Research Institute Japan, Japan; [3]Waseda University, Japan

Sun-P-7-2-10, Time: 13:30

Auditory scene analysis is critical in observing bio-diversity and understanding social behavior of animals in natural habitats because many animals and birds sing or call and environmental sounds are made. To understand acoustic interactions among songbirds, we need to collect spatiotemporal data for a long period of time during which multiple individuals and species are singing simultaneously. We are developing HARKBird, which is an easily-available and portable system to record, localize, and analyze bird songs. It is composed of a laptop PC with an open source robot audition system HARK (Honda Research Institute Japan Audition for Robots with Kyoto University) and a commercially available low-cost microphone array. HARKBird helps us annotate bird songs and grasp the soundscape around the microphone array by providing the direction of arrival (DOA) of each localized source and its separated sound automatically. In this paper, we briefly introduce our system and show an example analysis of a track recorded at the experimental forest of Nagoya University, in central Japan. We demonstrate that HARKBird can extract birdsongs successfully by combining multiple localization results with appropriate parameter settings that took account of ecological properties of environment around a microphone array and species-specific properties of bird songs.

Robust Detection of Multiple Bioacoustic Events with Repetitive Structures

Frank Kurth; Fraunhofer FKIE, Germany

Sun-P-7-2-11, Time: 13:30

In this paper we address the task of robustly detecting multiple bioacoustic events with repetitive structures in outdoor monitoring recordings. For this, we propose to use the shift-autocorrelation (shift-ACF) that was previously successfully applied to F0 estimation in speech processing and has subsequently led to a robust technique for speech activity detection. As a first contribution, we illustrate the potentials of various shift-ACF-based time-frequency representations adapted to repeated signal components in the context of bioacoustic pattern detection. Secondly, we investigate a method for automatically detecting multiple repeated events and present an application to a concrete bioacoustic monitoring scenario. As a third contribution, we provide a systematic evaluation of the shift-ACF-based feature extraction in representing multiple overlapping repeated events.

A Real-Time Parametric General-Purpose Mammalian Vocal Synthesiser

Roger K. Moore; University of Sheffield, UK

Sun-P-7-2-12, Time: 13:30

Although R&D into ‘speech synthesis’ has received a considerable amount of attention over many years, there has been remarkably little effort devoted to constructing vocal synthesisers for non-human animals. Of course, interest in synthesising human speech has been driven by the demand for practical applications such as reading machines for the blind or voice-operated assistants. Nevertheless, there are potential uses for non-human vocal synthesis: *e.g.* in education, robotics or ecological fieldwork. The latter is of particular interest, since it is common practice to use ‘playback’ methods (based on recorded samples) that do not easily facilitate parametric control over key experimental variables. Therefore, this paper presents the design and implementation of a real-time parametric general-purpose mammalian vocal synthesiser. The approach taken has been to decompose the overall sound production system into the relevant anatomical components (such as the lungs, vocal folds, tongue and mouth), and to implement a real-time simulation in ‘Pure Data’ — an open-source dataflow programming language. The software was successfully used to design an appropriate mammalian voice for the *MiRo* biomimetic robot, but there are potential applications in a number of areas. The software is available for free download at `http://www.dcs.shef.ac.uk/~roger/downloads.html`.

YIN-Bird: Improved Pitch Tracking for Bird Vocalisations

Colm O’Reilly, Nicola M. Marples, David J. Kelly, Naomi Harte; Trinity College Dublin, Ireland

Sun-P-7-2-13, Time: 13:30

Pitch is an important property of birdsong. Accurate and automatic tracking of pitch for large numbers of recordings would be useful for automatic analysis of birdsong. Currently, pitch trackers such as YIN can work with carefully tuned parameters but the characteristics of birdsong mean those optimal parameters can change quickly even within a single song. This paper presents YIN-bird, a modified version of YIN which exploits spectrogram properties to automatically set a minimum fundamental frequency parameter for YIN. This parameter is continuously updated without user intervention. A ground truth dataset of synthetic birdsong with known fundamental frequency is generated for evaluation of YIN-bird. Listener tests from expert birders described the synthetic samples as “sounding like original & can hardly tell it is synthetic”. Gross pitch error on whistles and trills were reduced by up to 4%. An analysis of nasal sounds shows the challenge in accurate pitch tracking for this syllable type.

Sun-P-7-3 : Learning, Education and Different Speech

Pacific Concourse – Poster C, 13:30–15:30, Sunday, 11 Sept. 2016
Chair: Thomas Pellegrini

Mispronunciation Detection Leveraging Maximum Performance Criterion Training of Acoustic Models and Decision Functions

Yao-Chi Hsu, Ming-Han Yang, Hsiao-Tsung Hung, Berlin Chen; National Taiwan Normal University, Taiwan

Sun-P-7-3-1, Time: 13:30

Mispronunciation detection is part and parcel of a computer assisted pronunciation training (CAPT) system, facilitating second-language (L2) learners to pinpoint erroneous pronunciations in a given utterance so as to improve their spoken proficiency. This paper presents a continuation of such a general line of research and the major contributions are twofold. First, we present an effective training approach that estimates the deep neural network based acoustic models involved in the mispronunciation detection process by optimizing an objective directly linked to the ultimate evaluation metric. Second, along the same vein, two disparate logistic sigmoid based decision functions with either phone- or senone-dependent parameterization are also inferred and used for enhanced mispronunciation detection. A series of experiments on a Mandarin mispronunciation detection task seem to show the performance merits of the proposed method.

Using Clinician Annotations to Improve Automatic Speech Recognition of Stuttered Speech

Peter A. Heeman[1], Rebecca Lunsford[1], Andy McMillin[2], J. Scott Yaruss[3]; [1]BioSpeech, USA; [2]Oregon Health & Science University, USA; [3]University of Pittsburgh, USA

Sun-P-7-3-2, Time: 13:30

In treating people who stutter, clinicians often have their clients read a story in order to determine their stuttering frequency. As the client is speaking, the clinician annotates each disfluency. For further analysis of the client's speech, it is useful to have a word transcription of what was said. However, as these are real-time annotations, they are not always correct, and they usually lag where the actual disfluency occurred. We have built a tool that rescores a word lattice taking into account the clinician's annotations. In the paper, we describe how we incorporate the clinician's annotations, and the improvement over a baseline version. This approach of leveraging clinician annotations can be used for other clinical tasks where a word transcription is useful for further or richer analysis.

Deep Neural Networks for Voice Quality Assessment Based on the GRBAS Scale

Simin Xie[1], Nan Yan[2], Ping Yu[3], Manwa L. Ng[4], Lan Wang[2], Zhuanzhuan Ji[2]; [1]Wuhan University of Technology, China; [2]Chinese Academy of Sciences, China; [3]PLA General Hospital, China; [4]University of Hong Kong, China

Sun-P-7-3-3, Time: 13:30

In the field of voice therapy, perceptual evaluation is widely used by expert listeners as a way to evaluate pathological and normal voice quality. This approach is understandably subjective as it is subject to listeners' bias which high inter- and intra-listeners variability can be found. As such, research on automatic assessment of pathological voices using a combination of subjective and objective analyses emerged. The present study aimed to develop a complementary automatic assessment system for voice quality based on the well-known GRBAS scale by using a battery of multidimensional acoustical measures through Deep Neural Networks. A total of 44 dimensionality parameters including Mel-frequency Cepstral Coefficients, Smoothed Cepstral Peak Prominence and Long-Term Average Spectrum was adopted. In addition, the state-of-the-art automatic assessment system based on Modulation Spectrum (MS) features and GMM classifiers was used as comparison system. The classification results using the proposed method revealed a moderate correlation with subjective GRBAS scores of dysphonic severity, and yielded a better performance than MS-GMM system, with the best accuracy around 81.53%. The findings indicate that such assessment system can be used as an appropriate evaluation tool in determining the presence and severity of voice disorders.

Automated Screening of Speech Development Issues in Children by Identifying Phonological Error Patterns

Lauren Ward[1], Alessandro Stefani[1], Daniel Smith[1], Andreas Duenser[1], Jill Freyne[1], Barbara Dodd[2], Angela Morgan[3]; [1]CSIRO, Australia; [2]University of Melbourne, Australia; [3]MCRI, Australia

Sun-P-7-3-4, Time: 13:30

A proof of concept system is developed to provide a broad assessment of speech development issues in children. It has been designed to enable non-experts to complete an initial screening of children's speech with the aim of reducing the workload on Speech Language Pathology services. The system was composed of an acoustic model trained by neural networks with split temporal context features and a constrained HMM encoded with the knowledge of Speech Language Pathologists. Results demonstrated the system was able to improve PER by 33% compared with standard HMM decoders, with a minimum PER of 19.03% achieved. Identification of Phonological Error Patterns with up to 94% accuracy was achieved despite utilizing only a small corpus of disordered speech from Australian children. These results indicate the proposed system is viable and the direction of further development are outlined in the paper.

Automatic Pronunciation Evaluation of Non-Native Mandarin Tone by Using Multi-Level Confidence Measures

Ju Lin, Yanlu Xie, Jinsong Zhang; BLCU, China

Sun-P-7-3-5, Time: 13:30

Automatic evaluation of tonal production plays an important role in a tonal language Computer-Assisted Pronunciation Training (CAPT) system. In this paper, we propose an automatic evaluation method for non-native Mandarin tones. The method applied multi-level confidence measures generated from Deep Neural Network (DNN). The confidence measures consisted of Log Posterior Ratios (LPR), Average Frame-level Log Posteriors (AFLP) and Segment-level Log Posteriors (SLP). The LPR was calculated between the correct tone model and competing tone models. The AFLP and LPR were obtained from frame-level scores. And the SLP was directly derived from segment-level scores. The multi-level confidence measures were modeled with a support vector machine (SVM) classifier. For comparison, three experiments were conducted according to different features: AFLP+LPR, SLP only and AFLP+LPR+SLP. The experimental results showed that the performance of the system which used multi-level confidence measures was the best, achieving a FRR of 5.63% and a DA of 82.45%, which demonstrated the efficiency of the proposed method.

Dysarthric Speech Recognition Using Kullback-Leibler Divergence-Based Hidden Markov Model

Myungjong Kim[1], Jun Wang[1], Hoirin Kim[2]; [1]University of Texas at Dallas, USA; [2]KAIST, Korea

Sun-P-7-3-6, Time: 13:30

Dysarthria is a neuro-motor speech disorder that impedes the physical production of speech. Patients with dysarthria often have trouble in pronouncing certain sounds, resulting in undesirable phonetic variation. Current automatic speech recognition systems designed for the general public are ineffective for dysarthric sufferers due to the phonetic variation. In this paper, we investigate dysarthric

NOTES

speech recognition using Kullback-Leibler divergence-based hidden Markov models. In the model, the emission probability of state is modeled by a categorical distribution using phoneme posterior probabilities from a deep neural network, and therefore, it can effectively capture the phonetic variation of dysarthric speech. Experimental evaluation on a database of several hundred words uttered by 30 speakers consisting of 12 mildly dysarthric, 8 moderately dysarthric, and 10 control speakers showed that our approach provides substantial improvement over the conventional Gaussian mixture model and deep neural network based speech recognition systems.

Detection of Total Syllables and Canonical Syllables in Infant Vocalizations

Anne S. Warlaumont [1], Heather L. Ramsdell-Hudock [2]; [1] University of California at Merced, USA; [2] Idaho State University, USA

Sun-P-7-3-7, Time: 13:30

During the first two years of life, human infants produce increasing numbers of speech-like (canonical) syllables. Both basic research on child speech development and clinical work assessing a child's pre-speech capabilities stand to benefit from efficient, accurate, and consistent methods for counting the syllables present in a given infant utterance. To date, there have been only a few attempts to perform syllable counting in infant vocalizations automatically, and thorough comparisons to human listener counts are lacking. We apply four existing, openly available systems for detecting syllabic, consonant, or vowel elements in vocalizations and apply them to a set of infant utterances individually and in combination. With the automated methods, we obtain canonical syllable counts that correlate well enough with trained human listener counts to replicate the pattern of increasing canonical syllable frequency as infants get older. However, agreement between the automated methods and human listener canonical syllable counts is considerably weaker than human listeners' agreement with each other. On the other hand, automatic identification of syllable-like units of any type (canonical and non-canonical both included) match human listeners' judgments quite well. Interestingly, these total syllable counts also increase with infant age.

Improving Automatic Recognition of Aphasic Speech with AphasiaBank

Duc Le, Emily Mower Provost; University of Michigan, USA

Sun-P-7-3-8, Time: 13:30

Automatic recognition of aphasic speech is challenging due to various speech-language impairments associated with aphasia as well as a scarcity of training data appropriate for this speaker population. AphasiaBank, a shared database of multimedia interactions primarily used by clinicians to study aphasia, offers a promising source of data for Deep Neural Network acoustic modeling. In this paper, we establish the first large-vocabulary continuous speech recognition baseline on AphasiaBank and study recognition accuracy as a function of diagnoses. We investigate several out-of-domain adaptation methods and show that AphasiaBank data can be leveraged to significantly improve the recognition rate on a smaller aphasic speech corpus. This work helps broaden the understanding of aphasic speech recognition, demonstrates the potential of AphasiaBank, and guides researchers who wish to use this database for their own work.

Pronunciation Assessment of Japanese Learners of French with GOP Scores and Phonetic Information

Vincent Laborde [1], Thomas Pellegrini [1], Lionel Fontan [1], Julie Mauclair [1], Halima Sahraoui [2], Jérôme Farinas [1]; [1] IRIT, France; [2] Octogone-Lordat (EA4156), France

Sun-P-7-3-9, Time: 13:30

In this paper, we report automatic pronunciation assessment experiments at phone-level on a read speech corpus in French, collected from 23 Japanese speakers learning French as a foreign language. We compare the standard approach based on Goodness Of Pronunciation (GOP) scores and phone-specific score thresholds to the use of logistic regressions (LR) models. French native speech corpus, in which artificial pronunciation errors were introduced, was used as training set. Two typical errors of Japanese speakers were considered: /ʀ/ and /v/ often mispronounced as [l] and [b], respectively. The LR classifier achieved a 64.4% accuracy similar to the 63.8% accuracy of the baseline threshold method, when using GOP scores and the expected phone identity as input features only. A significant performance gain of 20.8% relative was obtained by adding phonetic and phonological features as input to the LR model, leading to a 77.1% accuracy. This LR model also outperformed another baseline approach based on linear discriminant models trained on raw f-BANK coefficient features.

Pronunciation Error Detection for New Language Learners

Sean Robertson, Cosmin Munteanu, Gerald Penn; University of Toronto, Canada

Sun-P-7-3-10, Time: 13:30

Existing pronunciation error detection research assumes that second language learners' speech is advanced enough that its segments are generally well articulated. However, learners just beginning their studies, especially when those studies are organized according to western, dialogue-driven pedagogies, are unlikely to abide by those assumptions. This paper presents an evaluation of pronunciation error detectors on the utterances of second language learners just beginning their studies. A corpus of nonnative speech data is collected through an experimental application teaching beginner French. Word-level binary labels are acquired through successive pairwise comparisons made by language experts with years of experience teaching. Six error detectors are trained to classify these data: a classifier inspired by phonetic distance algorithms; the Goodness of Pronunciation classifier [1]; and four GMM-based discriminative classifiers modelled after [2]. Three partitioning strategies for 4-fold cross-validation are tested: one based on corpus distribution, another leaving speakers out, and another leaving annotators out. The best error detector, a log-likelihood ratio of native versus nonnative GMMs, achieved detector-annotator agreement of up to $\kappa = .41$, near the expected between-annotator agreement.

L2 English Rhythm in Read Speech by Chinese Students

Hongwei Ding [1], Xinping Xu [2]; [1] Shanghai Jiao Tong University, China; [2] Shanghai East High School, China

Sun-P-7-3-11, Time: 13:30

L2 English speech produced by Mandarin Chinese speakers is usually perceived to be intermediate between stress-timed and syllable-timed in rhythm. However, previous studies seldom employed comparable

NOTES

data of target language, source language and L2 interlanguage in one investigation, which may lead to discrepant results. Thus, in this study we conducted a contrastive investigation of 10 Chinese students and 10 native English speakers. We measured the rhythmic correlates in passage readings of Mandarin and L2 English produced by the native Chinese subjects, and those of English by the native British speakers. Comparison of the widely used rhythmic metrics *%V*, ΔC, ΔV, *nPVI*, *rPVI*, *VarcoV*, and *VarcoC* confirmed that Mandarin Chinese is a highly syllable-timed language. Results suggested that vowel-related metrics were better indexes to classify L2 English rhythm produced by Chinese speakers as being more syllable-timed than stress-timed. Analysis showed that vowel epenthesis, non-reduction of vowels, and no stressed/unstressed contrast could contribute to the auditory impression of syllable-timed rhythm of their L2 English. This investigation could shed some light on the Chinese accent of L2 English and provided support to facilitate the rhythmic acquisition of stress-timed languages for Chinese students.

Sun-P-7-4 : Dialogue Systems and Analysis of Dialogue

Pacific Concourse – Poster D, 13:30–15:30, Sunday, 11 Sept. 2016
Chair: Joakim Gustafson

Improving the Probabilistic Framework for Representing Dialogue Systems with User Response Model

Miao Li, Zhipeng Chen, Ji Wu; Tsinghua University, China

Sun-P-7-4-1, Time: 13:30

A probabilistic framework for goal-driven spoken dialogue systems (SDSs) has been proposed by us in a previous work. In the framework, a target distribution, instead of the frame structure, is used to represent the dialogue state at each turn. The target-based state tracking algorithm enables the system to handle uncertainties in the dialogue. By summarizing the target-based state, information from the back-end database can be exploited to develop efficient dialogue strategies. To extend our probabilistic framework and adapt our approach to real application scenarios, a user response model is investigated and integrated into the probabilistic framework to enhance the dialogue policy in this paper. Experiments in both ideal setting and real user test setting are conducted to test the enhanced dialogue policy. The results show that despite an unavoidable mismatch between the user response model based on prior knowledge and real users' behaviors in the experiment, the enhanced dialogue policy works robustly and efficiently. The results further demonstrate that the probabilistic framework is quite flexible and amenable to the integration of additional factors and models of real-world dialogue problems.

Dialogue Session Segmentation by Embedding-Enhanced TextTiling

Yiping Song[1], Lili Mou[1], Rui Yan[1], Li Yi[2], Zinan Zhu[2], Xiaohua Hu[2], Ming Zhang[1]; [1]Peking University, China; [2]Central China Normal University, China

Sun-P-7-4-2, Time: 13:30

In human-computer conversation systems, the context of a user-issued utterance is particularly important because it provides useful background information of the conversation. However, it is unwise to track all previous utterances in the current session as not all of them are equally important. In this paper, we address the problem of session segmentation. We propose an embedding-enhanced TextTiling approach, inspired by the observation that conversation utterances are highly noisy, and that word embeddings provide a robust way of capturing semantics. Experimental results show that our approach achieves better performance than the TextTiling, MMD approaches.

Target-Based State and Tracking Algorithm for Spoken Dialogue System

Miao Li, Zhiyang He, Ji Wu; Tsinghua University, China

Sun-P-7-4-3, Time: 13:30

Conventional spoken dialogue systems use frame structure to represent dialogue state. In this paper, we argue that using target distribution to represent dialogue state is much better than using frame structure. Based on the proposed target-based state, two target-based state tracking algorithms are introduced. Experiments in an end-to-end spoken dialogue system with real users are conducted to compare the performance between the target-based state trackers and frame-based state trackers. The experimental results show that the proposed target-based state tracker achieve 97% of dialogue success rate, comparing to 81% of frame-based state tracker, which suggests the advantage of target-based state.

Neural Attention Models for Sequence Classification: Analysis and Application to Key Term Extraction and Dialogue Act Detection

Sheng-syun Shen, Hung-Yi Lee; National Taiwan University, Taiwan

Sun-P-7-4-4, Time: 13:30

Recurrent neural network architectures combining with attention mechanism, or neural attention model, have shown promising performance recently for the tasks including speech recognition, image caption generation, visual question answering and machine translation. In this paper, neural attention model is applied on two sequence labeling tasks, dialogue act detection and key term extraction. In the sequence labeling tasks, the model input is a sequence, and the output is the label of the input sequence. The major difficulty of sequence labeling is that when the input sequence is long, it can include many noisy or irrelevant part. If the information in the whole sequence is treated equally, the noisy or irrelevant part may degrade the classification performance. The attention mechanism is helpful for sequence classification task because it is capable of highlighting important part among the entire sequence for the classification task. The experimental results show that with the attention mechanism, discernible improvements were achieved in the sequence labeling task considered here. The roles of the attention mechanism in the tasks are further analyzed and visualized in this paper.

NOTES

Objective Language Feature Analysis in Children with Neurodevelopmental Disorders During Autism Assessment

Manoj Kumar[1], Rahul Gupta[1], Daniel Bone[1], Nikolaos Malandrakis[1], Somer Bishop[2], Shrikanth S. Narayanan[1]; [1]University of Southern California, USA; [2]University of California at San Francisco, USA

Sun-P-7-4-5, Time: 13:30

Lexical planning is an important part of communication and is reflective of a speaker's internal state that includes aspects of affect, mood, as well as mental health. Within the study of developmental disorders such as autism spectrum disorder (ASD), language acquisition and language use have been studied to assess disorder severity and expressive capability as well as to support diagnosis. In this paper, we perform a language analysis of children focusing on word usage, social and cognitive linguistic word counts, and a few recently proposed psycho-linguistic norms. We use data from conversational samples of verbally fluent children obtained during Autism Diagnostic Observation Schedule (ADOS) sessions. We extract the aforementioned lexical cues from transcripts of session recordings and demonstrate their role in differentiating children diagnosed with Autism Spectrum Disorder from the rest. Further, we perform a correlation analysis between the lexical norms and ASD symptom severity. The analysis reveals an increased affinity by the interlocutor towards use of words with greater feminine association and negative valence.

Improving Generalisation to New Speakers in Spoken Dialogue State Tracking

Iñigo Casanueva, Thomas Hain, Phil Green; University of Sheffield, UK

Sun-P-7-4-6, Time: 13:30

Users with disabilities can greatly benefit from personalised voice-enabled environmental-control interfaces, but for users with speech impairments (e.g. dysarthria) poor ASR performance poses a challenge to successful dialogue. Statistical dialogue management has shown resilience against high ASR error rates, hence making it useful to improve the performance of these interfaces. However, little research was devoted to dialogue management personalisation to specific users so far. Recently, data driven discriminative models have been shown to yield the best performance in dialogue state tracking (the inference of the user goal from the dialogue history). However, due to the unique characteristics of each speaker, training a system for a new user when user specific data is not available can be challenging due to the mismatch between training and working conditions. This work investigates two methods to improve the performance with new speakers of a LSTM-based personalised state tracker: The use of speaker specific acoustic and ASR-related features; and dropout regularisation. It is shown that in an environmental control system for dysarthric speakers, the combination of both techniques yields improvements of 3.5% absolute in state tracking accuracy. Further analysis explores the effect of using different amounts of speaker specific data to train the tracking system.

Towards Machine Comprehension of Spoken Content: Initial TOEFL Listening Comprehension Test by Machine

Bo-Hsiang Tseng, Sheng-syun Shen, Hung-Yi Lee, Lin-Shan Lee; National Taiwan University, Taiwan

Sun-P-7-4-7, Time: 13:30

Multimedia or spoken content presents more attractive information than plain text content, but it's more difficult to display on a screen and be selected by a user. As a result, accessing large collections of the former is much more difficult and time-consuming than the latter for humans. It's highly attractive to develop a machine which can automatically understand spoken content and summarize the key information for humans to browse over. In this endeavor, we propose a new task of machine comprehension of spoken content. We define the initial goal as the listening comprehension test of TOEFL, a challenging academic English examination for English learners whose native language is not English. We further propose an Attention-based Multi-hop Recurrent Neural Network (AMRNN) architecture for this task, achieving encouraging results in the initial tests. Initial results also have shown that word-level attention is probably more robust than sentence-level attention for this task with ASR errors.

Sun-O-8-1 : Topics in Speech Recognition

Grand Ballroom A, 16:00–18:00, Sunday, 11 Sept. 2016
Chairs: Karen Livescu, Eric McDermott

How Neural Network Depth Compensates for HMM Conditional Independence Assumptions in DNN-HMM Acoustic Models

Suman Ravuri[1], Steven Wegmann[2]; [1]ICSI, USA; [2]Semantic Machines, USA

Sun-O-8-1-1, Time: 16:00

While DNN-HMM acoustic models have replaced GMM-HMMs in the standard ASR pipeline due to performance improvements, one unrealistic assumption that remains in these models is the conditional independence assumption of the Hidden Markov Model (HMM). In this work, we explore the extent to which depth of neural networks helps compensate for these poor conditional independence assumptions. Using a bootstrap resampling framework that allows us to control the amount of data dependence in the test set while still using real observations from the data, we can determine how robust neural networks, and particularly deeper models, are to data dependence. Our conclusions are that if the data were to match the conditional independence assumptions of the HMM, there would be little benefit from using deeper models. It is only when data become more dependent that depth improves ASR performance. That performance substantially degrades, however, as the data becomes more realistic suggests that better temporal modeling is still needed for ASR.

Jointly Learning to Locate and Classify Words Using Convolutional Networks

Dimitri Palaz[1], Gabriel Synnaeve[2], Ronan Collobert[1]; [1]Facebook, USA; [2]Facebook, France

Sun-O-8-1-2, Time: 16:20

In this paper, we propose a novel approach for weakly-supervised

NOTES

word recognition. Most state of the art automatic speech recognition systems are based on frame-level labels obtained through forced alignments or through a sequential loss. Recently, weakly-supervised trained models have been proposed in vision, that can learn which part of the input is relevant for classifying a given pattern [1]. Our system is composed of a convolutional neural network and a temporal score aggregation mechanism. For each sentence, it is trained using as supervision only some of the words (most frequent) that are present in a given sentence, without knowing their order nor quantity. We show that our proposed system is able to jointly classify and localize words. We also evaluate the system on a keyword spotting task, and show that it can yield similar performance to strong supervised HMM/GMM baseline.

On the Efficient Representation and Execution of Deep Acoustic Models

Raziel Alvarez, Rohit Prabhavalkar, Anton Bakhtin; Google, USA

Sun-O-8-1-3, Time: 16:40

In this paper we present a simple and computationally efficient quantization scheme that enables us to reduce the resolution of the parameters of a neural network from 32-bit floating point values to 8-bit integer values. The proposed quantization scheme leads to significant memory savings and enables the use of optimized hardware instructions for integer arithmetic, thus significantly reducing the cost of inference. Finally, we propose a 'quantization aware' training process that applies the proposed scheme during network training and find that it allows us to recover most of the loss in accuracy introduced by quantization. We validate the proposed techniques by applying them to a long short-term memory-based acoustic model on an open-ended large vocabulary speech recognition task.

Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI

Daniel Povey[1], Vijayaditya Peddinti[1], Daniel Galvez[2], Pegah Ghahremani[1], Vimal Manohar[1], Xingyu Na[3], Yiming Wang[1], Sanjeev Khudanpur[1]; [1]Johns Hopkins University, USA; [2]Cornell University, USA; [3]Lele Innovation & Intelligence Technology, China

Sun-O-8-1-4, Time: 17:00

In this paper we describe a method to perform sequence-discriminative training of neural network acoustic models without the need for frame-level cross-entropy pre-training. We use the lattice-free version of the maximum mutual information (MMI) criterion: LF-MMI. To make its computation feasible we use a phone n-gram language model, in place of the word language model. To further reduce its space and time complexity we compute the objective function using neural network outputs at one third the standard frame rate. These changes enable us to perform the computation for the forward-backward algorithm on GPUs. Further the reduced output frame-rate also provides a significant speed-up during decoding.

We present results on 5 different LVCSR tasks with training data ranging from 100 to 2100 hours. Models trained with LF-MMI provide a relative word error rate reduction of ~11.5%, over those trained with cross-entropy objective function, and ~8%, over those trained with cross-entropy and sMBR objective functions. A further reduction of ~2.5%, relative, can be obtained by fine tuning these models with the word-lattice based sMBR objective function.

Virtual Adversarial Training Applied to Neural Higher-Order Factors for Phone Classification

Martin Ratajczak[1], Sebastian Tschiatschek[2], Franz Pernkopf[1]; [1]Technische Universität Graz, Austria; [2]ETH Zürich, Switzerland

Sun-O-8-1-5, Time: 17:20

We explore virtual adversarial training (VAT) applied to neural higher-order conditional random fields for sequence labeling. VAT is a recently introduced regularization method promoting local distributional smoothness: It counteracts the problem that predictions of many state-of-the-art classifiers are unstable to adversarial perturbations. Unlike random noise, adversarial perturbations are minimal and bounded perturbations that flip the predicted label. We utilize VAT to regularize neural higher-order factors in conditional random fields. These factors are for example important for phone classification where phone representations strongly depend on the context phones. However, without using VAT for regularization, the use of such factors was limited as they were prone to overfitting. In extensive experiments, we successfully apply VAT to improve performance on the TIMIT phone classification task. In particular, we achieve a phone error rate of 13.0%, exceeding the state-of-the-art performance by a wide margin.

Sequence Student-Teacher Training of Deep Neural Networks

Jeremy H.M. Wong, Mark J.F. Gales; University of Cambridge, UK

Sun-O-8-1-6, Time: 17:40

The performance of automatic speech recognition can often be significantly improved by combining multiple systems together. Though beneficial, ensemble methods can be computationally expensive, often requiring multiple decoding runs. An alternative approach, appropriate for deep learning schemes, is to adopt student-teacher training. Here, a student model is trained to reproduce the outputs of a teacher model, or ensemble of teachers. The standard approach is to train the student model on the frame posterior outputs of the teacher. This paper examines the interaction between student-teacher training schemes and sequence training criteria, which have been shown to yield significant performance gains over frame-level criteria. There are several possible options for integrating sequence training, including training of the ensemble and further training of the student. This paper also proposes an extension to the student-teacher framework, where the student is trained to emulate the hypothesis posterior distribution of the teacher, or ensemble of teachers. This sequence student-teacher training approach allows the benefit of student-teacher training to be directly combined with sequence training schemes. These approaches are evaluated on two speech recognition tasks: a Wall Street Journal based task and a low-resource Tok Pisin conversational telephone speech task from the IARPA Babel programme.

NOTES

Sun-O-8-2 : Special Session: Realism in Robust Speech Processing

Grand Ballroom BC, 16:00–18:00, Sunday, 11 Sept. 2016
Chairs: Dayana Ribas, Emmanuel Vincent, John Hansen

Robustness in Speech, Speaker, and Language Recognition: "You've Got to Know Your Limitations"

John H.L. Hansen, Hynek Bořil; University of Texas at Dallas, USA

Sun-O-8-2-1, Time: 16:00

In the field of speech, speaker and language recognition, significant gains have and are being made with new machine learning strategies along with the availability of new and emerging speech corpora. However, many of the core scientific principles required for effective speech processing research appear to be drifting to the sidelines with the assumptions that access to larger amounts of data can address a growing range of issues relating to new speech/speaker/language recognition scenarios. This study focuses on exploring several challenging domains in formulating effective solutions in realistic speech data, and in particular the notion of using naturalistic data to better reflect the potential effectiveness of new algorithms. Our main focus is on mismatch/speech variability issues due to (i) differences in noisy speech with and without Lombard effect and a communication factor, (ii) realistic field data in noisy/increased cognitive load conditions, and (iii) dialect identification using found data. Finally, we study speaker–noise and speaker–speaker interactions in a newly established, fully naturalistic Prof-Life-Log corpus. The specific outcomes from this study include an analysis of the strengths and weaknesses of simulated vs. actual speech data collection for research.

The Use of Read versus Conversational Lombard Speech in Spectral Tilt Modeling for Intelligibility Enhancement in Near-End Noise Conditions

Emma Jokinen, Ulpu Remes, Paavo Alku; Aalto University, Finland

Sun-O-8-2-2, Time: 16:15

Intelligibility of speech in adverse near-end noise conditions can be enhanced with post-processing. Recently, a post-processing method based on statistical mapping of the spectral tilt of normal speech to that of Lombard speech was proposed. However, previous intelligibility improvement studies utilizing Lombard speech have mainly gathered data from read sentences which might result in a less pronounced Lombard effect. Having a mild Lombard effect in the training data weakens the statistical normal-to-Lombard mapping of the spectral tilt which in turn deteriorates performance of intelligibility enhancement. Therefore, a database containing both conversational and read Lombard speech was recorded in several background noise conditions in this study. Statistical models for normal-to-Lombard mapping of the spectral tilt were then trained using the obtained conversational and read speech data and evaluated using an objective intelligibility metric. The results suggest that the conversational data contains a more pronounced Lombard effect and could be used to obtain better statistical models for intelligibility enhancement.

Corpora for the Evaluation of Robust Speaker Recognition Systems

Douglas E. Sturim, Pedro A. Torres-Carrasquillo, Joseph P. Campbell; MIT Lincoln Laboratory, USA

Sun-O-8-2-3, Time: 16:30

The goal of this paper is to describe significant corpora available to support speaker recognition research and evaluation, along with details about the corpora collection and design. We describe the attributes of high-quality speaker recognition corpora. Considerations of the application, domain, and performance metrics are also discussed. Additionally, a literature survey of corpora used in speaker recognition research over the last 10 years is presented. Finally we show the most common corpora used in the research community and review them on their success in enabling meaningful speaker recognition research.

A French Corpus for Distant-Microphone Speech Processing in Real Homes

Nancy Bertin[1], Ewen Camberlein[1], Emmanuel Vincent[2], Romain Lebarbenchon[1], Stéphane Peillon[3], Éric Lamande[3], Sunit Sivasankaran[2], Frédéric Bimbot[1], Irina Illina[4], Ariane Tom[5], Sylvain Fleury[5], Éric Jamet[5]; [1]IRISA, France; [2]Inria, France; [3]VoiceBox Technologies, France; [4]LORIA, France; [5]CRPCC (EA 1285), France

Sun-O-8-2-4, Time: 16:45

We introduce a new corpus for distant-microphone speech processing in domestic environments. This corpus includes reverberated, noisy speech signals spoken by native French talkers in a lounge and recorded by an 8-microphone device at various angles and distances and in various noise conditions. Room impulse responses and noise-only signals recorded in various real rooms and homes and baseline speaker localization and enhancement software are also provided. This corpus stands apart from other corpora in the field by the number of rooms and homes considered and by the fact that it is publicly available at no cost. We describe the corpus specifications and annotations and the data recorded so far. We report baseline results.

Realistic Multi-Microphone Data Simulation for Distant Speech Recognition

Mirco Ravanelli, Piergiorgio Svaizer, Maurizio Omologo; FBK, Italy

Sun-O-8-2-5, Time: 17:00

The availability of realistic simulated corpora is of key importance for the future progress of distant speech recognition technology. The reliability, flexibility and low computational cost of a data simulation process may ultimately allow researchers to train, tune and test different techniques in a variety of acoustic scenarios, avoiding the laborious effort of directly recording real data from the targeted environment.

In the last decade, several simulated corpora have been released to the research community, including the data-sets distributed in the context of projects and international challenges, such as CHiME and REVERB. These efforts were extremely useful to derive baselines and common evaluation frameworks for comparison purposes. At the same time, in many cases they highlighted the need of a better

Notes

coherence between real and simulated conditions.

In this paper, we examine this issue and we describe our approach to the generation of realistic corpora in a domestic context. Experimental validation, conducted in a multi-microphone scenario, shows that a comparable performance trend can be observed with both real and simulated data across different recognition frameworks, acoustic models, as well as multi-microphone processing techniques.

Synthesis of Device-Independent Noise Corpora for Realistic ASR Evaluation

Hannes Gamper, Mark R.P. Thomas, Lyle Corbin, Ivan Tashev; Microsoft, USA

Sun-O-8-2-6, Time: 17:15

In order to effectively evaluate the accuracy of automatic speech recognition (ASR) with a novel capture device, it is important to create a realistic test data corpus that is representative of real-world noise conditions. Typically, this involves either recording the output of a device under test (DUT) in a noisy environment, or synthesizing an environment over loudspeakers in a way that simulates realistic signal-to-noise ratios (SNRs), reverberation times, and spatial noise distributions. Here we propose a method that aims at combining the realism of in-situ recordings with the convenience and repeatability of synthetic corpora. A device-independent spatial recording containing noise and speech is combined with the measured directivity pattern of a DUT to generate a synthetic test corpus for evaluating the performance of an ASR system. This is achieved by a spherical harmonic decomposition of both the sound field and the DUT's directivity patterns. Experimental results suggest that the proposed method can be a viable alternative to costly and cumbersome device-dependent measurements. The proposed simulation method predicted the SNR of the DUT response to within about 3 dB and the word error rate (WER) to within about 20%, across a range of test SNRs, target source directions, and noise types.

Speaker Recognition Using Real vs Synthetic Parallel Data for DNN Channel Compensation

Fred Richardson, Michael Brandstein, Jennifer Melot, Douglas Reynolds; MIT Lincoln Laboratory, USA

Sun-O-8-2-7, Time: 17:30

Recent work has shown large performance gains using denoising DNNs for speech processing tasks under challenging acoustic conditions. However, training these DNNs requires large amounts of parallel multichannel speech data which can be impractical or expensive to collect. The effective use of synthetic parallel data as an alternative has been demonstrated for several speech technologies including automatic speech recognition and speaker recognition (SR). This paper demonstrates that denoising DNNs trained with real Mixer 2 multichannel data perform only slightly better than DNNs trained with synthetic multichannel data for microphone SR on Mixer 6. Large reductions in pooled error rates of 50% EER and 30% min DCF are achieved using DNNs trained on real Mixer 2 data. Nearly the same performance gains are achieved using synthetic data generated with a limited number of room impulse responses (RIRs) and noise sources derived from Mixer 2. Using RIRs from three publicly available sources used in the Kaldi ASpIRE recipe yields somewhat lower pooled gains of 34% EER and 25% min DCF. These results confirm the effective use of synthetic parallel data for DNN channel compensation even when the RIRs used for synthesizing the data are not particularly well matched to the task.

Discussion

Dayana Ribas[1], Emmanuel Vincent[2], John H.L. Hansen[3], Emma Jokinen[4], Mirco Ravanelli[5], Hannes Gamper[6], Fred Richardson[7]; [1]CENATAV, Cuba; [2]Inria, France; [3]University of Texas at Dallas, USA; [4]Aalto University, Finland; [5]FBK, Italy; [6]Microsoft, USA; [7]MIT Lincoln Laboratory, USA

Sun-O-8-2-8, Time: 17:45

(No abstract available at the time of publication)

Sun-O-8-3 : Spoken Word Recognition

Bayview A, 16:00–18:00, Sunday, 11 Sept. 2016
Chairs: Michael McAuliffe, Odette Scharenborg

Combining Data-Oriented and Process-Oriented Approaches to Modeling Reaction Time Data

Louis ten Bosch, Lou Boves, M. Ernestus; Radboud Universiteit Nijmegen, The Netherlands

Sun-O-8-3-1, Time: 16:00

This paper combines two different approaches to modeling reaction time data from lexical decision experiments, viz. a data-oriented statistical analysis by means of a linear mixed effects model, and a process-oriented computational model of human speech comprehension.

The linear mixed effect model is implemented by lmer in R. As computational model we apply DIANA, an end-to-end computational model which aims at modeling the cognitive processes underlying speech comprehension. DIANA takes as input the speech signal, and provides as output the orthographic transcription of the stimulus, a word/non-word judgment and the associated reaction time. Previous studies have shown that DIANA shows good results for large-scale lexical decision experiments in Dutch and North-American English.

We investigate whether predictors that appear significant in an lmer analysis and processes implemented in DIANA can be related and inform both approaches. Predictors such as 'previous reaction time' can be related to a process description; other predictors, such as 'lexical neighborhood' are hard-coded in lmer and emergent in DIANA. The analysis focuses on the interaction between subject variables and task variables in lmer, and the ways in which these interactions can be implemented in DIANA.

Do Listeners Learn Better from Natural Speech?

Michael McAuliffe[1], Molly Babel[2], Charlotte Vaughn[3]; [1]McGill University, Canada; [2]University of British Columbia, Canada; [3]University of Oregon, USA

Sun-O-8-3-2, Time: 16:20

Perceptual learning of novel pronunciations is a seemingly robust and efficient process for adapting to unfamiliar speech patterns. In this study we compare perceptual learning of /s/ words where a medially occurring /s/ is substituted with /ʃ/, rendering, for example, *castle* as /kæʃl/ instead of /kæsl/. Exposure to the novel pronunciations is presented in the guise of a lexical decision task. Perceptual learning is assessed in a categorization task where listeners are presented with minimal pair continua (e.g., *sock-shock*). Given recent suggestions that perceptual learning may be more robust with

NOTES

natural as opposed to synthesized speech, we compare perceptual learning in groups that either receive natural /s/-to-/ʃ/ words or resynthesized /s/-to-/ʃ/ words. Despite low word endorsement rates in the lexical decision task, both groups of listeners show robust generalization in perceptual learning to the novel minimal pair continua, thereby indicating that at least with high quality resynthesis, perceptual learning in natural and synthesized speech is roughly equivalent.

Processing and Adaptation to Ambiguous Sounds during the Course of Perceptual Learning

Polina Drozdova, Roeland van Hout, Odette Scharenborg; Radboud Universiteit Nijmegen, The Netherlands

Sun-O-8-3-3, Time: 16:40

Listeners use their lexical knowledge to interpret ambiguous sounds, and retune their phonetic categories to include this ambiguous sound. Although there is ample evidence for lexically-guided retuning, the adaptation process is not fully understood. Using a lexical decision task with an embedded auditory semantic priming task, the present study investigates whether words containing an ambiguous sound are processed in the same way as “natural” words and whether adaptation to the ambiguous sound tends to equalize the processing of “ambiguous” and natural words. Analyses of the yes/no responses and reaction times to natural and “ambiguous” words showed that words containing an ambiguous sound were accepted as words less often and were processed slower than the same words without ambiguity. The difference in acceptance disappeared after exposure to approximately 15 ambiguous items. Interestingly, lower acceptance rates and slower processing did not have an effect on the processing of semantic information of the following word. However, lower acceptance rates of ambiguous primes predict slower reaction times of these primes, suggesting an important role of stimulus-specific characteristics in triggering lexically-guided perceptual learning.

The Effect of Background Noise on the Activation of Phonological and Semantic Information During Spoken-Word Recognition

Florian Hintz, Odette Scharenborg; Radboud Universiteit Nijmegen, The Netherlands

Sun-O-8-3-4, Time: 17:00

During spoken-word recognition, listeners experience phonological competition between multiple word candidates, which increases, relative to optimal listening conditions, when speech is masked by noise. Moreover, listeners activate semantic word knowledge during the word’s unfolding. Here, we replicated the effect of background noise on phonological competition and investigated to which extent noise affects the activation of semantic information in phonological competitors. Participants’ eye movements were recorded when they listened to sentences containing a target word and looked at three types of displays. The displays either contained a picture of the target word, or a picture of a phonological onset competitor, or a picture of a word semantically related to the onset competitor, each along with three unrelated distractors. The analyses revealed that, in noise, fixations to the target and to the phonological onset competitor were delayed and smaller in magnitude compared to the clean listening condition, most likely reflecting enhanced phonological competition. No evidence for the activation of semantic information in the phonological competitors was observed in noise and, surprisingly, also not in the clear. We discuss the implications of the lack of an effect and differences between the present and earlier studies.

Relationships Between Functional Load and Auditory Confusability Under Different Speech Environments

Shinae Kang[1], Clara Cohen[2]; [1]Georgetown University, USA; [2]Pennsylvania State University, USA

Sun-O-8-3-5, Time: 17:20

Functional load (FL) is an information-theoretic measure that captures a phoneme’s contribution to successful word identification. Experimental findings have shown that it can help explain patterns in perceptual accuracy. Here, we ask whether the relationship between FL and perception has larger consequences for the structure of a language’s lexicon. Since reducing FL minimizes the risk of misidentifying a word in the case where a listener inaccurately perceives the initial phoneme, we predicted that in spoken language, where perceptual accuracy is important for successful communication, the lexicon will be structured to reduce FL in auditorily confusable initial phonemes more than in written language. To test this prediction, we compared FL of all initial phonemes in spoken and academic written genres of the COCA corpus. We found that FL in phoneme pairs in the spoken corpus is overall higher and more variable than in the academic corpus, a natural consequence of the smaller lexical inventory characteristic of spoken language. In auditorily confusable pairs, however, this difference is relatively reduced, such that spoken FL decreases relative to academic FL. We argue that this reflects a pressure in spoken language to use words for which inaccurate perception does minimal damage to word identification.

The Role of Pitch in Punjabi Word Identification

Jasmeen Kanwal[1], Amanda Ritchart[2]; [1]University of Edinburgh, UK; [2]University of California at San Diego, USA

Sun-O-8-3-6, Time: 17:40

Previous work has argued that one class of consonants in Punjabi — those thought to be historically voiced aspirated — have now lost aspiration in all contexts and voicing in certain contexts. Word initially, these consonants are realized as voiceless unaspirated and are differentiated from other voiceless unaspirated consonants by a falling pitch on the following vowel. In this study, we investigate, using a two-alternative forced choice task, whether listeners make use of a falling pitch word-initially to distinguish between these two types of consonants that are otherwise phonetically identical. Our results show that, regardless of talker or listener, differences in falling pitch on the vowel following an unaspirated voiceless consonant are indeed sufficient for listeners to distinguish between words beginning with these consonants. These results provide further evidence that, in word-initial contexts, pitch may be in the process of phonologization in at least some dialects of Punjabi.

NOTES

Sun-O-8-4 : Speech Synthesis Oral: High Level Linguistic Features

Bayview B, 16:00–18:00, Sunday, 11 Sept. 2016
Chairs: Kishore Prahallad, Rob Clark

Improving TTS with Corpus-Specific Pronunciation Adaptation

Marie Tahon, Raheel Qader, Gwénolé Lecorvé, Damien Lolive; IRISA, France

Sun-O-8-4-1, Time: 16:00

Text-to-speech (TTS) systems are built on speech corpora which are labeled with carefully checked and segmented phonemes. However, phoneme sequences generated by automatic grapheme-to-phoneme converters during synthesis are usually inconsistent with those from the corpus, thus leading to poor quality synthetic speech signals. To solve this problem, the present work aims at adapting automatically generated pronunciations to the corpus. The main idea is to train corpus-specific phoneme-to-phoneme conditional random fields with a large set of linguistic, phonological, articulatory and acoustic-prosodic features. Features are first selected in cross-validation condition, then combined to produce the final best feature set. Pronunciation models are evaluated in terms of phoneme error rate and through perceptual tests. Experiments carried out on a French speech corpus show an improvement in the quality of speech synthesis when pronunciation models are included in the phonetization process. Apart from improving TTS quality, the presented pronunciation adaptation method also brings interesting perspectives in terms of expressive speech synthesis.

Deep Bidirectional Long Short-Term Memory Recurrent Neural Networks for Grapheme-to-Phoneme Conversion Utilizing Complex Many-to-Many Alignments

Amr El-Desoky Mousa, Björn Schuller; Universität Passau, Germany

Sun-O-8-4-2, Time: 16:20

Efficient grapheme-to-phoneme (G2P) conversion models are considered indispensable components to achieve the state-of-the-art performance in modern automatic speech recognition (ASR) and text-to-speech (TTS) systems. The role of these models is to provide such systems with a means to generate accurate pronunciations for unseen words. Recent work in this domain is based on recurrent neural networks (RNN) that are capable of translating grapheme sequences into phoneme sequences taking into account the full context of graphemes. To achieve high performance with these models, utilizing explicit alignment information is found essential. The quality of the G2P model heavily depends on the imposed alignment constraints.

In this paper, a novel approach is proposed using complex many-to-many G2P alignments to improve the performance of G2P models based on deep bidirectional long short-term memory (BLSTM) RNNs. Extensive experiments cover models with different numbers of hidden layers, projection layer, input splicing windows, and varying alignment schemes. One observes that complex alignments significantly improve the performance on the publicly available CMUDict US English dataset. We compare our results with previously published results.

Predicting Pronunciations with Syllabification and Stress with Recurrent Neural Networks

Daan van Esch, Mason Chua, Kanishka Rao; Google, USA

Sun-O-8-4-3, Time: 16:40

Word pronunciations, consisting of phoneme sequences and the associated syllabification and stress patterns, are vital for both speech recognition and text-to-speech (TTS) systems. For speech recognition phoneme sequences for words may be learned from audio data. We train recurrent neural network (RNN) based models to predict the syllabification and stress pattern for such pronunciations making them usable for TTS. We find these RNN models significantly outperform naive rule-based models for almost all languages we tested. Further, we find additional improvements to the stress prediction model by using the spelling as features in addition to the phoneme sequence. Finally, we train a single RNN model to predict the phoneme sequence, syllabification and stress for a given word. For several languages, this single RNN outperforms similar models trained specifically for either phoneme sequence or stress prediction. We report an exhaustive comparison of these approaches for twenty languages.

Adaptive Latency for Part-of-Speech Tagging in Incremental Text-to-Speech Synthesis

Maël Pouget, Olha Nahorna, Thomas Hueber, Gérard Bailly; GIPSA, France

Sun-O-8-4-4, Time: 17:00

Incremental text-to-speech systems aim at synthesizing a text 'on-the-fly', while the user is typing a sentence. In this context, this article addresses the problem of the part-of-speech tagging (POS, i.e. lexical category) which is a critical step for accurate grapheme-to-phoneme conversion and prosody estimation. Here, the main challenge is to estimate the POS of a given word without knowing its 'right context' (i.e. the following words which are not available yet). To address this issue, we propose a method based on a set of decision trees estimating online whether a given POS tag is likely to be modified when more right-contextual information becomes available. In such a case, the synthesis is delayed until POS stability is guaranteed. This results in delivering the synthetic voice in word chunks of variable length. Objective evaluation on French shows that the proposed method is able to estimate POS tags with more than a 92% accuracy (compared to a non-incremental system) while minimizing the synthesis latency (between 1 and 4 words). Perceptual evaluation (ranking test) is then carried in the context of HMM-based speech synthesis. Experimental results show that the word grouping resulting from the proposed method is rated more acceptable than word-by-word incremental synthesis.

Redefining the Linguistic Context Feature Set for HMM and DNN TTS Through Position and Parsing

Rasmus Dall[1], Kei Hashimoto[2], Keiichiro Oura[2], Yoshihiko Nankaku[2], Keiichi Tokuda[2]; [1]University of Edinburgh, UK; [2]Nagoya Institute of Technology, Japan

Sun-O-8-4-5, Time: 17:20

In this paper we present an investigation of a number of alternative linguistic feature context sets for HMM and DNN text-to-speech synthesis. The representation of positional values is explored

NOTES

through two alternatives to the standard set of absolute values, namely relational and categorical values. In a preference test the categorical representation was found to be preferred for both HMM and DNN synthesis. Subsequently, features based on probabilistic context free grammar and dependency parsing are presented. These features represent the phrase level relations between words in the sentences, and in a preference evaluation it was found that these features all improved upon the base set, with a combination of both parsing methods best overall. As the features primarily affected the F0 prediction, this illustrates the potential of syntactic structure to improve prosody in TTS.

Enhance the Word Vector with Prosodic Information for the Recurrent Neural Network Based TTS System

Xin Wang, Shinji Takaki, Junichi Yamagishi; NII, Japan

Sun-O-8-4-6, Time: 17:40

Word embedding, which is a dense and low-dimensional vector representation of word, is recently used to replace of the conventional prosodic context as an input feature to the acoustic model of a TTS system. However, these word vectors trained from text data may encode insufficient information related to speech. This paper presents a post-filtering approach to enhance the raw word vectors with prosodic information for the TTS task. Based on a publicly available speech corpus with manual prosodic annotation, a post-filter can be trained to transform the raw word vectors. Experiment shows that using the enhanced word vectors as an input to the neural network-based acoustic model improves the accuracy of the predicted F0 trajectory. Besides, we also show that the enhanced vectors provide better initial values than the raw vectors for error back-propagation of the network, which results in further improvement.

Sun-O-8-5 : Speech Enhancement

Seacliff BCD, 16:00–18:00, Sunday, 11 Sept. 2016
Chairs: Dimitrios Dimitriadis, Martin Cooke

Local Sparsity Based Online Dictionary Learning for Environment-Adaptive Speech Enhancement with Nonnegative Matrix Factorization

Kwang Myung Jeon, Hong Kook Kim; GIST, Korea

Sun-O-8-5-1, Time: 16:00

In this paper, a nonnegative matrix factorization (NMF)-based speech enhancement method robust to real and diverse noise is proposed by online NMF dictionary learning without relying on prior knowledge of noise. Conventional NMF-based methods have used a fixed noise dictionary, which often results in performance degradation when the NMF noise dictionary cannot cover noise types that occur in real-life recording. Thus, the noise dictionary needs to be learned from noises according to the variation of recording environments. To this end, the proposed method first estimates noise spectra and then performs online noise dictionary learning by a discriminative NMF learning framework. In particular, the noise spectra are estimated from minimum mean squared error filtering, which is based on the local sparsity defined by a posteriori signal-to-noise ratio (SNR) estimated from the NMF separation of the previous analysis frame. The effectiveness of the proposed speech enhancement method is demonstrated by adding six different realistic noises to clean speech signals with various SNRs. Consequently, it is shown that the proposed method outperforms comparative methods in terms of signal-to-distortion ratio (SDR) and perceptual evaluation of speech quality (PESQ) for all kinds of simulated noise and SNR conditions.

Noise Aware and Combined Noise Models for Speech Denoising in Unknown Noise Conditions

Pavlos Papadopoulos, Colin Vaz, Shrikanth S. Narayanan; University of Southern California, USA

Sun-O-8-5-2, Time: 16:20

Traditional denoising schemes require prior knowledge or statistics of the noise corrupting the signal, or estimate the noise from noise-only portions of the signal, which requires knowledge of speech boundaries. Extending denoising methods to perform well in unknown noise conditions can facilitate processing of data captured in different real life environments, and relax rigid data acquisition protocols. In this paper we propose two methods for denoising speech signals in unknown noise conditions. The first method has two stages. In the first stage we use Long Term Signal Variability features to decide which noise model to use from a pool of available models. Once we determine the noise type, we use Nonnegative Matrix Factorization with a dictionary trained on that noise to denoise the signal. In the second method, we create a combined noise dictionary from different types of noise, and use that dictionary in the denoising phase. Both of our systems improve signal quality, as measured by PESQ scores, for all the noise types we tested, and for different Signal to Noise Ratio levels.

Causal Speech Enhancement Combining Data-Driven Learning and Suppression Rule Estimation

Seyedmahdad Mirsamadi[1], Ivan Tashev[2]; [1]University of Texas at Dallas, USA; [2]Microsoft, USA

Sun-O-8-5-3, Time: 16:40

The problem of single-channel speech enhancement has been traditionally addressed by using statistical signal processing algorithms that are designed to suppress time-frequency regions affected by noise. We study an alternative data-driven approach which uses deep neural networks (DNNs) to learn the transformation from noisy and reverberant speech to clean speech, with a focus on real-time applications which require low-latency causal processing. We examine several structures in which deep learning can be used within an enhancement system. These include end-to-end DNN regression from noisy to clean spectra, as well as less intervening approaches which estimate a suppression gain for each time-frequency bin instead of directly recovering the clean spectral features. We also propose a novel architecture in which the general structure of a conventional noise suppressor is preserved, but the sub-tasks are independently learned and carried out by separate networks. It is shown that DNN-based suppression gain estimation outperforms the regression approach in the causal processing mode and for noise types that are not seen during DNN training.

NOTES

A Phase-Based Time-Frequency Masking for Multi-Channel Speech Enhancement in Domestic Environments

Alessio Brutti[1], Antigoni Tsiami[2], Athanasios Katsamanis[3], Petros Maragos[2]; [1]FBK, Italy; [2]NTUA, Greece; [3]Athena RIC, Greece

Sun-O-8-5-4, Time: 17:00

This paper introduces a novel time-frequency masking approach for speech enhancement, based on the consistency of the phase of the cross-spectrum observed at multiple microphones. The proposed approach is derived from solutions commonly adopted in spatial source separation and can be used as a post-filter in traditional multi-channel speech enhancement schemes. Since it is not based on a modeling of the coherence of diffuse noise, the proposed method complements traditional post-filters implementations, targeting non diffuse/coherent sources. It is particularly effective in domestic scenarios where microphones in a given room capture interfering coherent sources active in adjacent rooms.

An experimental analysis on the DIRHA-GRID corpus shows that the proposed method considerably improves the signal-to-interference-ratio and can be used on top of state-of-the-art multi-channel speech enhancement methods.

Generalizing Steady State Suppression for Enhanced Intelligibility Under Reverberation

Petko N. Petkov, Yannis Stylianou; Toshiba Research Europe, UK

Sun-O-8-5-5, Time: 17:20

Speech intelligibility in reverberant environments decreases due to overlap-masking. Unlike additive noise, the masking signal is not independent from the information bearing signal. A mathematical framework for intelligibility-enhancing signal modification prior to presentation in reverberant environments is presented in this paper. The optimal solution generalizes steady state suppression and adjusts the short-term signal power as a function of late reverberation power and signal importance. The signal modification operates in a full-band setting and preserves the time scale of the unmodified signal. Gain smoothing based on an adaptive rate-of-change constraint reduces processing artifacts and enhances performance. Subjective validation shows that the proposed method effectively reduces the impact of overlap-masking. Speech intelligibility at a reverberation time of 1.8 s was improved significantly compared to unmodified and steady-state-suppressed speech.

Speech Intelligibility Prediction Based on the Envelope Power Spectrum Model with the Dynamic Compressive Gammachirp Auditory Filterbank

Katsuhiko Yamamoto[1], Toshio Irino[1], Toshie Matsui[1], Shoko Araki[2], Keisuke Kinoshita[2], Tomohiro Nakatani[2]; [1]Wakayama University, Japan; [2]NTT, Japan

Sun-O-8-5-6, Time: 17:40

In this study, we develop a new method to realize speech intelligibility prediction of synthetic sounds processed by nonlinear speech enhancement algorithms. A speech envelope power spectrum model (sEPSM) was proposed to account for subjective results on a spectral subtraction, but it is untested by recent state-of-the-art speech enhancement algorithms. We introduce a dynamic compressive gammachirp auditory filterbank as the front-end of the sEPSM (dcGC-sEPSM) to improve the predictability. We perform subjective experiments on speech intelligibility (SI) of noise-reduced sounds processed by the spectral subtraction and a recently developed Wiener filter algorithm. We compare the subjective SI scores with the objective SI scores predicted by the proposed dcGC-sEPSM, the original GT-sEPSM, the three-level coherence SII (CSII), and the short-time objective intelligibility (STOI). The results show that the proposed dcGC-sEPSM performs better than the conventional models.

Sun-O-8-6 : Dialogue: Backchannels and Turntaking

Seacliff A, 16:00–18:00, Sunday, 11 Sept. 2016
Chairs: Zofia Malisz, David Traum

Prediction and Generation of Backchannel Form for Attentive Listening Systems

Tatsuya Kawahara[1], Takashi Yamaguchi[1], Koji Inoue[1], Katsuya Takanashi[1], Nigel Ward[2]; [1]Kyoto University, Japan; [2]University of Texas at El Paso, USA

Sun-O-8-6-1, Time: 16:00

In human-human dialogue, especially in attentive listening such as counseling, backchannels are important not only for smooth communication but also for establishing rapport. Despite several studies on when to backchannel, most of the current spoken dialogue systems generate the same pattern of backchannels, giving monotonous impressions to users. In this work, we investigate generation of a variety of backchannel forms according to the dialogue context. We first show the feasibility of choosing appropriate backchannel forms based on machine learning, and the synergy of using linguistic and prosodic features. For generation of backchannels, a framework based on a set of binary classifiers is adopted to effectively make a "not-to-generate" decision. The proposed model achieved better prediction accuracy than a baseline which always outputs the same backchannel form and another baseline which randomly generates backchannels. Finally, evaluations by human subjects demonstrate that the proposed method generates backchannels as naturally as human choices, giving impressions of understanding and empathy.

Measuring Turn-Taking Offsets in Human-Human Dialogues

Rebecca Lunsford, Peter A. Heeman, Emma Rennie; Oregon Health & Science University, USA

Sun-O-8-6-2, Time: 16:20

This paper examines the pauses, gaps and overlaps associated with turn-taking in order to better understand how people engage in this activity, which should lead to more natural and effective spoken dialogue systems. This paper makes three advances in studying these durations. First, we take into account the type of turn-taking event, carefully treating interruptions, dual starts, and delayed backchannels, as these can make it appear that turn-taking is more disorderly than it really is. Second, we do not view turn-transitions in isolation, but consider turn-transitions and turn-continuations together, as equal alternatives of what could have occurred. Third, we use the distributions of turn-transition and turn-continuation

Notes

offsets (gaps, overlaps, and pauses) to shed light on the extent to which turn-taking is negotiated by the two conversants versus controlled by the current speaker.

Using Past Speaker Behavior to Better Predict Turn Transitions

Tomer Meshorer, Peter A. Heeman; Oregon Health & Science University, USA

Sun-O-8-6-3, Time: 16:40

This paper explores using a summary of past speaker behavior to better predict turn transitions. We computed two types of summary features that represent the current speaker's past turn-taking behavior: relative turn length and relative floor control. Relative turn length measures the current turn length so far (in time and words) relative to the speaker's average turn length. Relative floor control measures the speaker's control of the conversation floor (in time and words) relative to the total conversation length. The features are recomputed for each dialog act based on past turns of the speaker within the current conversation. Using the switchboard corpus, we trained two models to predict turn transitions: one with just local features (e.g., current speech act, previous speech act) and one that added the summary features. Our results shows that using the summary features improve turn transitions prediction.

Quantitative Analysis of Backchannels Uttered by an Interviewer During Neuropsychological Tests

Gérard Bailly[1], Frédéric Elisei[1], Alexandra Juphard[2], Olivier Moreaud[2]; [1]GIPSA, France; [2]CHU de Grenoble, France

Sun-O-8-6-4, Time: 17:00

This paper examines in detail the backchannels uttered by a French professional interviewer during a neuropsychological test of verbal memories. These backchannels are short utterances such as *oui, d'accord, uhm,* etc. They are mainly produced here to encourage subjects to retrieve a set of words after their controlled encoding. We show that the choice of lexical items, their production rates and their associated prosodic contours are influenced by the subject performance and conditioned by the protocol.

Predicting User Satisfaction from Turn-Taking in Spoken Conversations

Shammur Absar Chowdhury, Evgeny A. Stepanov, Giuseppe Riccardi; Università di Trento, Italy

Sun-O-8-6-5, Time: 17:20

User satisfaction is an important aspect of the user experience while interacting with objects, systems or people. Traditionally user satisfaction is evaluated a-posteriori via spoken or written questionnaires or interviews. In automatic behavioral analysis we aim at measuring the user emotional states and its descriptions as they unfold during the interaction. In our approach, *user satisfaction* is modeled as the final state of a sequence of emotional states and given ternary values `positive, negative, neutral`. In this paper, we investigate the discriminating power of turn-taking in predicting user satisfaction in spoken conversations. Turn-taking is used for discourse organization of a conversation by means of explicit phrasing, intonation, and pausing. In this paper, we train different characterization of turn-taking, such as competitiveness of the speech overlaps. To extract turn-taking features we design a turn segmentation and labeling system that incorporates lexical and acoustic information. Given a human-human spoken dialog, our system automatically infers any of the three values of the state of the user satisfaction. We evaluate the classification system on real-life call-center human-human dialogs. The comparative performance analysis shows that the contribution of the turn-taking features outperforms both prosodic and lexical features.

Towards Building an Attentive Artificial Listener: On the Perception of Attentiveness in Feedback Utterances

Catharine Oertel[1], Joakim Gustafson[1], Alan W. Black[2]; [1]KTH, Sweden; [2]Carnegie Mellon University, USA

Sun-O-8-6-6, Time: 17:40

Current speech synthesizers typically lack backchannel tokens. Those synthesiser, which include backchannels, typically only support a limited set of stereotypical functions. However, this does not mirror the subtleties of backchannels in spontaneous conversations. If we want to be able to build an artificial listener, that can display degrees of attentiveness, we need a speech synthesizer with more fine-grained control of the prosodic realisations of its backchannels.

In the current study we used a corpus of three-party face-to-face discussions to sample backchannels produced under varying conversational dynamics. We wanted to understand i) which prosodic cues are relevant for the perception of varying degrees of attentiveness ii) how much of a difference is necessary for people to perceive a difference in attentiveness iii) whether a preliminary classifier could be trained to distinguish between more and less attentive backchannel token.

Sun-P-8-1 : Language Recognition

Pacific Concourse – Poster A, 16:00–18:00, Sunday, 11 Sept. 2016
Chair: Honza Cernocky

Language Recognition via Sparse Coding

Youngjune L. Gwon[1], William M. Campbell[1], Douglas E. Sturim[1], H.T. Kung[2]; [1]MIT Lincoln Laboratory, USA; [2]Harvard University, USA

Sun-P-8-1-1, Time: 16:00

Spoken language recognition requires a series of signal processing steps and learning algorithms to model distinguishing characteristics of different languages. In this paper, we present a sparse discriminative feature learning framework for language recognition. We use sparse coding, an unsupervised method, to compute efficient representations for spectral features from a speech utterance while learning basis vectors for language models. Differentiated from existing approaches in sparse representation classification, we introduce a maximum a posteriori (MAP) adaptation scheme based on online learning that further optimizes the discriminative quality of sparse-coded speech features. We empirically validate the effectiveness of our approach using the NIST LRE 2015 dataset.

Notes

A Feature Normalisation Technique for PLLR Based Language Identification Systems

Sarith Fernando, Vidhyasaharan Sethu, Eliathamby Ambikairajah; University of New South Wales, Australia

Sun-P-8-1-2, Time: 16:00

Phone log-likelihood ratio (PLLR) features have been shown to be effective in language identification systems. However, PLLR feature distributions are bounded and this may contradict assumptions of Gaussianity and consequently lead to reduced language recognition rates. In this paper, we propose a feature normalisation technique for the PLLR feature space and demonstrate that it can outperform conventional normalisation and decorrelation techniques such as mean-variance normalisation, feature warping, discrete cosine transform and principal component analysis. Experimental results on the NIST LRE 2007 and the NIST LRE 2015 databases show that the proposed method outperforms other normalisation methods by at least 9.3% in terms of %Cavg. Finally, unlike PCA which needs to be estimated from all the training data, the proposed technique can be applied on each utterance independently.

An Investigation of Deep Neural Network Architectures for Language Recognition in Indian Languages

Mounika K.V., Sivanand Achanta, Lakshmi H. R., Suryakanth V. Gangashetty, Anil Kumar Vuppala; IIIT Hyderabad, India

Sun-P-8-1-3, Time: 16:00

In this paper, deep neural networks are investigated for language identification in Indian languages. Deep neural networks (DNN) have been recently proposed for this task. However many architectural choices and training aspects that have been made while building such systems have not been studied carefully. We perform several experiments on a dataset consisting of 12 Indian languages with a total training data of about 120 hours in evaluating the effect of such choices.

While DNN based approach is inherently a frame based one, we propose an attention mechanism based DNN architecture for utterance level classification there by efficiently making use of the context. Evaluation of models were performed on 30 hours of testing data with 2.5 hours for each language. In our results, we find that deeper architectures outperform shallower counterparts. Also, DNN with attention mechanism outperforms the regular DNN models indicating the effectiveness of attention mechanism.

Automatic Dialect Detection in Arabic Broadcast Speech

Ahmed Ali[1], Najim Dehak[2], Patrick Cardinal[3], Sameer Khurana[1], Sree Harsha Yella[2], James Glass[2], Peter Bell[4], Steve Renals[4]; [1]Hamad Bin Khalifa University, Qatar; [2]MIT, USA; [3]École de Technologie Supérieure, Canada; [4]University of Edinburgh, UK

Sun-P-8-1-4, Time: 16:00

In this paper, we investigate different approaches for dialect identification in Arabic broadcast speech. These methods are based on phonetic and lexical features obtained from a speech recognition system, and bottleneck features using the i-vector framework. We studied both generative and discriminative classifiers, and we combined these features using a multi-class Support Vector Machine (SVM). We validated our results on an Arabic/English language identification task, with an accuracy of 100%. We also evaluated these features in a binary classifier to discriminate between Modern Standard Arabic (MSA) and Dialectal Arabic, with an accuracy of 100%. We further reported results using the proposed methods to discriminate between the five most widely used dialects of Arabic: namely Egyptian, Gulf, Levantine, North African, and MSA, with an accuracy of 59.2%. We discuss dialect identification errors in the context of dialect code-switching between Dialectal Arabic and MSA, and compare the error pattern between manually labeled data, and the output from our classifier. All the data used on our experiments have been released to the public as a language identification corpus.

Combining Weak Tokenisers for Phonotactic Language Recognition in a Resource-Constrained Setting

Raymond W.M. Ng, Bhusan Chettri, Thomas Hain; University of Sheffield, UK

Sun-P-8-1-5, Time: 16:00

In the phonotactic approach for language recognition, a phone tokeniser is normally used to transform the audio signal into acoustic tokens. The language identity of the speech is modelled by the occurrence statistics of the decoded tokens. The performance of this approach depends heavily on the quality of the audio tokeniser. A high-quality tokeniser in matched condition is not always available for a language recognition task. This study investigated into the performance of a phonotactic language recogniser in a resource-constrained setting, following NIST LRE 2015 specification. An ensemble of phone tokenisers was constructed by applying unsupervised sequence training on different target languages followed by a score-based fusion. This method gave 5–7% relative performance improvement to baseline system on LRE 2015 eval set. This gain was retained when the ensemble phonotactic system was further fused with an acoustic iVector system.

End-to-End Language Identification Using Attention-Based Recurrent Neural Networks

Wang Geng, Wenfu Wang, Yuanyuan Zhao, Xinyuan Cai, Bo Xu; Chinese Academy of Sciences, China

Sun-P-8-1-6, Time: 16:00

This paper proposes a novel attention-based recurrent neural network (RNN) to build an end-to-end automatic language identification (LID) system. Inspired by the success of attention mechanism on a range of sequence-to-sequence tasks, this work introduces the attention mechanism with long short term memory (LSTM) encoder to the sequence-to-tag LID task. This unified architecture extends the end-to-end training method to LID system and dramatically boosts the system performance. Firstly, a language category embedding module is used to provide attentional vector which guides the derivation of the utterance level representation. Secondly, two attention approaches are explored: a soft attention which attends all source frames and a hard one that focuses on a subset of the sequential input. Thirdly, a hybrid test method which traverses all gold labels is adopted in the inference phase. Experimental results show that 8.2% relative equal error rate (EER) reduction is obtained compared with the LSTM-based frame level system by the soft approach and 34.33% performance improvement is observed compared to the conventional i-Vector system.

Notes

Enhancing Multilingual Recognition of Emotion in Speech by Language Identification

Hesam Sagha [1], Pavel Matějka [2], Maryna Gavryukova [1], Filip Povolny [2], Erik Marchi [1], Björn Schuller [1]; [1] Universität Passau, Germany; [2] Phonexia Brno, Czech Republic

Sun-P-8-1-7, Time: 16:00

We investigate, for the first time, if applying model selection based on automatic language identification (LID) can improve multilingual recognition of emotion in speech. Six emotional speech corpora from three language families (Germanic, Romance, Sino-Tibetan) are evaluated. The emotions are represented by the quadrants in the arousal/valence plane, i. e., positive/ negative arousal/valence. Four selection approaches for choosing an optimal training set depending on the current language are compared: within the same language family, across language family, use of all available corpora, and selection based on the automatic LID. We found that, on average, the proposed LID approach for selecting training corpora is superior to using all the available corpora when the spoken language is not known.

Sun-P-8-2 : Speech and Audio Segmentation and Classification

Pacific Concourse – Poster B, 16:00–18:00, Sunday, 11 Sept. 2016
Chair: Martine Adda-Decker

Deep Neural Network Bottleneck Features for Acoustic Event Recognition

Seongkyu Mun [1], Suwon Shon [1], Wooil Kim [2], Hanseok Ko [1]; [1] Korea University, Korea; [2] Incheon National University, Korea

Sun-P-8-2-1, Time: 16:00

Bottleneck features have been shown to be effective in improving the accuracy of speaker recognition, language identification and automatic speech recognition. However, few works have focused on bottleneck features for acoustic event recognition. This paper proposes a novel acoustic event recognition framework using bottleneck features derived from a Deep Neural Network (DNN). In addition to conventional features (MFCC, Mel-spectrum, etc.), this paper employs rhythm, timbre, and spectrum-statistics features for effectively extracting acoustic characteristics from audio signals. The effectiveness of the proposed method is demonstrated on a database of real life recordings via experiments, and its robust performance is verified by comparing to conventional methods.

Combining Energy and Cross-Entropy Analysis for Nuclear Segments Detection

Antonio Origlia, Francesco Cutugno; Università di Napoli Federico II, Italy

Sun-P-8-2-2, Time: 16:00

Features related to rhythmic patterns are involved in the representation of the intonational content for spoken language analysis. Among others, speech rate is one of the most used measures extracted by systems using prosodic analysis and is typically measured in syllables per second. Automatic approaches designed to estimate this measure in absence of manual annotations usually mark the position of syllable nuclei as a single point in time. Approaches extracting duration features using automatic segmentation in units shorter than words but larger than phones tend to detect syllables. To represent the prosodic contents of an utterance, especially from the rhythmic point of view, automatic positioning of nuclear boundaries may, however, be more informative than syllable boundaries. In this paper we present a method combining the analysis of the energy envelope and of the cross-entropy profile to obtain a segmentation into nuclear and inter-nuclear segments, showing that the proposed method can be used to obtain a reliable estimate of speech rate and that accuracy in nuclear boundary positioning allows the extraction of segmental features useful for automatic prosodic analysis.

Anchored Speech Detection

Roland Maas, Sree Hari Krishnan Parthasarathi, Brian King, Ruitong Huang, Björn Hoffmeister; Amazon.com, USA

Sun-P-8-2-3, Time: 16:00

We propose two new methods of speech detection in the context of voice-controlled far-field appliances. While conventional detection methods are designed to differentiate between speech and nonspeech, we aim at distinguishing *desired speech*, which we define as speech originating from the person interacting with the device, from background noise and interfering talkers. Our two proposed methods use the first word spoken by the desired talker, the “anchor” word, as a reference to learn characteristics about that speaker. In the first method, we estimate the mean of the anchor word segment and subtract it from the subsequent feature vectors. In the second, we use an encoder-decoder network with features that are normalized by applying conventional log amplitude causal mean subtraction. The experimental results reveal that both techniques achieve around 10% relative reduction in frame classification error rate over a baseline feed-forward network with conventionally normalized features.

Towards Smart-Cars That Can Listen: Abnormal Acoustic Event Detection on the Road

Mahesh Kumar Nandwana, Taufiq Hasan; Robert Bosch, USA

Sun-P-8-2-4, Time: 16:00

Even with the recent technological advancements in smart-cars, safety is still a major challenge in autonomous driving. State-of-the-art self-driving vehicles mostly rely on visual, ultrasonic and radar sensors to assess the surroundings and make decisions. However, in certain driving scenarios, the best modality for context awareness is environmental sound. In this study, we propose an acoustic event recognition framework for detecting abnormal audio events on the road. We consider five classes of audio events, namely, ambulance siren, railroad crossing bell, tire screech, car honk, and glass break. We explore various generative and discriminative back-end classifiers, utilizing Gaussian Mixture Models (GMM), GMM mean supervectors and the I-vector framework. Evaluation results using the proposed strategy validate the effectiveness of the proposed system.

Notes

Hierarchical Classification of Speaker and Background Noise and Estimation of SNR Using Sparse Representation

K.V. Vijay Girish[1], *A.G. Ramakrishnan*[1], *T.V. Ananthapadmanabha*[2]; [1]*Indian Institute of Science, India;* [2]*Voice and Speech Systems, India*

Sun-P-8-2-5, Time: 16:00

In the analysis of recordings of conversations, one of the motivations is to be able to identify the nature of background noise as a means of identifying the possible geographical location of a speaker. In a high noise environment, to minimize manual analysis of the recording, it is also desirable to automatically locate only the segments of the recording, which contain speech. The next task is to identify if the speech is from one of the known people. A dictionary learning and block sparsity based source recovery approach has been used to estimate the SNR of a noisy speech recording, simulated at different SNRs using ten different noise sources. Given a test utterance, a noise label is assigned using block sparsity approach, and subsequently, the speaker is classified using sum of weights recovered from the concatenation of speaker dictionaries and the identified noise source dictionary. Using the dictionaries of the identified speaker and noise sources, framewise speech and noise energy are estimated using a source recovery method. The energy estimates are then used to identify the segments, where speech is present. We obtain 100% accuracy for background classification and around 90% for speaker classification at a SNR of 10 dB.

Robust Sound Event Detection in Continuous Audio Environments

Haomin Zhang[1], *Ian McLoughlin*[2], *Yan Song*[1]; [1]*USTC, China;* [2]*University of Kent, UK*

Sun-P-8-2-6, Time: 16:00

Sound event detection in real world environments has attracted significant research interest recently because of it's applications in popular fields such as machine hearing and automated surveillance, as well as in sound scene understanding. This paper considers continuous robust sound event detection, which means multiple overlapped sound events in different types of interfering noise. First, a standard evaluation task is outlined based upon existing testing data sets for the sound event classification of isolated sounds. This paper then proposes and evaluates the use of spectrogram image features employing an energy detector to segment sound events, before developing a novel segmentation method making use of a Bayesian inference criteria. At the back end, a convolutional neural network is used to classify detected regions, and this combination is compared to several alternative approaches. The proposed method is shown capable of achieving very good performance compared with current state-of-the-art techniques.

Deep Convolutional Neural Networks and Data Augmentation for Acoustic Event Recognition

Naoya Takahashi[1], *Michael Gygli*[2], *Beat Pfister*[2], *Luc Van Gool*[2]; [1]*Sony, Japan;* [2]*ETH Zürich, Switzerland*

Sun-P-8-2-7, Time: 16:00

We propose a novel method for Acoustic Event Recognition (AER). In contrast to speech, sounds coming from acoustic events may be produced by a wide variety of sources. Furthermore, distinguishing them often requires analyzing an extended time period due to the lack of a clear sub-word unit. In order to incorporate the long-time frequency structure for AER, we introduce a convolutional neural network (CNN) with a large input field. In contrast to previous works, this enables to train audio event detection end-to-end. Our architecture is inspired by the success of VGGNet [1] and uses small, 3×3 convolutions, but more depth than previous methods in AER. In order to prevent over-fitting and to take full advantage of the modeling capabilities of our network, we further propose a novel data augmentation method to introduce data variation. Experimental results show that our CNN significantly outperforms state of the art methods including Bag of Audio Words (BoAW) and classical CNNs, achieving a 16% absolute improvement.

Artificial Neural Network-Based Feature Combination for Spatial Voice Activity Detection

Stefan Meier, Walter Kellermann; FAU Erlangen-Nürnberg, Germany

Sun-P-8-2-8, Time: 16:00

For many applications in speech communications and speech-based human-machine interaction, a reliable Voice Activity Detection (VAD) is crucial. Conventional methods for VAD typically differentiate between a target speaker and background noise by exploiting characteristic properties of speech signals. If a target speaker should be distinguished from other speech sources, these conventional concepts are no longer applicable, and other methods, typically exploiting the spatial diversity of the individual sources, are required. Often, it is beneficial to combine several features in order to improve the overall decision. Optimum combinations of features, however, depend strongly on the scenario, especially on the position of the target source, the characteristics of noise and interference and the Signal-to-Interference Ratio (SIR). Moreover, choosing detection thresholds which are robust to changing scenarios is often a difficult problem. In this paper, these issues are addressed by introducing Artificial Neural Networks (ANNs) for spatial voice activity detection, which allow to combine several features with background information. The experimental results show that already small ANNs can significantly and robustly improve the detection rates, offering a valuable tool for VAD.

HAPPY Team Entry to NIST OpenSAD Challenge: A Fusion of Short-Term Unsupervised and Segment i-Vector Based Speech Activity Detectors

Tomi Kinnunen[1], *Alexey Sholokhov*[1], *Elie Khoury*[2], *Dennis Alexander Lehmann Thomsen*[3], *Md. Sahidullah*[1], *Zheng-Hua Tan*[3]; [1]*University of Eastern Finland, Finland;* [2]*Pindrop, USA;* [3]*Aalborg University, Denmark*

Sun-P-8-2-9, Time: 16:00

Speech activity detection (SAD), the task of locating speech segments from a given recording, remains challenging under acoustically degraded conditions. In 2015, National Institute of Standards and Technology (NIST) coordinated OpenSAD bench-mark. We summarize "HAPPY" team effort to OpenSAD. SADs come in both unsupervised and supervised flavors, the latter requiring a labeled training set. Our solution fuses six base SADs (2 supervised and 4 unsupervised). The individually best SAD, in terms of detection cost function (DCF), is supervised and uses adaptive segmentation with

NOTES

i-vectors to represent the segments. Fusion of the six base SADs yields a relative decrease of 9.3% in DCF over this SAD. Further, relative decrease of 17.4% is obtained by incorporating channel detection side information.

Manual versus Automated: The Challenging Routine of Infant Vocalisation Segmentation in Home Videos to Study Neuro(mal)development

Florian B. Pokorny[1], Robert Peharz[1], Wolfgang Roth[2], Matthias Zöhrer[2], Franz Pernkopf[3], Peter B. Marschik[1], Björn Schuller[4]; [1]Medizinische Universität Graz, Austria; [2]Technische Universität Graz, Austria; [3]BioTechMed-Graz, Austria; [4]Universität Passau, Germany

Sun-P-8-2-10, Time: 16:00

In recent years, voice activity detection has been a highly researched field, due to its importance as input stage in many real-world applications. Automated detection of vocalisations in the very first year of life is still a stepchild of this field. On our quest defining acoustic parameters in pre-linguistic vocalisations as markers for neuro(mal)development, we are confronted with the challenge of manually segmenting and annotating hours of variable quality home video material for sequences of infant voice/vocalisations. While in total our corpus comprises video footage of typically developing infants and infants with various neurodevelopmental disorders of more than a year running time, only a small proportion has been processed so far. This calls for automated assistance tools for detecting and/or segmenting infant utterances from real-live video recordings. In this paper, we investigated several approaches of infant voice detection and segmentation, including a rule-based voice activity detector, hidden Markov models with Gaussian mixture observation models, support vector machines, and random forests. Results indicate that the applied methods could be well applied in a semi-automated retrieval of infant utterances from highly non-standardised footage. At the same time, our results show that, a fully automated approach for this problem is yet to come.

Minimizing Annotation Effort for Adaptation of Speech-Activity Detection Systems

Luciana Ferrer[1], Martin Graciarena[2]; [1]Universidad de Buenos Aires, Argentina; [2]SRI International, USA

Sun-P-8-2-11, Time: 16:00

Annotating audio data for the presence and location of speech is a time-consuming and therefore costly task. This is mostly because annotation precision greatly affects the performance of the speech-activity detection (SAD) systems trained with this data, which means that the annotation process must be careful and detailed. Although significant amounts of data are already annotated for speech presence and are available to train SAD systems, these systems are known to perform poorly on channels that are not well-represented by the training data. However obtaining representative audio samples from a new channel is relative easy and this data can be used for training a new SAD system or adapting one trained with larger amounts of mismatched data. This paper focuses on the problem of selecting the best-possible subset of available audio data given a budgeted time for annotation. We propose simple approaches for selection that lead to significant gains over naïve methods that merely select N full files at random. An approach that uses the frame-level scores from a baseline system to select regions such that the score distribution is uniformly sampled gives the best trade-off across a variety of channel groups.

Sun-P-8-3 : New Products and Services

Pacific Concourse – Poster C, 16:00–18:00, Sunday, 11 Sept. 2016
Chair: Chuck Wooters

Progress and Prospects for Spoken Language Technology: What Ordinary People Think

Roger K. Moore, Hui Li, Shih-Hao Liao; University of Sheffield, UK

Sun-P-8-3-1, Time: 16:00

Arguably the most significant milestone (so far) in the spoken language technology field was the appearance in November 2011 of *Siri* — Apple's voice-based 'personal assistant and knowledge navigator' for the iPhone. *Siri* brought the potential of spoken language technology to the attention of the wider general public, and speech finally became "*mainstream*". This meant that ordinary people suddenly had an informed opinion about the merits (or otherwise) of using their voice to access information, send messages and control their smart devices. So, this paper presents the results of two surveys that were conducted in order to find out what ordinary people think about contemporary spoken language technology. The first used a modified version of the surveys conducted every six years at the IEEE ASRU series of workshops, and the second addressed questions about the awareness and usage of speech technology by members of the general public. The overall results suggest that ordinary people are more optimistic than the experts about what spoken language technology might have to offer, but usage patterns reveal that the majority of end users still prefer typing to talking, with accuracy, privacy and online accessibility cited as the main impediments to wider take-up.

Progress and Prospects for Spoken Language Technology: Results from Four Sexennial Surveys

Roger K. Moore, Ricard Marxer; University of Sheffield, UK

Sun-P-8-3-2, Time: 16:00

Since 1997, a survey has been conducted every six years at the IEEE workshop on *Automatic Speech Recognition and Understanding* (ASRU) in order to ascertain the research community's perspective on future progress and prospects in spoken language technology. These surveys have been based on a set of 'statements', each of which portray a possible future scenario, and respondents are asked to estimate the year in which each given scenario might become true. Many of the statements have appeared in several of the surveys, hence it is possible to track changes in opinion over time. This paper presents the combined results of all four surveys, the most recent of which was conducted at ASRU-2015. The results give an insight into the key trends that are taking place in the spoken language technology field, and reveal the realism that pervades the research community. They also suggest that there is growing confidence that some of the scenarios will indeed be realised at some point in the future.

On Employing a Highly Mismatched Crowd for Speech Transcription

Purushotam Radadia, Rahul Kumar, Kanika Kalra, Shirish Karande, Sachin Lodha; Tata Consultancy Services, India

Sun-P-8-3-3, Time: 16:00

Crowd sourcing provides a cheap and fast way to obtain speech transcriptions. The crowd size available for a task is inversely proportional to the skill requirements. Hence, there has been recent interest in studying the utility of mismatched crowd workers, who provide transcriptions even without knowing the source language. Nevertheless, these studies have required that the worker be capable of providing a transcription in Roman script. We believe that if the script constraint is removed, then countries like India can provide significantly larger crowd base. With this as a motivation, in this paper, we consider transcription of spoken Russian words by a rural Indian crowd that is unfamiliar with Russian and has very limited knowledge of English. The crowd we employ knew Gujarati, Marathi, Telugu and used the scripts of these languages to provide their transcriptions. We utilized an insertion-deletion-substitution channel to model the transcription errors. With a parallel channel model we can easily combine the crowd inputs. We show that the 4 transcriptions in Indic scripts (2 Gujarati, 1 Marathi, 1 Telugu) provide an accuracy of 73.77 (vs. 47% for ROVER algorithm) and a 4-best accuracy of 86.48%, even without employing any worker filtering.

Sage: The New BBN Speech Processing Platform

Roger Hsiao, Ralf Meermeier, Tim Ng, Zhongqiang Huang, Maxwell Jordan, Enoch Kan, Tanel Alumäe, Jan Silovsky, William Hartmann, Francis Keith, Omer Lang, Manhung Siu, Owen Kimball; Raytheon BBN Technologies, USA

Sun-P-8-3-4, Time: 16:00

To capitalize on the rapid development of Speech-to-Text (STT) technologies and the proliferation of open source machine learning toolkits, BBN has developed Sage, a new speech processing platform that integrates technologies from multiple sources, each of which has particular strengths. In this paper, we describe the design of Sage, which allows the easy interchange of STT components from different sources. We also describe our approach for fast prototyping with new machine learning toolkits, and a framework for sharing STT components across different applications. Finally, we report Sage's state-of-the-art performance on different STT tasks.

DNN-Based Feature Enhancement Using Joint Training Framework for Robust Multichannel Speech Recognition

Kang Hyun Lee, Tae Gyoon Kang, Woo Hyun Kang, Nam Soo Kim; Seoul National University, Korea

Sun-P-8-3-5, Time: 16:00

Ever since the deep neural network (DNN) appeared in the speech signal processing society, the recognition performance of automatic speech recognition (ASR) has been greatly improved. Due to this achievement, the demands on various applications in distant-talking environment also have been increased. However, ASR performance in such environments is still far from that in close-talking environments due to various problems. In this paper, we propose a novel multichannel-based feature mapping technique combining conventional beamformer, DNN and its joint training scheme. Through the experiments using multichannel wall street journal audio visual (MC-WSJ-AV) corpus, it has been shown that the proposed technique models the complicated relationship between the array inputs and clean speech features effectively via employing intermediate target. The proposed method outperformed the conventional DNN system.

Deep Neural Network Frontend for Continuous EMG-Based Speech Recognition

Michael Wand, Jürgen Schmidhuber; IDSIA, Switzerland

Sun-P-8-3-6, Time: 16:00

We report on a *Deep Neural Network* frontend for a continuous speech recognizer based on Surface Electromyography (EMG). Speech data is obtained by facial electrodes capturing the electric activity generated by the articulatory muscles, thus allowing speech processing without making use of the acoustic signal. The electromyographic signal is preprocessed and fed into the neural network, which is trained on framewise targets; the output layer activations are further processed by a Hidden Markov sequence classifier. We show that such a neural network frontend can be trained on EMG data and yields substantial improvements over previous systems, despite the fact that the available amount of data is very small, just amounting to a few tens of sentences: on the *EMG-UKA* corpus, we obtain average evaluation set Word Error Rate improvements of more than 32% relative on context-independent phone models and 13% relative on versatile *Bundled Phonetic feature* (BDPF) models, compared to a conventional system using Gaussian Mixture Models. In particular, on simple context-independent phone models, the new system yields results which are almost as good as with BDPF models, which were specifically designed to cope with small amounts of training data.

Overcoming Data Sparsity in Acoustic Modeling of Low-Resource Language by Borrowing Data and Model Parameters from High-Resource Languages

Basil Abraham, S. Umesh, Neethu Mariam Joy; IIT Madras, India

Sun-P-8-3-7, Time: 16:00

In this paper, we propose two techniques to improve the acoustic model of a low-resource language by: (i) Pooling data from closely related languages using a phoneme mapping algorithm to build acoustic models like subspace Gaussian mixture model (SGMM), phone cluster adaptive training (Phone-CAT), deep neural network (DNN) and convolutional neural network (CNN). Using the low-resource language data, we then adapt the afore mentioned models towards that language. (ii) Using models built from high-resource languages, we first borrow subspace model parameters from SGMM/Phone-CAT; or hidden layers from DNN/CNN. The language specific parameters are then estimated using the low-resource language data. The experiments were performed on four Indian languages namely Assamese, Bengali, Hindi and Tamil. Relative improvements of 10 to 30% were obtained over corresponding monolingual models in each case.

NOTES

Multi-Language Neural Network Language Models

Anton Ragni, Edgar Dakin, Xie Chen, Mark J.F. Gales, Kate M. Knill; University of Cambridge, UK

Sun-P-8-3-8, Time: 16:00

In recent years there has been considerable interest in neural network based language models. These models typically consist of vocabulary dependent input and output layers and one, or more, hidden layers. A standard problem with these networks is that large quantities of training data are needed to robustly estimate the model parameters. This poses a challenge when only limited data is available for the target language. One way to address this issue is to make use of overlapping vocabularies between related languages. However this is only applicable to a small set of languages, and the impact is expected to be limited for more general applications. This paper describes a general solution that allows data from any language to be used. Here, only the input and output layers are vocabulary dependent whilst hidden layers are shared, language independent. This multi-task training set-up allows the quantity of data available to train the hidden layers to be increased. This multi-language network can be used in a range of configurations, including as initialisation for previously unseen languages. As a proof of concept this paper examines multilingual recurrent neural network language models. Experiments are conducted using language packs released within the IARPA Babel program.

Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration

Ottokar Tilk[1], Tanel Alumäe[2]; [1]Tallinn University of Technology, Estonia; [2]Raytheon BBN Technologies, USA

Sun-P-8-3-9, Time: 16:00

Automatic speech recognition systems generally produce unpunctuated text which is difficult to read for humans and degrades the performance of many downstream machine processing tasks. This paper introduces a bidirectional recurrent neural network model with attention mechanism for punctuation restoration in unsegmented text. The model can utilize long contexts in both directions and direct attention where necessary enabling it to outperform previous state-of-the-art on English (IWSLT2011) and Estonian datasets by a large margin.

TheanoLM — An Extensible Toolkit for Neural Network Language Modeling

Seppo Enarvi, Mikko Kurimo; Aalto University, Finland

Sun-P-8-3-10, Time: 16:00

We present a new tool for training neural network language models (NNLMs), scoring sentences, and generating text. The tool has been written using Python library Theano, which allows researcher to easily extend it and tune any aspect of the training process. Regardless of the flexibility, Theano is able to generate extremely fast native code that can utilize a GPU or multiple CPU cores in order to parallelize the heavy numerical computations. The tool has been evaluated in difficult Finnish and English conversational speech recognition tasks, and significant improvement was obtained over our best back-off n-gram models. The results that we obtained in the Finnish task were compared to those from existing RNNLM and RWTHLM toolkits, and found to be as good or better, while training times were an order of magnitude shorter.

Selection of Multi-Genre Broadcast Data for the Training of Automatic Speech Recognition Systems

P. Lanchantin, Mark J.F. Gales, Penny Karanasou, X. Liu, Y. Qian, L. Wang, P.C. Woodland, C. Zhang; University of Cambridge, UK

Sun-P-8-3-11, Time: 16:00

This paper compares schemes for the selection of multi-genre broadcast data and corresponding transcriptions for speech recognition model training. Selections of the same amount of data (700 hours) from lightly supervised alignments based on the same original subtitle transcripts are compared. Data segments were selected according to a maximum phone matched error rate between the lightly supervised decoding and the original transcript. The data selected with an improved lightly supervised system yields lower word error rates (WERs). Detailed comparisons of the data selected on carefully transcribed development data show how the selected portions match the true phone error rate for each genre. From a broader perspective, it is shown that for different genres, either the original subtitles or the lightly supervised output should be used for model training and a suitable combination yields further reductions in final WER.

Manipulating Word Lattices to Incorporate Human Corrections

Yashesh Gaur, Florian Metze, Jeffrey P. Bigham; Carnegie Mellon University, USA

Sun-P-8-3-12, Time: 16:00

Automatic Speech Recognition (ASR) is not perfect and even advanced statistical models make errors that render its output difficult to understand. We are therefore interested in having Humans correct ASR output efficiently. A naive approach, in which Humans manually “edit” the ASR output, may work when the recognition is done offline, but fails in on-line scenarios when Humans cannot keep up. To address this problem, our prior work introduced an approach that combines ASR and keyword search (KWS) to allow Humans to simply type corrections for the errors they observe, while the system positioned each correction using KWS and then “stitches” in the correction. In this paper, we present an improved “stitching” algorithm that works at the lattice level (rather than on the first-best string). We show that this algorithm drastically improves the word error rate (WER) of a TED system when applied to a new corpus of CS lectures that has not been carefully prepared for ASR experiments. We also show that the system can fix annoying repeat errors from just a single correction, making it suitable for post-processing of large amounts of data from limited corrections.

Context-Aware Restaurant Recommendation for Natural Language Queries: A Formative User Study in the Automotive Domain

Philipp Fischer[1], Cornelius Styp von Rekowski[2], Andreas Nürnberger[2]; [1]Mercedes-Benz R&D North America, USA; [2]Otto-von-Guericke-Universität Magdeburg, Germany

Sun-P-8-3-13, Time: 16:00

In this paper, the authors describe an extension to an approach previously discussed for personalization of a natural language system in the automotive domain that allows reasoning under uncertainty with incomplete preference structures. Therefore, the concept of an

NOTES

"information stream" is defined as an underlying model for real-time recommendation learned from previous speech queries. The stream captures contextual data based on implicit feedback from the user's speech utterances.

Furthermore, a formative user study is discussed. Each study iteration has been based on a prototype that allows the user to utter natural language queries in the restaurant domain. The system responds with a ranked list of restaurant recommendations in relation to the user's context. Several driving scenarios with varying contexts have been analyzed (e.g. weekday/ weekend, route destinations, traffic). Users could inspect the result lists and indicate the most preferred item. In addition to quantitative data gained from this interaction, feedback on relevance of context features and on the UI concept was collected in a post-study interview for each iteration. Based on the study findings, we outline the contextual features found to be most relevant for speech-based interaction in automotive applications. These findings will be integrated into an existing hybrid recommendation model.

Teaming Up: Making the Most of Diverse Representations for a Novel Personalized Speech Retrieval Application

Stephanie Pancoast[1], Murat Akbacak[2]; [1]Airbnb, USA; [2]Apple, USA

Sun-P-8-3-14, Time: 16:00

In addition to the increasing number of publicly available multimedia documents generated and searched every day, there is also a large corpora of personalized videos, images and spoken recordings, stored on users' private devices and/or in their personal accounts in the cloud. Retrieving spoken items via voice commonly involves supervised indexing approaches such as large vocabulary speech recognition. When these items are personalized recordings, diverse and personalized content causes recognition systems to experience mis-matches mostly in vocabulary and language model components, and sometimes even in the language users use. All of these contribute to retrieval task performing very poorly. Alternatively, common audio patterns can be captured and used for exampler-based retrieval in an unsupervised fashion but this approach has its limitations as well. In this work we explore supervised, unsupervised and fusion techniques to perform the retrieval of short personalized spoken utterances. On a small collection of personal recordings, we find that when fusing word, phoneme and unsupervised frame based systems, we can improve accuracy on the top retrieved item approximately 3% above the best performing individual system. Besides demonstrating this improvement on our initial collection, we hope to attract community's interest to such novel personalized retrieval applications.

Automatic Speech Transcription for Low-Resource Languages — The Case of Yoloxóchitl Mixtec (Mexico)

Vikramjit Mitra[1], Andreas Kathol[1], Jonathan D. Amith[2], Rey Castillo García[3]; [1]SRI International, USA; [2]Gettysburg College, USA; [3]Secretaría de Educación Pública, Mexico

Sun-P-8-3-15, Time: 16:00

The rate at which endangered languages can be documented has been highly constrained by human factors. Although digital recording of natural speech in endangered languages may proceed at a fairly robust pace, transcription of this material is not only time consuming but severely limited by the lack of native-speaker personnel proficient in the orthography of their mother tongue. Our NSF-funded project in the Documenting Endangered Languages (DEL) program proposes to tackle this problem from two sides: first via a tool that helps native speakers become proficient in the orthographic conventions of their language, and second by using automatic speech recognition (ASR) output that assists in the transcription effort for newly recorded audio data. In the present study, we focus exclusively on progress in developing speech recognition for the language of interest, Yoloxóchitl Mixtec (YM), an Oto-Manguean language spoken by fewer than 5000 speakers on the Pacific coast of Guerrero, Mexico. In particular, we present results from an initial set of experiments and discuss future directions through which better and more robust acoustic models for endangered languages with limited resources can be created.

Real-Time Presentation Tracking Using Semantic Keyword Spotting

Reza Asadi, Harriet J. Fell, Timothy Bickmore, Ha Trinh; Northeastern University, USA

Sun-P-8-3-16, Time: 16:00

Given presentation slides with detailed written speaking notes, automatic tracking of oral presentations can help speakers ensure they cover their planned content, and can reduce their anxiety during the speech. Tracking is a more complex problem than speech-to-text alignment, since presenters rarely follow their exact presentation notes, and it must be performed in real-time. In this paper, we propose a novel system that can track the current degree of coverage of each slide's contents. To do this, the presentation notes for each slide are segmented into sentences, and the words are filtered into keyword candidates. These candidates are then scored based on word specificity and semantic similarity measures to find the most useful keywords for the tracking task. Real-time automatic speech recognition results are matched against the keywords and their synonyms. Sentences are scored based on detected keywords, and the ones with scores higher than a threshold are tagged as covered. We manually and automatically annotated 150 slide presentation recordings to evaluate the system. A simple tracking method, matching speech recognition results against the notes, was used as the baseline. The results show that our approach led to higher accuracy measures compared to the baseline method.

Sun-P-8-4 : Low Resource Speech Recognition

Pacific Concourse – Poster D, 16:00–18:00, Sunday, 11 Sept. 2016
Chair: Kartik Audhkhasi

Deriving Phonetic Transcriptions and Discovering Word Segmentations for Speech-to-Speech Translation in Low-Resource Settings

Andrew Wilkinson, Tiancheng Zhao, Alan W. Black; Carnegie Mellon University, USA

Sun-P-8-4-1, Time: 16:00

We investigate speech-to-speech translation where one language does not have a well-defined written form. We use English-Spanish and Mandarin-English bitext corpora in order to provide both gold-standard text-based translations and experimental results for

Notes

different levels of automatically derived symbolic representations from speech. We constrain our experiments such that the methods developed can be extended to low-resource languages. We derive different phonetic representations of the source texts in order to model the kinds of transcriptions that can be learned from low-resource-language speech data. We experiment with different methods of clustering the elements of the phonetic representations together into word-like units. We train MT models on the resulting texts, and report BLEU scores for the different representations and clustering methods in order to compare their effectiveness. Finally, we discuss our findings and suggest avenues for future research.

Unsupervised Joint Estimation of Grapheme-to-Phoneme Conversion Systems and Acoustic Model Adaptation for Non-Native Speech Recognition

Satoshi Tsujioka, Sakriani Sakti, Koichiro Yoshino, Graham Neubig, Satoshi Nakamura; NAIST, Japan

Sun-P-8-4-2, Time: 16:00

Non-native speech differs significantly from native speech, often resulting in a degradation of the performance of automatic speech recognition (ASR). Hand-crafted pronunciation lexicons used in standard ASR systems generally fail to cover non-native pronunciations, and design of new ones by linguistic experts is time consuming and costly. In this work, we propose acoustic data-driven iterative pronunciation learning for non-native speech recognition, which automatically learns non-native pronunciations directly from speech using an iterative estimation procedure. Grapheme-to-Phoneme (G2P) conversion is used to predict multiple candidate pronunciations for each word, occurrence frequency of pronunciation variations is estimated from the acoustic data of non-native speakers, and these automatically estimated pronunciation variations are used to perform acoustic model adaptation. We investigate various cases such as learning (1) without knowledge of non-native pronunciation, and (2) when we adapt to the speaker's proficiency level. In experiments on speech from non-native speakers of various levels, the proposed method was able to achieve an 8.9% average improvement in accuracy.

Learning Personalized Pronunciations for Contact Name Recognition

Antoine Bruguier, Fuchun Peng, Françoise Beaufays; Google, USA

Sun-P-8-4-3, Time: 16:00

Automatic speech recognition that involves people's names is difficult because names follow a long-tail distribution and they have no commonly accepted spelling or pronunciation. This poses significant challenges to contact dialing by voice. We propose using personalized pronunciation learning: people can use their own pronunciations for their contact names. We achieve this by implicitly learning from users' corrections and within minutes making that pronunciation available for the next voice dialing. We show that personalized pronunciations significantly reduce word error for difficult contact names by 15% relatively.

Generation and Pruning of Pronunciation Variants to Improve ASR Accuracy

Zhenhao Ge, Aravind Ganapathiraju, Ananth N. Iyer, Scott A. Randal, Felix I. Wyss; Interactive Intelligence, USA

Sun-P-8-4-4, Time: 16:00

Speech recognition, especially name recognition, is widely used in phone services such as company directory dialers, stock quote providers or location finders. It is usually challenging due to pronunciation variations. This paper proposes an efficient and robust data-driven technique which automatically learns acceptable word pronunciations and updates the pronunciation dictionary to build a better lexicon without affecting recognition of other words similar to the target word. It generalizes well on datasets with various sizes, and reduces the error rate on a database with 13000+ human names by 42%, compared to a baseline with regular dictionaries already covering canonical pronunciations of 97%+ words in names, plus a well-trained spelling-to-pronunciation (STP) engine.

Optimizing Speech Recognition Evaluation Using Stratified Sampling

Janne Pylkkönen[1], Thomas Drugman[2], Max Bisani[2]; [1]Amazon.com, Finland; [2]Amazon.com, Germany

Sun-P-8-4-5, Time: 16:00

Producing large enough quantities of high-quality transcriptions for accurate and reliable evaluation of an automatic speech recognition (ASR) system can be costly. It is therefore desirable to minimize the manual transcription work for producing metrics with an agreed precision. In this paper we demonstrate how to improve ASR evaluation precision using stratified sampling. We show that by altering the sampling, the deviations observed in the error metrics can be reduced by up to 30% compared to random sampling, or alternatively, the same precision can be obtained on about 30% smaller datasets. We compare different variants for conducting stratified sampling, including a novel sample allocation scheme tailored for word error rate. Experimental evidence is provided to assess the effect of different sampling schemes to evaluation precision.

NOTES

Keynote 4: Dan Jurafsky

Grand Ballroom ABC, 08:30–09:30, Monday, 12 Sept. 2016
Chair: Panos Georgiou

Ketchup, Interdisciplinarity, and the Spread of Innovation in Speech and Language Processing

Dan Jurafsky; Stanford University, USA

Mon-Keynote-4, Time: 08:30

I show how natural language processing can help model the spread of innovation through scientific communities, with special focus on the history of speech and language processing, and the important role of interdisciplinarity.

Mon-SE-3 : Special Event: Speech Ventures

Grand Ballroom A, 10:00–12:00, Monday, 12 Sept. 2016
Chairs: Nicolas Scheffer, Korbinian Riedhammer, Alex Lebrun, David Suendermann-Oeft

Speech Ventures

Nicolas Scheffer[1], Korbinian Riedhammer[2], Alexandre Lebrun[1], David Suendermann-Oeft[3]; [1]Facebook, USA; [2]Remeeting, USA; [3]Educational Testing Service, USA

Mon-SE-3, Time: 10:00

Interspeech 2016, the world's largest conference on speech technologies to be held in San Francisco, the heart of Silicon Valley, provides a unique opportunity to present the most recent developments and ideas of both academia and industry. Located at the cross-section of the two, startups that are interested in using speech in their products or that want to share their experience in doing so, are invited to participate in the speech venture special event. This event provides a platform for participants to interact with the brightest speech researchers and present and discover new trends in spoken language technology.

The objective of this special event are two-fold: • Leverage the experience and stories of startups as they adopt speech in their products; • Enable startups to attend a day of the largest conference in speech and meet with researchers, companies, and research institutions.

Mon-O-9-2 : Special Session: Speech and Language Technologies for Human-Machine Conversation-Based Language Education

Grand Ballroom BC, 10:00–12:00, Monday, 12 Sept. 2016
Chairs: Yao Qian, Helen Meng, Frank Soong

Context Aware Mispronunciation Detection for Mandarin Pronunciation Training

*Rong Tong, Nancy F. Chen, Bin Ma, Haizhou Li; A*STAR, Singapore*

Mon-O-9-2-1, Time: 10:00

Mispronunciation detection is an important component in a computer-assisted language learning (CALL) system. Many CALL systems only provide pronunciation correctness as the single feedback, which is not very informative for language learners. This paper proposes a context aware multilayer framework for Mandarin mispronunciation detection. The proposed framework incorporates the context information in the detection process and providing phonetic, tonal and syllabic level feedback. In particular, the contribution of this work is twofold: 1) we propose to use a multilayer mispronunciation detection architecture to detect and provide mispronunciation feedback at the phonetic, tonal and syllabic levels. 2) we propose to incorporate the phonetic and tone context information in mispronunciation detection using vector space modelling. Our experiment results show that the proposed framework improves the mispronunciation detection performance in all three levels.

DNN Online with iVectors Acoustic Modeling and Doc2Vec Distributed Representations for Improving Automated Speech Scoring

Jidong Tao, Lei Chen, Chong Min Lee; Educational Testing Service, USA

Mon-O-9-2-2, Time: 10:15

When applying automated speech-scoring technology to the rating of globally administered real assessments, there are several practical challenges: (a) ASR accuracy on non-native spontaneous speech is generally low; (b) due to the data mismatch between an ASR systems training stage and its final usage, the recognition accuracy obtained in practice is even lower; (c) content-relevance was not widely used in the scoring models in operation due to various technical and logistical issues. For this paper, an ASR in a deep neural network (DNN) architecture of multi-splice with iVectors was trained and resulted in a performance at 19.1% word error rate (WER). Secondly, we applied language model (LM) adaptation for the prompts that were not covered in ASR training by using the spoken responses acquired from previous operational tests, and we were able to reduce the relative WER by more than 8%. The boosted ASR performance improves the scoring performance without any extra human annotation cost. Finally, the developed ASR system allowed us to apply content features in practice. Besides the conventional frequency-based approach, content vector analysis (CVA), we also explored distributed representations with Doc2Vec and found an improvement on content measurement.

Self-Adaptive DNN for Improving Spoken Language Proficiency Assessment

Yao Qian, Xinhao Wang, Keelan Evanini, David Suendermann-Oeft; Educational Testing Service, USA

Mon-O-9-2-3, Time: 10:30

Automated assessment of language proficiency of a test taker's spoken response regarding its content, vocabulary, grammar and context depends largely upon how well the input speech can be recognized. While state-of-the-art, deep neural net based acoustic models have significantly improved the recognition performance of native speaker's speech, good recognition is still challenging when the input speech consists of non-native spontaneous utterances. In this paper, we investigate how to train a DNN based ASR with a fairly large non-native English corpus and make it self-adaptive to a test speaker and a new task, namely a simulated conversation, which is different from them monologic speech in the training data. Automated assessment of language proficiency is evaluated according to

NOTES

both task completion (TC) and pragmatic competence (PC) rubrics. Experimental results show that self-adaptive DNNs trained with i-vectors can reduce absolute word error rate by 11.7% and deliver more accurate recognized word sequences for language proficiency assessment. Also, the recognition accuracy gain translates into a gain of automatic assessment performance on the test data. The correlations between automated scoring and expert scoring could be increased by 0.07 (TC) and 0.15 (PC), respectively.

Detecting Mispronunciations of L2 Learners and Providing Corrective Feedback Using Knowledge-Guided and Data-Driven Decision Trees

*Wei Li[1], Kehuang Li[1], Sabato Marco Siniscalchi[1], Nancy F. Chen[2], Chin-Hui Lee[1]; [1]Georgia Institute of Technology, USA; [2]A*STAR, Singapore*

Mon-O-9-2-4, Time: 10:45

We propose a novel decision tree based framework to detect phonetic mispronunciations produced by L2 learners caused by using inaccurate speech attributes, such as manner and place of articulation. Compared with conventional score-based CAPT (computer assisted pronunciation training) systems, our proposed framework has three advantages: (1) each mispronunciation in a tree can be interpreted and communicated to the L2 learners by traversing the corresponding path from a leaf node to the root node; (2) corrective feedback based on speech attribute features, which are directly used to describe how consonants and vowels are produced using related articulators, can be provided to the L2 learners; and (3) by building the phone-dependent decision tree, the relative importance of the speech attribute features of a target phone can be automatically learned and used to distinguish itself from other phones. This information can provide L2 learners speech attribute feedback that is ranked in order of importance. In addition to the abovementioned advantages, experimental results confirm that the proposed approach can detect most pronunciation errors and provide accurate diagnostic feedback.

Phoneme Set Design Considering Integrated Acoustic and Linguistic Features of Second Language Speech

Xiaoyun Wang, Tsuneo Kato, Seiichi Yamamoto; Doshisha University, Japan

Mon-O-9-2-5, Time: 11:00

Recognition of second language speech is still a challenging task even for state-of-the-art automatic speech recognition (ASR) systems. Considering that second language speech usually includes less fluent pronunciation and mispronunciation even when it is grammatically correct, we propose a novel phonetic decision tree (PDT) method considering integrated acoustic and linguistic features to derive the phoneme set for second language speech recognition. We verify the efficacy of the proposed method using second language speech collected with a translation game type dialogue-based English CALL system. Experimental results demonstrated that the derived phoneme set achieved higher accuracy recognition performance than the canonical one.

HMM-Based Non-Native Accent Assessment Using Posterior Features

Ramya Rasipuram, Milos Cernak, Mathew Magimai-Doss; Idiap Research Institute, Switzerland

Mon-O-9-2-6, Time: 11:15

Automatic non-native accent assessment has potential benefits in language learning and speech technologies. The three fundamental challenges in automatic accent assessment are to characterize, model and assess individual variation in speech of the non-native speaker. In our recent work, accentedness score was automatically obtained by comparing two phone probability sequences obtained through instances of non-native and native speech. Although automatic accentedness ratings of the approach correlated well with human accent ratings, the approach is critically constrained because of the requirement of native speech instance. In this paper, we build on the previous work and obtain the native latent symbol probability sequence through the word hypothesis modeled as a hidden Markov model (HMM). The latent symbols are either context-independent phonemes or clustered context-dependent phonemes. The advantage of the proposed approach is that it requires just reference text transcription instead of native speech recordings. Using the HMMs trained on an auxiliary native speech corpus, the proposed approach achieves a correlation of 0.68 with human accent ratings on the ISLE corpus. This is further interesting considering that the approach does not use any non-native data and human accent ratings at any stage of the system development.

Automatic Assessment and Error Detection of Shadowing Speech: Case of English Spoken by Japanese Learners

Shuju Shi[1], Yosuke Kashiwagi[1], Shohei Toyama[1], Junwei Yue[1], Yutaka Yamauchi[2], Daisuke Saito[1], Nobuaki Minematsu[1]; [1]University of Tokyo, Japan; [2]Tokyo International University, Japan

Mon-O-9-2-7, Time: 11:30

Shadowing is a task where the subject is required to repeat the presented speech as s/he hears it. Although shadowing is cognitively a challenging task, it is considered as an efficient way of language training since it includes processes of listening, speaking and comprehension simultaneously. Our previous study realized automatic assessment of shadowing speech using the average of Goodness of Pronunciation (GOP) scores. But the fact that shadowing often includes broken utterances makes this approach insufficient. This study attempts to improve automatic assessment and, at the same time, give corrective feedbacks to learners based on error detection. We first manually labeled shadowing speech of 10 female and 10 male speakers and defined ten typical error types including word omission, substitution etc.. Forced alignment with adjusted grammar and GOP scores are adopted to detect word omission errors and poorly pronounced words. In the experiments, GOP scores, Word Recognition Rate (WRR), silence ratio, forced alignment log-likelihood scores, word omission rate are used to predict the overall proficiency of the individual speakers. The mean correlation coefficient between automatic scores and the speaker's TOEIC scores is 0.81, improved by 13% relatively. The detection accuracy of word omission is 73%.

Mon-O-9-3 : Phonation and Voice Quality

Bayview A, 10:00–12:00, Monday, 12 Sept. 2016
Chairs: Paavo Alku, Marc Garellek

Multiplicity of the Acoustic Correlates of the Fortis-Lenis Contrast: Plosives in Aberystwyth English

Míša Hejná; Newcastle University, UK

Mon-O-9-3-1, Time: 10:00

Using evidence from Aberystwyth English, this study shows two points relevant for the phonetic implementation of the fortis-lenis contrast in plosives and two points concerning the diachronic scenarios proposed as ways in which pre-aspiration (one of the correlates of this contrast) innovates. Firstly, a wide range of acoustic features distinguishes the fortis and the lenis series. Release duration is a correlate of the contrast in three foot-positions (initial: **t**ot vs **d**ot; medial: co**tt**er vs co**dd**er; final: co**t** vs co**d**), as is vowel duration and the presence of voicing. Furthermore, pre-aspiration and breathiness differentiate the two series foot-medially and foot-finally. For one speaker, glottalisation rather than pre-aspiration distinguishes the series foot-finally. Secondly, whilst the plosives are frequently post-aspirated foot-initially, the release of /t/ and /d/ is realised variably with affrication and/or post-aspiration in all three positions: rather than presence or absence of affrication or post-aspiration then, it is release duration that distinguishes the series. Thirdly, the data is not supportive of the suggestion that pre-aspiration innovates in the fortis series as a consequence of the loss of voicing in the lenis series or the other way round [1] and [2] or, fourthly, as a step on a degemination trajectory [3], [4].

Automatic Measurement of Voice Onset Time and Prevoicing Using Recurrent Neural Networks

Yossi Adi[1], Joseph Keshet[1], Olga Dmitrieva[2], Matt Goldrick[3]; [1]Bar-Ilan University, Israel; [2]Purdue University, USA; [3]Northwestern University, USA

Mon-O-9-3-2, Time: 10:20

Voice onset time (VOT) is defined as the time difference between the onset of the burst and the onset of voicing. When voicing begins preceding the burst, the stop is called prevoiced, and the VOT is negative. When voicing begins following the burst the VOT is positive. While most of the work on automatic measurement of VOT has focused on positive VOT mostly evident in American English, in many languages the VOT can be negative. We propose an algorithm that estimates if the stop is prevoiced, and measures either positive or negative VOT, respectively. More specifically, the input to the algorithm is a speech segment of an arbitrary length containing a single stop consonant, and the output is the time of the burst onset, the duration of the burst, and the time of the prevoicing onset with a confidence. Manually labeled data is used to train a recurrent neural network that can model the dynamic temporal behavior of the input signal, and outputs the events' onset and duration. Results suggest that the proposed algorithm is superior to the current state-of-the-art both in terms of the VOT measurement and in terms of prevoicing detection.

L1-L2 Interference: The Case of Final Devoicing of French Voiced Fricatives in Final Position by German Learners

Sucheta Ghosh[1], Camille Fauth[2], Aghilas Sini[1], Yves Laprie[1]; [1]LORIA, France; [2]LiLPa, France

Mon-O-9-3-3, Time: 10:40

This work is dealing with a case of L1-L2 interference in language learning. The Germans learning French as a second language frequently produce unvoiced fricatives in word-final position instead of the expected voiced fricatives. We investigated the production of French fricatives for 16 non-native (8 beginner- and 8 advanced-learners) and 8 native speakers, and designed auditory feedback to help them realize the right voicing feature. The productions of all speakers were categorized either as voiced or unvoiced by experts. The same fricatives were also evaluated by non-experts in a perception experiment targeting VCs. We compare the ratings by experts and non-experts with the feature-based analysis. The ratio of locally unvoiced frames in the consonantal segment and also the ratio between consonantal duration and V1 duration were measured. The acoustic cues of neighboring sounds and pitch-based features play a significant role in the voicing judgment. As expected, we found that beginners face more difficulties to produce voiced fricatives than advanced learners. Also, the production becomes easier for the learners, especially for the beginners, if they practice repetition after a native speaker. We use these findings to design and develop feedback via speech analysis/synthesis technique TD-PSOLA using the learner's own voice.

Perceptual Salience of Voice Source Parameters in Signaling Focal Prominence

Irena Yanushevskaya, Andy Murphy, Christer Gobl, Ailbhe Ní Chasaide; Trinity College Dublin, Ireland

Mon-O-9-3-4, Time: 11:00

This paper describes listening tests investigating the perceptual role of voice source parameters (other than F0) in signaling focal prominence. Synthesized stimuli were constructed on the basis of an inverse filtered utterance 'We were away a year ago'. Voice source parameters were manipulated in the two potentially accentable syllables WAY and YEAR (in terms of the absolute magnitude and alignment of peaks) and to provide source deaccentuation of post-focal material. Participants in the first listening test were asked to decide whether the syllable WAY, YEAR or neither was deemed the most prominent: judgments on the degree of prominence and naturalness were also indicated on a continuous visual analogue scale. In the second test listeners indicated the degree of prominence for every syllable in the phrase. For WAY, voice source manipulations can cue focal accentuation, and both the magnitude of the source manipulation of the syllable and the presence of source deaccentuation contribute to the effect. However, for YEAR, listeners' perception of focal accentuation tended to show relatively minor increases in perceived prominence regardless of the source manipulations involved. It therefore appears that the source expression of focus is sensitive to the location of focus in the intonational phrase.

NOTES

Classification of Voice Modality Using Electroglottogram Waveforms

Michal Borsky[1], Daryush D. Mehta[2], Julius P. Gudjohnsen[1], Jon Gudnason[1]; [1]Reykjavik University, Iceland; [2]Massachusetts General Hospital, USA

Mon-O-9-3-5, Time: 11:20

It has been proven that the improper function of the vocal folds can result in perceptually distorted speech that is typically identified with various speech pathologies or even some neurological diseases. As a consequence, researchers have focused on finding quantitative voice characteristics to objectively assess and automatically detect non-modal voice types. The bulk of the research has focused on classifying the speech modality by using the features extracted from the speech signal. This paper proposes a different approach that focuses on analyzing the signal characteristics of the electroglottogram (EGG) waveform. The core idea is that modal and different kinds of non-modal voice types produce EGG signals that have distinct spectral/cepstral characteristics. As a consequence, they can be distinguished from each other by using standard cepstral-based features and a simple multivariate Gaussian mixture model. The practical usability of this approach has been verified in the task of classifying among modal, breathy, rough, pressed and soft voice types. We have achieved 83% frame-level accuracy and 91% utterance-level accuracy by training a speaker-dependent system.

Voice-Quality Difference Between the Vowels in Filled Pauses and Ordinary Lexical Items

Kikuo Maekawa[1], Hiroki Mori[2]; [1]NINJAL, Japan; [2]Utsunomiya University, Japan

Mon-O-9-3-6, Time: 11:40

Acoustic differences between the vowels in filled pauses and ordinary lexical items such as nouns and verbs were examined to know if there was systematic difference of voice-quality. Statistical test of material taken from the Corpus of Spontaneous Japanese showed that, in most cases, there was significant difference of acoustic features like F0, F1, F2, intensity, jitter, shimmer, TL, H1-H2, H1-A2, duration, etc. between the two classes of vowels. Random forest classification of open data sets showed higher than 0.8 F-values on average. It turned out intensity, F0, F1, jitter, and H1-H2 were the most important acoustic features for the expected voice-quality difference.

Mon-O-9-4 : Speech Synthesis Oral: Prosody and Expressive Speech

Bayview B, 10:00–12:00, Monday, 12 Sept. 2016
Chairs: Junichi Yamagishi, Maria Wolters

Generation of Emotion Control Vector Using MDS-Based Space Transformation for Expressive Speech Synthesis

Yan-You Chen, Chung-Hsien Wu, Yu-Fong Huang; National Cheng Kung University, Taiwan

Mon-O-9-4-1, Time: 10:00

In control vector-based expressive speech synthesis, the emotion/style control vector defined in the categorical (CAT) emotion space is uneasy to be precisely defined by the user to synthesize the speech with the desired emotion/style. This paper applies the arousal-valence (AV) space to the multiple regression hidden semi-Markov model (MRHSMM)-based synthesis framework for expressive speech synthesis. In this study, the user can designate a specific emotion by defining the AV values in the AV space. The multidimensional scaling (MDS) method is adopted to project the AV emotion space and the categorical (CAT) emotion space onto their corresponding orthogonal coordinate systems. A transformation approach is thus proposed to transform the AV values to the emotion control vector in CAT emotion space for MRHSMM-based expressive speech synthesis. In the synthesis phase given the input text and desired emotion, with the transformed emotion control vector, the speech with the desired emotion is generated from the MRHSMMs. Experimental result shows the proposed method is helpful for the user to easily and precisely determine the desired emotion for expressive speech synthesis.

Direct Expressive Voice Training Based on Semantic Selection

Igor Jauk, Antonio Bonafonte; Universitat Politècnica de Catalunya, Spain

Mon-O-9-4-2, Time: 10:20

This work aims at creating expressive voices from audiobooks using semantic selection. First, for each utterance of the audiobook an acoustic feature vector is extracted, including iVectors built on MFCC and on F0 basis. Then, the transcription is projected into a semantic vector space. A seed utterance is projected to the semantic vector space and the N nearest neighbors are selected. The selection is then filtered by selecting only acoustically similar data.

The proposed technique can be used to train emotional voices by using emotional keywords or phrases as seeds, obtaining training data semantically similar to the seed. It can also be used to read larger texts in an expressive manner, creating specific voices for each sentence. That later application is compared to a DNN predictor, which predicts acoustic features from semantic features. The selected data is used to adapt statistical speech synthesis models. The performance of the technique is analyzed objectively and in a perceptive experiment. In the first part of the experiment, subjects clearly show preference for particular expressive voices to synthesize semantically expressive utterances. In the second part, the proposed method is shown to achieve similar or better performance than the DNN based prediction.

Syllable-Level Representations of Suprasegmental Features for DNN-Based Text-to-Speech Synthesis

Manuel Sam Ribeiro, Oliver Watts, Junichi Yamagishi; University of Edinburgh, UK

Mon-O-9-4-3, Time: 10:40

A top-down hierarchical system based on deep neural networks is investigated for the modeling of prosody in speech synthesis. Suprasegmental features are processed separately from segmental features and a compact distributed representation of high-level units is learned at syllable-level. The suprasegmental representation is then integrated into a frame-level network. Objective measures show that balancing segmental and suprasegmental features can be useful for the frame-level network. Additional features incorporated into the hierarchical system are then tested. At the syllable-level, a bag-of-phones representation is proposed and, at the word-level, embeddings learned from text sources are used. It is shown that the

NOTES

hierarchical system is able to leverage new features at higher-levels more efficiently than a system which exploits them directly at the frame-level. A perceptual evaluation of the proposed systems is conducted and followed by a discussion of the results.

Pause Prediction from Text for Speech Synthesis with User-Definable Pause Insertion Likelihood Threshold

Norbert Braunschweiler, Ranniery Maia; Toshiba Research Europe, UK

Mon-O-9-4-4, Time: 11:00

Predicting the location of pauses from text is an important aspect for speech synthesizers. The accuracy of pause prediction can significantly influence both naturalness and intelligibility. Pauses which help listeners to better parse the synthesized speech into meaningful units are deemed to increase naturalness and intelligibility ratings, while pauses in unexpected or incorrect locations can reduce these ratings and cause confusion. This paper presents a multi-stage pause prediction approach including first prosodic chunk prediction, followed by a feature scoring algorithm and finally a pause sequence evaluation module. Preference tests showed that the new method outperformed a pauses-at-punctuation baseline while not yet matching human performance. In addition, the approach includes two more functionalities: (1) a user-specifiable pause insertion rate and (2) multiple output formats in the form of binary pauses, multi-level pauses or as a score reflecting pause strength.

A Hybrid System for Continuous Word-Level Emphasis Modeling Based on HMM State Clustering and Adaptive Training

Quoc Truong Do[1], Tomoki Toda[2], Graham Neubig[1], Sakriani Sakti[1], Satoshi Nakamura[1]; [1]NAIST, Japan; [2]Nagoya University, Japan

Mon-O-9-4-5, Time: 11:20

Emphasis is an important aspect of speech that conveys the focus of utterances, and modeling of this emphasis has been an active research field. Previous work has modeled emphasis using state clustering with an emphasis contextual factor indicating whether or not a word is emphasized. In addition, cluster adaptive training (CAT) makes it possible to directly optimize model parameters for clusters with different characteristics. In this paper, we first make a straightforward extension of CAT to emphasis adaptive training using continuous emphasis representations. We then compare it to state clustering, and propose a hybrid approach that combines both the emphasis contextual factor and adaptive training. Experiments demonstrated the effectiveness of adaptive training both stand-alone or combined with the state clustering approach (hybrid system) with it improving emphasis estimation by 2–5% F-measure and producing more natural audio.

Improving Prosodic Boundaries Prediction for Mandarin Speech Synthesis by Using Enhanced Embedding Feature and Model Fusion Approach

Yibin Zheng, Ya Li, Zhengqi Wen, Xingguang Ding, Jianhua Tao; Chinese Academy of Sciences, China

Mon-O-9-4-6, Time: 11:40

Hierarchical prosody structure generation is an important but challenging component for speech synthesis systems. In this paper, we investigate the use of enhanced embedding (joint learning of character and word embedding (CWE)) features and different model fusion approaches at both character and word level for Mandarin prosodic boundaries prediction. For CWE module, the internal structures of words and non-compositional words are considered in the word embedding, while the character ambiguity is addressed by multiple-prototype character embedding. For model fusion module, linear function (LF) and gradient boosting decision tree (GBDT), are investigated at the decision level respectively, with the important features selected by feature ranking module used as its input. Experiment results show the effectiveness of the proposed enhanced embedding features and the two model fusion approaches at both character and word level.

Mon-O-9-5 : Language Recognition

Seacliff BCD, 10:00–12:00, Monday, 12 Sept. 2016
Chairs: Jean-Francois Bonastre, Haizhou Li

Results of The 2015 NIST Language Recognition Evaluation

Hui Zhao[1], Désiré Bansé[1], George Doddington[1], Craig Greenberg[1], Jaime Hernández-Cordero[2], John Howard[1], Lisa Mason[2], Alvin Martin[1], Douglas Reynolds[3], Elliot Singer[3], Audrey Tong[1]; [1]NIST, USA; [2]DoD, USA; [3]MIT Lincoln Laboratory, USA

Mon-O-9-5-1, Time: 10:00

In 2015, NIST conducted the most recent in an ongoing series of Language Recognition Evaluations (LRE) meant to foster research in language recognition. The 2015 Language Recognition Evaluation featured 20 target languages grouped into 6 language clusters. The evaluation was focused on distinguishing languages within each cluster, without disclosing which cluster a test language belongs to.

The 2015 evaluation introduced several new aspects, such as using limited and specified training data and a wider range of durations for test segments. Unlike in past LRE's, systems were not required to output hard decisions for each test language and test segment, instead systems were required to provide a vector of log likelihood ratios to indicate the likelihood a test segment matches a target language. A total of 24 research organizations participated in this four-month long evaluation and combined they submitted 167 systems to be evaluated. The evaluation results showed that top-performing systems exhibited similar performance and there were wide variations in performance based on language clusters and within cluster language pairs. Among the 6 clusters, the French cluster was the hardest to recognize, with near random performance, and the Slavic cluster was the easiest to recognize.

The 2015 NIST Language Recognition Evaluation: The Shared View of I2R, Fantastic4 and SingaMS

*Kong Aik Lee [1], Haizhou Li [1], Li Deng [2], Ville Hautamäki [3], Wei Rao [4], Xiong Xiao [4], Anthony Larcher [5], Hanwu Sun [1], Trung Hieu Nguyen [1], Guangsen Wang [1], Aleksandr Sizov [1], Jianshu Chen [2], Ivan Kukanov [3], Amir Hossein Poorjam [3], Trung Ngo Trong [3], Cheng-Lin Xu [4], Haihua Xu [4], Bin Ma [1], Eng Siong Chng [4], Sylvain Meignier [5]; [1] A*STAR, Singapore; [2] Microsoft, USA; [3] University of Eastern Finland, Finland; [4] NTU, Singapore; [5] LIUM, France*

Mon-O-9-5-2, Time: 10:20

The series of *language recognition evaluations* (LRE's) conducted by the National Institute of Standards and Technology (NIST) have been one of the driving forces in advancing spoken language recognition technology. This paper presents a shared view of five institutions resulting from our collaboration toward LRE 2015 submissions under the names of I2R, Fantastic4, and SingaMS. Among others, LRE'15 emphasizes on language detection in the context of closely related languages, which is different from previous LRE's. From the perspective of language recognition system design, we have witnessed a major paradigm shift in adopting deep neural network (DNN) for both feature extraction and classifier. In particular, deep bottleneck features (DBF) have a significant advantage in replacing the shifted-delta-cepstral (SDC) which has been the only option in the past. We foresee deep learning is going to serve as a major driving force in advancing spoken language recognition system in the coming years.

Pair-Wise Distance Metric Learning of Neural Network Model for Spoken Language Identification

Xugang Lu [1], Peng Shen [1], Yu Tsao [2], Hisashi Kawai [1]; [1] NICT, Japan; [2] Academia Sinica, Taiwan

Mon-O-9-5-3, Time: 10:40

The i-vector representation and modeling technique has been successfully applied in spoken language identification (SLI). In modeling, a discriminative transform or classifier must be applied to emphasize variations correlated to language identity since the i-vector representation encodes most of the acoustic variations (e.g., speaker variation, transmission channel variation, etc.). Due to the strong nonlinear discriminative power of neural network (NN) modeling (including its deep form DNN), the NN has been directly used to learn the mapping function between the i-vector representation and language identity labels. In most studies, only the point-wise feature-label information is feeded to NN for parameter learning which may result in model overfitting, particularly when with limited training data. In this study, we propose to integrate pair-wise distance metric learning in NN parameter optimization. In the representation space of nonlinear transforms of hidden layers, a distance metric learning is explicitly designed for minimizing the pair-wise intra-class variation and maximizing the inter-class variation. With the distance metric as a constraint in the point-wise learning, the i-vectors are transformed to a new feature space which are much more discriminative for samples belonging to different languages while are much more similar for samples belonging to the same language. We tested the algorithm on a SLI task, encouraging results were obtained with more than 20% relative improvement on identification error rate.

Non-Iterative Parameter Estimation for Total Variability Model Using Randomized Singular Value Decomposition

Ruchir Travadi, Shrikanth S. Narayanan; University of Southern California, USA

Mon-O-9-5-4, Time: 11:00

In this paper, we address the problem of parameter estimation for the Total Variability Model (TVM) [1]. Typically, the estimation of the Total Variability Matrix requires several iterations of the Expectation Maximization (EM) algorithm [2], and can be considerably demanding computationally. As a result, fast and efficient parameter estimation remains a key challenge facing the model. We show that it is possible to reduce the Maximum Likelihood parameter estimation problem for TVM into a Singular Value Decomposition (SVD) problem by making some suitably justified approximations in the likelihood function. By using randomized algorithms for efficient computation of the SVD, it becomes possible to accelerate the parameter estimation task remarkably. In addition, we show that this method is able to increase the efficiency of the ivector extraction procedure, and also lends some interpretability to the extracted ivectors.

Stacked Long-Term TDNN for Spoken Language Recognition

Daniel Garcia-Romero, Alan McCree; Johns Hopkins University, USA

Mon-O-9-5-5, Time: 11:20

This paper introduces a stacked architecture that uses a time delay neural network (TDNN) to model long-term patterns for spoken language identification. The first component of the architecture is a feed-forward neural network with a bottleneck layer that is trained to classify context-dependent phone states (senones). The second component is a TDNN that takes the output of the bottleneck, concatenated over a long time span, and produces a posterior probability over the set of languages. The use of a TDNN architecture provides an efficient model to capture discriminative patterns over a wide temporal context. Experimental results are presented using the audio data from the language i-vector challenge (IVC) recently organized by NIST. The proposed system outperforms a state-of-the-art shifted delta cepstra i-vector system and provides complementary information to fuse with the new generation of bottleneck-based i-vector systems that model short-term dependencies.

A Divide-and-Conquer Approach for Language Identification Based on Recurrent Neural Networks

G. Gelly [1], Jean-Luc Gauvain [1], V.B. Le [2], A. Messaoudi [2]; [1] LIMSI, France; [2] Vocapia Research, France

Mon-O-9-5-6, Time: 11:40

This paper describes the design of an acoustic language recognition system based on BLSTM that can discriminate closely related languages and dialects of the same language. We introduce a *Divide-and-Conquer* (D&C) method to quickly and successfully train an RNN-based multi-language classifier. Experiments compare this approach to the straightforward training of the same RNN, as well as to two widely used LID techniques: a phonotactic system using DNN acoustic models and an i-vector system. Results are reported on two different data sets: the 14 languages of NIST LRE07 and the 20 closely related languages and dialects of NIST OpenLRE15.

NOTES

In addition to reporting the NIST Cavg metric which served as the primary metric for the LRE07 and OpenLRE15 evaluations, the EER and LER are provided. When used with BLSTM, the D&C training scheme significantly outperformed the classical training method for multi-class RNNs. On the OpenLRE15 data set, this method also outperforms classical LID techniques and combines very well with a phonotactic system.

Mon-O-9-6 : Spoken Language Understanding Systems

Seacliff A, 10:00–12:00, Monday, 12 Sept. 2016
Chairs: Helen Meng, Gokhan Tur

Context-Sensitive and Role-Dependent Spoken Language Understanding Using Bidirectional and Attention LSTMs

Chiori Hori, Takaaki Hori, Shinji Watanabe, John R. Hershey; MERL, USA

Mon-O-9-6-1, Time: 10:00

To understand speaker intentions accurately in a dialog, it is important to consider the context of the surrounding sequence of dialog turns. Furthermore, each speaker may play a different role in the conversation, such as agent versus client, and thus features related to these roles may be important to the context. In previous work, we proposed context-sensitive spoken language understanding (SLU) using role-dependent long short-term memory (LSTM) recurrent neural networks (RNNs), and showed improved performance at predicting concept tags representing the intentions of agent and client in a human-human hotel reservation task. In the present study, we use bidirectional and attention-based LSTMs to train a role-dependent context-sensitive model to jointly represent both the local word-level context within each utterance, and the left and right context within the dialog. The different roles of client and agent are modeled by switching between role-dependent layers. We evaluated label accuracies in the hotel reservation task using a variety of models, including logistic regression, RNNs, LSTMs, and the proposed bidirectional and attention-based LSTMs. The bidirectional and attention-based LSTMs yield significantly better performance in this task.

A Step Beyond Local Observations with a Dialog Aware Bidirectional GRU Network for Spoken Language Understanding

Vedran Vukotić[1], Christian Raymond[1], Guillaume Gravier[2]; [1]INSA de Rennes, France; [2]Inria, France

Mon-O-9-6-2, Time: 10:20

Architectures of Recurrent Neural Networks (RNN) recently become a very popular choice for Spoken Language Understanding (SLU) problems; however, they represent a big family of different architectures that can furthermore be combined to form more complex neural networks. In this work, we compare different recurrent networks, such as simple Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) networks, Gated Memory Units (GRU) and their bidirectional versions, on the popular ATIS dataset and on MEDIA, a more complex French dataset. Additionally, we propose a novel method where information about the presence of relevant word classes in the dialog history is combined with a bidirectional GRU, and we show that combining relevant word classes from the dialog history improves the performance over recurrent networks that work by solely analyzing the current sentence.

End-to-End Memory Networks with Knowledge Carryover for Multi-Turn Spoken Language Understanding

Yun-Nung Chen[1], Dilek Hakkani-Tür[2], Gokhan Tur[2], Jianfeng Gao[2], Li Deng[2]; [1]National Taiwan University, Taiwan; [2]Microsoft, USA

Mon-O-9-6-3, Time: 10:40

Spoken language understanding (SLU) is a core component of a spoken dialogue system. In the traditional architecture of dialogue systems, the SLU component treats each utterance independent of each other, and then the following components aggregate the multi-turn information in the separate phases. However, there are two challenges: 1) errors from previous turns may be propagated and then degrade the performance of the current turn; 2) knowledge mentioned in the long history may not be carried into the current turn. This paper addresses the above issues by proposing an architecture using end-to-end memory networks to model knowledge carryover in multi-turn conversations, where utterances encoded with intents and slots can be stored as embeddings in the memory and the decoding phase applies an attention model to leverage previously stored semantics for intent prediction and slot tagging simultaneously. The experiments on Microsoft Cortana conversational data show that the proposed memory network architecture can effectively extract salient semantics for modeling knowledge carryover in the multi-turn conversations and outperform the results using the state-of-the-art recurrent neural network framework (RNN) designed for single-turn SLU.

Sequential Convolutional Neural Networks for Slot Filling in Spoken Language Understanding

Ngoc Thang Vu; Universität Stuttgart, Germany

Mon-O-9-6-4, Time: 11:00

We investigate the usage of convolutional neural networks (CNNs) for the slot filling task in spoken language understanding. We propose a novel CNN architecture for sequence labeling which takes into account the previous context words with preserved order information and pays special attention to the current word with its surrounding context. Moreover, it combines the information from the past and the future words for classification. Our proposed CNN architecture outperforms even the previously best ensembling recurrent neural network model and achieves state-of-the-art results with an F1-score of 95.61% on the ATIS benchmark dataset without using any additional linguistic knowledge and resources.

A New Pre-Training Method for Training Deep Learning Models with Application to Spoken Language Understanding

Asli Celikyilmaz, Ruhi Sarikaya, Dilek Hakkani-Tür, Xiaohu Liu, Nikhil Ramesh, Gokhan Tur; Microsoft, USA

Mon-O-9-6-5, Time: 11:20

We propose a simple and efficient approach for pre-training deep learning models with application to slot filling tasks in spoken language understanding. The proposed approach leverages unlabeled

NOTES

data to train the models and is generic enough to work with any deep learning model. In this study, we consider the CNN2CRF architecture that contains Convolutional Neural Network (CNN) with Conditional Random Fields (CRF) as top layer, since it has shown great potential for learning useful representations for supervised sequence learning tasks. The proposed pre-training approach with this architecture learns the feature representations from both labeled and unlabeled data at the CNN layer, covering features that would not be observed in limited labeled data. At the CRF layer, the unlabeled data uses predicted classes of words as latent sequence labels together with labeled sequences. Latent labeled sequences, in principle, has the regularization effect on the labeled sequences, yielding a better generalized model. This allows the network to learn representations that are useful for not only slot tagging using labeled data but also learning dependencies both within and between latent clusters of unseen words. The proposed pre-training method with the CRF2CNN architecture achieves significant gains with respect to the strongest semi-supervised baseline.

Joint Syntactic and Semantic Analysis with a Multitask Deep Learning Framework for Spoken Language Understanding

Jeremie Tafforeau, Frederic Bechet, Thierry Artiere, Benoit Favre; LIF (UMR 7279), France

Mon-O-9-6-6, Time: 11:40

Spoken Language Understanding (SLU) models have to deal with Automatic Speech Recognition outputs which are prone to contain errors. Most of SLU models overcome this issue by directly predicting semantic labels from words without any deep linguistic analysis. This is acceptable when enough training data is available to train SLU models in a supervised way. However for open-domain SLU, such annotated corpus is not easily available or very expensive to obtain, and generic syntactic and semantic models, such as dependency parsing, Semantic Role Labeling (SRL) or FrameNet parsing are good candidates if they can be applied to noisy ASR transcriptions with enough robustness. To tackle this issue we present in this paper an RNN-based architecture for performing joint syntactic and semantic parsing tasks on noisy ASR outputs. Experiments carried on a corpus of French spoken conversations collected in a telephone call-centre are reported and show that our strategy brings an improvement over the standard pipeline approach while allowing a lot more flexibility in the model design and optimization.

Mon-P-9-1 : Language Recognition

Pacific Concourse – Poster A, 10:00–12:00, Monday, 12 Sept. 2016
Chair: Xiong Xiao

Exploiting Hidden-Layer Responses of Deep Neural Networks for Language Recognition

Ruizhi Li[1], Sri Harish Mallidi[1], Lukáš Burget[2], Oldřich Plchot[2], Najim Dehak[1]; [1]Johns Hopkins University, USA; [2]Brno University of Technology, Czech Republic

Mon-P-9-1-1, Time: 10:00

The most popular way to apply Deep Neural Network (DNN) for Language IDentification (LID) involves the extraction of bottleneck features from a network that was trained on automatic speech recognition task. These features are modeled using a classical I-vector system. Recently, a more direct DNN approach was proposed, it consists of estimating the language posteriors directly from a stacked frames input. The final decision score is based on averaging the scores for all the frames for a given speech segment. In this paper, we extended the direct DNN approach by modeling all hidden-layer activations rather than just averaging the output scores. One super-vector per utterance is formed by concatenating all hidden-layer responses. The dimensionality of this vector is then reduced using a Principal Component Analysis (PCA). The obtained reduce vector summarizes the most discriminative features for language recognition based on the trained DNNs. We evaluated this approach in NIST 2015 language recognition evaluation. The performances achieved by the proposed approach are very competitive to the classical I-vector baseline.

Out of Set Language Modelling in Hierarchical Language Identification

*Saad Irtza[1], Vidhyasaharan Sethu[1], Sarith Fernando[1], Eliathamby Ambikairajah[1], Haizhou Li[2]; [1]University of New South Wales, Australia; [2]A*STAR, Singapore*

Mon-P-9-1-2, Time: 10:00

This paper proposes a novel approach to the open set language identification task by introducing out of set (OOS) language modelling in a Hierarchical Language Identification (HLID) framework. Most recent language identification systems make use of data sources from other than target languages to model OOS languages. The proposed approach does not require such data to model OOS languages, instead it only uses data from target languages. Additionally, a diverse language selection method is incorporated to further improve OOS language modelling. This work also proposes the use of a new training data selection method to develop compact models in a hierarchical framework. Experiments are conducted on the recent NIST LRE 2015 data set. The overall results show relative improvements of 32.9% and 30.1% in terms of C_{avg} with and without the diverse language selection method respectively over the corresponding baseline systems, when using the proposed hierarchical OOS modelling.

Language Identification Based on Generative Modeling of Posteriorgram Sequences Extracted from Frame-by-Frame DNNs and LSTM-RNNs

Ryo Masumura, Taichi Asami, Hirokazu Masataki, Yushi Aono, Sumitaka Sakauchi; NTT, Japan

Mon-P-9-1-3, Time: 10:00

This paper aims to enhance spoken language identification methods based on direct discriminative modeling of language labels using deep neural networks (DNNs) and long short-term memory recurrent neural networks (LSTM-RNNs). In conventional methods, frame-by-frame DNNs or LSTM-RNNs are used for utterance-level classification. Although they have strong frame-level classification performance and real-time efficiency, they are not optimized for variable length utterance-level classification since the classification is conducted by simply averaging frame-level prediction results. In addition, the simple classification methodology cannot fully utilize the combination of DNNs and LSTM-RNNs. To address these issues, our idea is to combine the frame-by-frame DNNs and LSTM-RNNs with a sequential generative model based classifier. In the proposed method, we regard posteriorgram sequences generated from a frame-by-frame classifier as feature sequences, and model them with respect to each language using language modeling technologies. The generative model based classifier does not model an identification boundary, so we can flexibly deal with variable length utterances without loss of conventional advantages. Furthermore, the proposed method can support the combination of DNNs and LSTMs using joint posteriorgram sequences, those of generative modeling can capture differences between two posteriorgram sequences. Experiments conducted using the GlobalPhone database demonstrate the proposed method's effectiveness.

Gating Recurrent Enhanced Memory Neural Networks on Language Identification

Wang Geng, Yuanyuan Zhao, Wenfu Wang, Xinyuan Cai, Bo Xu; Chinese Academy of Sciences, China

Mon-P-9-1-4, Time: 10:00

This paper proposes a novel memory neural network structure, namely gating recurrent enhanced memory network (GREMN), to model long-range dependency in temporal series on language identification (LID) task at the acoustic frame level. The proposed GREMN is a stacking gating recurrent neural network (RNN) equipped with a learnable enhanced memory block near the classifier. It aims at capturing the long-span history and certain future contextual information of the sequential input. In addition, two optimization strategies of coherent SortaGrad-like training mechanism and a hard sample score acquisition approach are proposed. The proposed optimization policies drastically boost this memory network based LID system, especially on the large disparity training materials. It is confirmed by the experimental results that the proposed GREMN possesses strong ability of sequential modeling and generalization, where about 5% relative equal error rate (EER) reduction is obtained comparing with the approximate-sized gating RNNs and 38.5% performance improvements is observed compared to conventional i-Vector based LID system.

Sequence Summarizing Neural Networks for Spoken Language Recognition

Jan Pešán, Lukáš Burget, Jan Černocký; Brno University of Technology, Czech Republic

Mon-P-9-1-5, Time: 10:00

This paper explores the use of Sequence Summarizing Neural Networks (SSNNs) as a variant of deep neural networks (DNNs) for classifying sequences. In this work, it is applied to the task of spoken language recognition. Unlike other classification tasks in speech processing where the DNN needs to produce a per-frame output, language is considered constant during an utterance. We introduce a summarization component into the DNN structure producing one set of language posteriors per utterance. The training of the DNN is performed by an appropriately modified gradient-descent algorithm. In our initial experiments, the SSNN results are compared to a single state-of-the-art i-vector based baseline system with a similar complexity (i.e. no system fusion, etc.). For some conditions, SSNNs is able to provide performance comparable to the baseline system. Relative improvement up to 30% is obtained with the score level fusion of the baseline and the SSNN systems.

The Role of Spectral Resolution in Foreign-Accented Speech Perception

Michelle R. Kapolowicz, Vahid Montazeri, Peter F. Assmann; University of Texas at Dallas, USA

Mon-P-9-1-6, Time: 10:00

Several studies have shown that diminished spectral resolution leads to poorer speech recognition in adverse listening conditions such as competing background noise or in cochlear implants. Although intelligibility is also reduced when the talker has a foreign accent, it is unknown how limited spectral resolution interacts with foreign-accent perception. It is hypothesized that limited spectral resolution will further impair perception of foreign-accented speech. To test this, we assessed the contribution of spectral resolution to the intelligibility of foreign-accented speech by varying the number of spectral channels in a tone vocoder. We also examined listeners' abilities to discriminate between native and foreign-accented speech in each condition to determine the effect of reduced spectral resolution on accent detection. Results showed that increasing the spectral resolution improves intelligibility for foreign-accented speech while also improving listeners' ability to detect a foreign accent but not to the level of accuracy for broadband speech. Results also reveal a correlation between intelligibility and accent detection. Overall, results suggest that greater spectral resolution is needed for perception of foreign-accented speech compared to native speech.

THU-EE System Description for NIST LRE 2015

Liang He, Yao Tian, Yi Liu, Jiaming Xu, Weiwei Liu, Cai Meng, Jia Liu; Tsinghua University, China

Mon-P-9-1-7, Time: 10:00

This paper describes the systems developed by the Department of Electronic Engineering of Tsinghua University for the NIST Language Recognition Evaluation 2015. We submitted one primary and three alternative systems for the fixed training data evaluation and didn't take part in the open training data evaluation for our limited data resources and computation capability. Both the primary system and three alternative systems are fusions of multiple subsystems. The primary system and alternative systems are identical except for the training, development and fusion data. The subsystems are different in feature, statistical modeling or backend approach. The features of our subsystems include MFCC, PLP, TFC, PNCC and Fbank. The statistical modeling of our subsystems can be roughly categorized into four types: i-vector, deep neural network, multiple coordinate sequence kernel (MCSK) and phoneme recognizer followed by vector space models (PR-VSM). The backend approach includes LDA-Gaussian, SVM and extreme learning machine (ELM). Finally, these subsystems are fused by the FoCal toolkit. Our primary system is presented and briefly discussed. Post-key analyses are also addressed, including comparison of different features, modeling backend approaches and a study of their contribution to the whole performance. The processing speed for each subsystem is also given in the paper.

Variation in Spoken North Sami Language

Kristiina Jokinen[1], Trung Ngo Trong[1], Ville Hautamäki[2]; [1]University of Helsinki, Finland; [2]University of Eastern Finland, Finland

Mon-P-9-1-8, Time: 10:00

The paper sets to investigate the amount of variation between the

NOTES

North Sami speakers living in two different majority language contexts: Finnish, spoken in Finland, and Norwegian Bokmål, spoken in Norway. We hypothesize that the majority language is a significant factor in recognizing variation of the North Sami language. Although North Sami is the biggest of the nine currently spoken Sami languages and it has become a lingua franca among the Sami speakers, there are clear differences in the pronunciation of the North Sami spoken in Finland and Norway, so that the difference can be used to recognize which majority language region the speaker comes from. Using a corpus of spoken North Sami collected in locations in Finland and Norway, we experimented in classifying the speech samples into categories based on the two majority languages. We used the i-vector methodology to model both intra- and between-dialect variations, and achieved the average recognition of about 17.31% EER for classifying the Sami speech samples. The results support our hypothesis that the variation is due to the majority language, i.e. Finnish or Norwegian, spoken in the given context, rather than individual variation.

Mon-P-9-2 : Music, Audio, and Source Separation

Pacific Concourse – Poster B, 10:00–12:00, Monday, 12 Sept. 2016
Chair: Yang Liu

Improved Music Genre Classification with Convolutional Neural Networks

Weibin Zhang, Wenkang Lei, Xiangmin Xu, Xiaofeng Xing; SCUT, China

Mon-P-9-2-1, Time: 10:00

In recent years, deep neural networks have been shown to be effective in many classification tasks, including music genre classification. In this paper, we proposed two ways to improve music genre classification with convolutional neural networks: 1) combining max- and average-pooling to provide more statistical information to higher level neural networks; 2) using shortcut connections to skip one or more layers, a method inspired by residual learning method. The input of the CNN is simply the short time Fourier transforms of the audio signal. The output of the CNN is fed into another deep neural network to do classification. By comparing two different network topologies, our preliminary experimental results on the GTZAN data set show that the above two methods can effectively improve the classification accuracy, especially the second one.

Enhanced Harmonic Content and Vocal Note Based Predominant Melody Extraction from Vocal Polyphonic Music Signals

Gurunath Reddy M., K. Sreenivasa Rao; IIT Kharagpur, India

Mon-P-9-2-2, Time: 10:00

A method based on the production mechanism of the vocals in the composite vocal polyphonic music signal is proposed for vocal melody extraction. In the proposed method, initially the non-pitched percussive source is suppressed by observing its wideband spectral characteristics to emphasise the harmonic content in the mixture signal. Further, the harmonic enhanced signal is segmented into vocal and non-vocal regions by thresholding the salience energy contour. The vocal regions are further divided into vocal note like regions by their spectral transition cues in the frequency domain. The melody contour in each vocal note is extracted by detecting the locations of instant of significant excitation by passing it through adaptive zero frequency filtering (ZFF) in the time domain. The experimental results showed that the proposed method is indeed comparable to the state-of-the-art saliency based melody extraction method.

Long Short-Term Memory for Speaker Generalization in Supervised Speech Separation

Jitong Chen, DeLiang Wang; Ohio State University, USA

Mon-P-9-2-3, Time: 10:00

Speech separation can be formulated as a supervised learning problem where a time-frequency mask is estimated by a learning machine from acoustic features of noisy speech. Deep neural networks (DNNs) have been successful for noise generalization in supervised separation. However, real world applications desire a trained model to perform well with both unseen speakers and unseen noises. In this study we investigate speaker generalization for noise-independent models and propose a separation model based on long short-term memory to account for the temporal dynamics of speech. Our experiments show that the proposed model significantly outperforms a DNN in terms of objective speech intelligibility for both seen and unseen speakers. Compared to feedforward networks, the proposed model is more capable of modeling a large number of speakers, and represents an effective approach for speaker- and noise-independent speech separation.

Phonotactic Language Identification for Singing

Anna M. Kruspe; Fraunhofer IDMT, Germany

Mon-P-9-2-4, Time: 10:00

In the past decades, many successful approaches for language identification have been published. However, almost none of these approaches were developed with singing in mind. Singing has a lot of characteristics that differ from speech, such as a wider variance of fundamental frequencies and phoneme durations, vibrato, pronunciation differences, and different semantic content.

We present a new phonotactic language identification system for singing based on phoneme posteriorgrams. These posteriorgrams were extracted using acoustic models trained on English speech (*TIMIT*) and on an unannotated English-language a-capella singing dataset (*DAMP*). SVM models were then trained on phoneme statistics.

The models are evaluated on a set of amateur singing recordings from *YouTube*, and, for comparison, on the *OGI Multilanguage* corpus.

While the results on a-capella singing are somewhat worse than the ones previously obtained using i-vector extraction, this approach is easier to implement. Phoneme posteriorgrams need to be extracted for many applications, and can easily be employed for language identification using this approach. The results on singing improve significantly when the utilized acoustic models have also been trained on singing. Interestingly, the best results on the *OGI* speech corpus are also obtained when acoustic models trained on singing are used.

NOTES

Comparing the Influence of Spectro-Temporal Integration in Computational Speech Segregation

Thomas Bentsen, Tobias May, Abigail A. Kressner, Torsten Dau; Technical University of Denmark, Denmark

Mon-P-9-2-5, Time: 10:00

The goal of computational speech segregation systems is to automatically segregate a target speaker from interfering maskers. Typically, these systems include a feature extraction stage in the front-end and a classification stage in the back-end. A spectro-temporal integration strategy can be applied in either the front-end, using the so-called delta features, or in the back-end, using a second classifier that exploits the posterior probability of speech from the first classifier across a spectro-temporal window. This study systematically analyzes the influence of such stages on segregation performance, the error distributions and intelligibility predictions. Results indicated that it could be problematic to exploit context in the back-end, even though such a spectro-temporal integration stage improves the segregation performance. Also, the results emphasized the potential need of a single metric that comprehensively predicts computational segregation performance and correlates well with intelligibility. The outcome of this study could help to identify the most effective spectro-temporal integration strategy for computational segregation systems.

Blind Speech Separation with GCC-NMF

Sean U.N. Wood, Jean Rouat; Université de Sherbrooke, Canada

Mon-P-9-2-6, Time: 10:00

We introduce a blind source separation algorithm named GCC-NMF that combines unsupervised dictionary learning via non-negative matrix factorization (NMF) with spatial localization via the generalized cross correlation (GCC) method. Dictionary learning is performed on the mixture signal, with separation subsequently achieved by grouping dictionary atoms, over time, according to their spatial origins. Separation quality is evaluated using publicly available data from the SiSEC signal separation evaluation campaign consisting of stereo recordings of 3 and 4 concurrent speakers in reverberant environments. Performance is quantified using perceptual and SNR-based measures with the PEASS and BSS Eval toolkits, respectively. We compare our approach with other NMF-based speech separation algorithms including unsupervised and semi-supervised approaches. GCC-NMF outperforms the unsupervised model-based approach that combines NMF with spatial covariance mixture models, and compares favourably to semi-supervised approaches that leverage prior knowledge and information, despite being purely unsupervised itself.

Effects of Cochlear Hearing Loss on the Benefits of Ideal Binary Masking

Vahid Montazeri, Shaikat Hossain, Peter F. Assmann; University of Texas at Dallas, USA

Mon-P-9-2-7, Time: 10:00

Ideal Binary Masking (IdBM) is considered as the primary goal of computational auditory scene analysis. This binary masking criterion provides a time-frequency representation of noisy speech and retains regions where the speech dominates the noise while discarding regions where the noise is dominant. Several studies have shown the benefits of IdBM for normal hearing and hearing-impaired listeners as well as cochlear implant recipients. In this study, we evaluate the effects of simulated moderate and severe hearing loss on the masking release resulting from IdBM. Speech-shaped noise was added to IEEE sentences; the stimuli were processed using a tone-vocoder with 32 bandpass filters. The bandwidths of the filters were adjusted to account for impaired frequency selectivity observed in individuals with moderate and severe hearing loss. Following envelope extraction, the IdBM processing was then applied to the envelopes. The processed stimuli were presented to nineteen normal hearing listeners and their intelligibility scores were measured. Statistical analysis indicated that participants' benefit from IdBM was significantly reduced with impaired frequency selectivity (spectral smearing). Results show that the masking release obtained from IdBM is highly dependent on the listeners' hearing loss.

Combining Mask Estimates for Single Channel Audio Source Separation Using Deep Neural Networks

Emad M. Grais, Gerard Roma, Andrew J.R. Simpson, Mark D. Plumbley; University of Surrey, UK

Mon-P-9-2-8, Time: 10:00

Deep neural networks (DNNs) are usually used for single channel source separation to predict either soft or binary time frequency masks. The masks are used to separate the sources from the mixed signal. Binary masks produce separated sources with more distortion and less interference than soft masks. In this paper, we propose to use another DNN to combine the estimates of binary and soft masks to achieve the advantages and avoid the disadvantages of using each mask individually. We aim to achieve separated sources with low distortion and low interference between each other. Our experimental results show that combining the estimates of binary and soft masks using DNN achieves lower distortion than using each estimate individually and achieves as low interference as the binary mask.

Monaural Source Separation Using a Random Forest Classifier

Cosimo Riday, Saurabh Bhargava, Richard H.R. Hahnloser, Shih-Chii Liu; Universität Zürich, Switzerland

Mon-P-9-2-9, Time: 10:00

We address the problem of separating two audio sources from a single channel mixture recording. A novel method called Multi Layered Random Forest (MLRF) that learns a binary mask for both the sources is presented. Random Forest (RF) classifiers are trained for each frequency band of a source spectrogram. A specialized set of linear transformations are applied to a local time-frequency (T-F) neighborhood of the mixture that captures relevant local statistics. A sampling method is presented that efficiently samples T-F training bins in each frequency band. We draw equal numbers of dominant (more power) training samples from the two sources for RF classifiers that estimate the Ideal Binary Mask (IBM). An estimated IBM in a given layer is used to train a RF classifier in the next higher layer of the MLRF hierarchy. On average, MLRF performs better than deep Recurrent Neural Networks (RNNs) and Non-Negative Sparse Coding (NNSC) in signal-to-noise ratio (SNR) of reconstructed audio, overall T-F bin classification accuracy, as well as PESQ and STOI scores. Additionally, we demonstrate the ability of the MLRF to correctly reconstruct T-F bins of the target even when the latter has lower power in that frequency band.

NOTES

Adaptive Group Sparsity for Non-Negative Matrix Factorization with Application to Unsupervised Source Separation

Xu Li, Ziteng Wang, Xiaofei Wang, Qiang Fu, Yonghong Yan; Chinese Academy of Sciences, China

Mon-P-9-2-10, Time: 10:00

Non-negative matrix factorization (NMF) is an appealing technique for many audio applications, such as automatic music transcription, source separation and speech enhancement. Sparsity constraints are commonly used on the NMF model to discover a small number of dominant patterns. Recently, group sparsity has been proposed for NMF based methods, in which basis vectors belonging to a same group are permitted to activate together, while activations across groups are suppressed. However, most group sparsity models penalize all groups using a same parameter without considering the relative importance of different groups for modeling the input data. In this paper, we propose adaptive group sparsity to model the relative importance of different groups with adaptive penalty parameters and investigate its potential benefit to separate speech from other sound sources. Experimental results show that the proposed adaptive group sparsity improves the performance over regular group sparsity in unsupervised settings where neither the speaker identity nor the type of noise is known in advance.

A Robust Dual-Microphone Speech Source Localization Algorithm for Reverberant Environments

Yanmeng Guo[1], Xiaofei Wang[1], Chao Wu[1], Qiang Fu[1], Ning Ma[2], Guy J. Brown[2]; [1]Chinese Academy of Sciences, China; [2]University of Sheffield, UK

Mon-P-9-2-11, Time: 10:00

Speech source localization (SSL) using a microphone array aims to estimate the direction-of-arrival (DOA) of the speech source. However, its performance often degrades rapidly in reverberant environments. In this paper, a novel dual-microphone SSL algorithm is proposed to address this problem. First, the time-frequency regions dominated by direct sound are extracted by tracking the envelopes of speech, reverberation and background noise. The time-difference-of-arrival (TDOA) is then estimated by considering only these reliable regions. Second, a bin-wise de-aliasing strategy is introduced to make better use of the DOA information carried at high frequencies, where the spatial resolution is higher and there is typically less corruption by diffuse noise. Our experiments show that when compared with other widely-used algorithms, the proposed algorithm produces more reliable performance in realistic reverberant environments.

Speech Localisation in a Multitalker Mixture by Humans and Machines

Ning Ma, Guy J. Brown; University of Sheffield, UK

Mon-P-9-2-12, Time: 10:00

Speech localisation in multitalker mixtures is affected by the listener's expectations about the spatial arrangement of the sound sources. This effect was investigated via experiments with human listeners and a machine system, in which the task was to localise a female-voice target among four spatially distributed male-voice maskers. Two configurations were used: either the masker locations were fixed or the locations varied from trial-to-trial. The machine system uses deep neural networks (DNNs) to learn the relationship between binaural cues and source azimuth, and exploits top-down knowledge about the spectral characteristics of the target source. Performance was examined in both anechoic and reverberant conditions. Our experiments show that the machine system outperformed listeners in some conditions. Both the machine and listeners were able to make use of *a priori* knowledge about the spatial configuration of the sources, but the effect for headphone listening was smaller than that previously reported for listening in a real room.

Reverberation-Robust One-Bit TDOA Based Moving Source Localization for Automatic Camera Steering

Harshavardhan Sundar, Gokul Deepak Manavalan, T.V. Sreenivas, Chandra Sekhar Seelamantula; Indian Institute of Science, India

Mon-P-9-2-13, Time: 10:00

We address the problem of moving acoustic source localization and automatic camera steering using one-bit measurement of the time-difference of arrival (TDOA) between two microphones in a given array. Given that the camera has a finite field of view (FoV), an algorithm with a coarse estimate of the source location would suffice for the purpose. We use a microphone array and develop an algorithm to obtain a coarse estimate of the source using only one-bit information of the TDOA, the sign of it, to be precise. One advantage of the one-bit approach is that the computational complexity is lower, which aids in real-time adaptation and localization of the moving source. We carried out experiments in a reverberant enclosure with a 60 dB reverberation time of 600 ms (RT60 = 600 ms). We analyzed the performance of the proposed approach using a circular microphone array. We report comparisons with a point source localization-based automatic camera steering algorithm proposed in the literature. The proposed algorithm turned out to be more accurate in terms of always having the moving speaker within the field of view.

Multi-Talker Speech Recognition Based on Blind Source Separation with ad hoc Microphone Array Using Smartphones and Cloud Storage

Keiko Ochi[1], Nobutaka Ono[1], Shigeki Miyabe[2], Shoji Makino[2]; [1]NII, Japan; [2]University of Tsukuba, Japan

Mon-P-9-2-14, Time: 10:00

In this paper, we present a multi-talker speech recognition system based on blind source separation with an ad hoc microphone array, which consists of smartphones and cloud storage. In this system, a mixture of voices from multiple speakers is recorded by each speaker's smartphone, which is automatically transferred to online cloud storage. Our prototype system is realized using iPhone and Dropbox. Although the signals recorded by different iPhones are not synchronized, the blind synchronization technique compensates both the differences in the time offset and the sampling frequency mismatch. Then, auxiliary-function-based independent vector analysis separates the synchronized mixture into each speaker's voice. Finally, automatic speech recognition is applied to transcribe the speech. By experimental evaluation of the multi-talker speech recognition system using Julius, we confirm that it effectively reduces the speech overlap and improves the speech recognition performance.

NOTES

Mon-P-9-3 : Acoustic Modeling with Neural Networks

Pacific Concourse – Poster C, 10:00–12:00, Monday, 12 Sept. 2016
Chair: Samuel Thomas

Phase-Aware Signal Processing for Automatic Speech Recognition

Johannes Fahringer, Tobias Schrank, Johannes Stahl, Pejman Mowlaee, Franz Pernkopf; Technische Universität Graz, Austria

Mon-P-9-3-1, Time: 10:00

Conventional automatic speech recognition (ASR) often neglects the spectral phase information in its front-end and feature extraction stages. The aim of this paper is to show the impact that enhancement of the noisy spectral phase has on ASR accuracy when dealing with speech signals corrupted with additive noise. Apart from proof-of-concept experiments using clean spectral phase, we also present a phase enhancement method as a phase-aware front-end and modified group delay as a phase-aware feature extractor, and the combination thereof. In experiments, we demonstrate the improved performance for each individual component and their combination, compared to the conventional phase-unaware Mel Frequency Cepstral Coefficients (MFCCs)-based ASR. We observe that the estimated phase information used in the front-end or feature extraction component improves the ASR word accuracy rate (WAR) by 20.98% absolute for noise corrupted speech (averaged over SNRs ranging from 0 to 20 dB).

Unsupervised Deep Auditory Model Using Stack of Convolutional RBMs for Speech Recognition

Hardik B. Sailor, Hemant A. Patil; DA-IICT, India

Mon-P-9-3-2, Time: 10:00

Recently, we have proposed an unsupervised filterbank learning model based on Convolutional RBM (ConvRBM). This model is able to learn auditory-like subband filters using speech signals as an input. In this paper, we propose two-layer Unsupervised Deep Auditory Model (UDAM) by stacking two ConvRBMs. The first layer ConvRBM learns filterbank from speech signals and hence, it represents early auditory processing. The hidden units' responses of the first layer are pooled as short-time spectral representation to train another ConvRBM using greedy layer-wise method. The ConvRBM in second layer trained on spectral representation learns Temporal Receptive Field (TRF) which represent temporal properties of the auditory cortex in human brain. To show the effectiveness of the proposed UDAM, speech recognition experiments were conducted on TIMIT and AURORA 4 databases. We have shown that features extracted from second layer when added to filterbank features of first layer performs better than first layer features alone (and their delta features as well). For both databases, our proposed two-layer deep auditory features improve speech recognition performance over Mel filterbank features. Further improvements can be achieved by system-level combination of both UDAM features and Mel filterbank features.

Interpretation of Low Dimensional Neural Network Bottleneck Features in Terms of Human Perception and Production

Philip Weber, Linxue Bai, Martin Russell, Peter Jančovič, Stephen Houghton; University of Birmingham, UK

Mon-P-9-3-3, Time: 10:00

Low-dimensional 'bottleneck' features extracted from neural networks have been shown to give phoneme recognition accuracy similar to that obtained with higher-dimensional MFCCs, using GMM-HMM models. Such features have also been shown to preserve well the assumptions of speech trajectory dynamics made by dynamic models of speech such as Continuous-State HMMs. However, little is understood about how networks derive these features and how and whether they can be interpreted in terms of human speech perception and production.

We analyse three-dimensional bottleneck features. We show that for vowels, their spatial representation is very close to the familiar F_1:F_2 vowel quadrilateral. For other classes of phonemes the features can similarly be related to phonetic and acoustic spatial representations presented in the literature. This suggests that these networks derive representations specific to particular phonetic categories, with properties similar to those used by human perception. The representation of the full set of phonemes in the bottleneck space is consistent with a hypothesized comprehensive model of speech perception and also with models of speech perception such as prototype theory.

Compact Feedforward Sequential Memory Networks for Large Vocabulary Continuous Speech Recognition

Shiliang Zhang[1], Hui Jiang[2], Shifu Xiong[1], Si Wei[1], Li-Rong Dai[1]; [1]USTC, China; [2]York University, Canada

Mon-P-9-3-4, Time: 10:00

In acoustic modeling for large vocabulary continuous speech recognition, it is essential to model long term dependency within speech signals. Usually, recurrent neural network (RNN) architectures, especially the long short term memory (LSTM) models, are the most popular choice. Recently, a novel architecture, namely feedforward sequential memory networks (FSMN), provides a non-recurrent architecture to model long term dependency in sequential data and has achieved better performance over RNNs on acoustic modeling and language modeling tasks. In this work, we propose a compact feedforward sequential memory networks (cFSMN) by combining FSMN with low-rank matrix factorization. We also make a slight modification to the encoding method used in FSMNs in order to further simplify the network architecture. On the Switchboard task, the proposed new cFSMN structures can reduce the model size by 60% and speed up the learning by more than 7 times while the models still significantly outperform the popular bidirection LSTMs for both frame-level cross-entropy (CE) criterion based training and MMI based sequence training.

Future Context Attention for Unidirectional LSTM Based Acoustic Model

Jian Tang[1], Shiliang Zhang[1], Si Wei[2], Li-Rong Dai[1]; [1]USTC, China; [2]iFLYTEK, China

Mon-P-9-3-5, Time: 10:00

Recently, feedforward sequential memory networks (FSMN) has shown strong ability to model past and future long-term depen-

NOTES

dency in speech signals without using recurrent feedback, and has achieved better performance than BLSTM in acoustic modeling. However, the encoding coefficients in FSMN is context-independent while context-dependent weights are commonly supposed to be more reasonable in acoustic modeling. In this paper, we propose a novel architecture called attention-based LSTM, which employs context-dependent scores or context-dependent weights to encode temporal future context information with the help of a kind of attention mechanism for unidirectional LSTM based acoustic model. Preliminary experimental results on TIMIT corpus have shown that the proposed attention-based LSTM achieves a phone error rate (PER) of 20.8% while PER is 20.1% for BLSTM. We have also presented a lot of experiments to evaluate different context attention methods.

Hybrid Accelerated Optimization for Speech Recognition

Jen-Tzung Chien[1], Pei-Wen Huang[1], Tan Lee[2]; [1]National Chiao Tung University, Taiwan; [2]Chinese University of Hong Kong, China

Mon-P-9-3-6, Time: 10:00

Optimization procedure is crucial to achieve desirable performance for speech recognition based on deep neural networks (DNNs). Conventionally, DNNs are trained by using mini-batch stochastic gradient descent (SGD) which is stable but prone to be trapped into local optimum. A recent work based on Nesterov's accelerated gradient descent (NAG) algorithm is developed by merging the current momentum information into correction of SGD updating. NAG less likely jumps into local minimum so that convergence rate is improved. In general, optimization based on SGD is more stable while that based on NAG is faster and more accurate. This study aims to boost the performance of speech recognition by combining complimentary SGD and NAG. A new hybrid optimization is proposed by integrating the SGD with momentum and the NAG by using an interpolation scheme which is continuously run in each mini-batch according to the change rate of cost function in consecutive two learning epochs. Tradeoff between two algorithms can be balanced for mini-batch optimization. Experiments on speech recognition using CUSENT and Aurora-4 show the effectiveness of the hybrid accelerated optimization in DNN acoustic model.

On Online Attention-Based Speech Recognition and Joint Mandarin Character-Pinyin Training

William Chan, Ian Lane; Carnegie Mellon University, USA

Mon-P-9-3-7, Time: 10:00

In this paper, we explore the use of attention-based models for online speech recognition without the usage of language models or searching. Our model is based on an attention-based neural network which directly emits English/Mandarin characters as outputs. The model jointly learns the pronunciation, acoustic and language model. We evaluate the model for online speech recognition on English and Mandarin. On English, we achieve a 33.0% WER on the WSJ task, or a 5.4% absolute reduction in WER compared to an online CTC based system. We also introduce a new training method and show how we can learn joint Mandarin Character-Pinyin models. Our Mandarin character only model achieves a 72% CER on the GALE Phase 2 evaluation, and with our joint Mandarin Character-Pinyin model, we achieve 59.3% CER or 12.7% absolute improvement over the character only model.

GMM-Free Flat Start Sequence-Discriminative DNN Training

Gábor Gosztolya[1], Tamás Grósz[2], László Tóth[1]; [1]MTA-SZTE RGAI, Hungary; [2]University of Szeged, Hungary

Mon-P-9-3-8, Time: 10:00

Recently, attempts have been made to remove Gaussian mixture models (GMM) from the training process of deep neural network-based hidden Markov models (HMM/DNN). For the GMM-free training of a HMM/DNN hybrid we have to solve two problems, namely the initial alignment of the frame-level state labels and the creation of context-dependent states. Although flat-start training via iteratively realigning and retraining the DNN using a frame-level error function is viable, it is quite cumbersome. Here, we propose to use a sequence-discriminative training criterion for flat start. While sequence-discriminative training is routinely applied only in the final phase of model training, we show that with proper caution it is also suitable for getting an alignment of context-independent DNN models. For the construction of tied states we apply a recently proposed KL-divergence-based state clustering method, hence our whole training process is GMM-free. In the experimental evaluation we found that the sequence-discriminative flat start training method is not only significantly faster than the straightforward approach of iterative retraining and realignment, but the word error rates attained are slightly better as well.

Open-Domain Audio-Visual Speech Recognition: A Deep Learning Approach

Yajie Miao, Florian Metze; Carnegie Mellon University, USA

Mon-P-9-3-9, Time: 10:00

Automatic speech recognition (ASR) on video data naturally has access to two modalities: audio and video. In previous work, audio-visual ASR, which leverages visual features to help ASR, has been explored on restricted domains of videos. This paper aims to extend this idea to open-domain videos, for example videos uploaded to YouTube. We achieve this by adopting a unified deep learning approach. First, for the visual features, we propose to apply segment- (utterance-) level features, instead of highly restrictive frame-level features. These visual features are extracted using deep learning architectures which have been pre-trained on computer vision tasks, e.g., object recognition and scene labeling. Second, the visual features are incorporated into ASR under deep learning based acoustic modeling. In addition to simple feature concatenation, we also apply an adaptive training framework to incorporate visual features in a more flexible way. On a challenging video transcribing task, audio-visual ASR using our proposed approach gets notable improvements in terms of word error rates (WERs), compared to ASR merely using speech features.

Multidimensional Residual Learning Based on Recurrent Neural Networks for Acoustic Modeling

Yuanyuan Zhao, Shuang Xu, Bo Xu; Chinese Academy of Sciences, China

Mon-P-9-3-10, Time: 10:00

Theoretical and empirical evidences indicate that the depth of neural networks is crucial to acoustic modeling in speech recognition tasks. Unfortunately, the situation in practice always is that with the depth increasing, the accuracy gets saturated and then degrades rapidly. In this paper, a novel multidimensional residual learning architecture is proposed to address this degradation of deep recurrent neural networks (RNNs) on acoustic modeling by further exploring the spatial and temporal dimensions. In the spatial dimension, shortcut connections are introduced to RNNs, along which the information can flow across several layers without attenuation. In the temporal dimension, we cope with the degradation problem by regulating temporal granularity, namely, splitting the input sequence into several parallel sub-sequences, which can ensure information flowing across the time axis unimpededly. Finally, we place a row convolution layer on the top of all recurrent layers to comprehend appropriate information from several parallel sub-sequences to feed to the classifier. Experiments are illustrated on two quite different speech recognition tasks and 10% relative performance improvements are observed.

Towards Online-Recognition with Deep Bidirectional LSTM Acoustic Models

Albert Zeyer, Ralf Schlüter, Hermann Ney; RWTH Aachen University, Germany

Mon-P-9-3-11, Time: 10:00

Online-Recognition requires the acoustic model to provide posterior probabilities after a limited time delay given the online input audio data. This necessitates unidirectional modeling and the standard solution is to use unidirectional long short-term memory (LSTM) recurrent neural networks (RNN) or feed-forward neural networks (FFNN).

It is known that bidirectional LSTMs are more powerful and perform better than unidirectional LSTMs. To demonstrate the performance difference, we start by comparing several different bidirectional and unidirectional LSTM topologies.

Furthermore, we apply a modification to bidirectional RNNs to enable online-recognition by moving a window over the input stream and perform one forwarding through the RNN on each window. Then, we combine the posteriors of each forwarding and we renormalize them. We show in experiments that the performance of this online-enabled bidirectional LSTM performs as good as the offline bidirectional LSTM and much better than the unidirectional LSTM.

Advances in Very Deep Convolutional Neural Networks for LVCSR

Tom Sercu, Vaibhava Goel; IBM, USA

Mon-P-9-3-12, Time: 10:00

Very deep CNNs with small 3×3 kernels have recently been shown to achieve very strong performance as acoustic models in hybrid NN-HMM speech recognition systems. In this paper we investigate how to efficiently scale these models to larger datasets. Specifically, we address the design choice of pooling and padding along the time dimension which renders convolutional evaluation of sequences highly inefficient. We propose a new CNN design without timepadding and without timepooling, which is slightly suboptimal for accuracy, but has two significant advantages: it enables sequence training and deployment by allowing efficient convolutional evaluation of full utterances, and, it allows for batch normalization to be straightforwardly adopted to CNNs on sequence data. Through batch normalization, we recover the lost peformance from removing the time-pooling, while keeping the benefit of efficient convolutional evaluation.

We demonstrate the performance of our models both on larger scale data than before, and after sequence training. Our very deep CNN model sequence trained on the 2000h switchboard dataset obtains 9.4 word error rate on the Hub5 test-set, matching with a single model the performance of the 2015 IBM system combination, which was the previous best published result.

Acoustic Modelling from the Signal Domain Using CNNs

Pegah Ghahremani, Vimal Manohar, Daniel Povey, Sanjeev Khudanpur; Johns Hopkins University, USA

Mon-P-9-3-13, Time: 10:00

Most speech recognition systems use spectral features based on fixed filters, such as MFCC and PLP. In this paper, we show that it is possible to achieve state of the art results by making the feature extractor a part of the network and jointly optimizing it with the rest of the network. The basic approach is to start with a convolutional layer that operates on the signal (say, with a step size of 1.25 milliseconds), and aggregate the filter outputs over a portion of the time axis using a network in network architecture, and then down-sample to every 10 milliseconds for use by the rest of the network. We find that, unlike some previous work on learned feature extractors, the objective function converges as fast as for a network based on traditional features.

Because we found that iVector adaptation is less effective in this framework, we also experiment with a different adaptation method that is part of the network, where activation statistics over a medium time span (around a second) are computed at intermediate layers. We find that the resulting ‘direct-from-signal’ network is competitive with our state of the art networks based on conventional features with iVector adaptation.

Distilling Knowledge from Ensembles of Neural Networks for Speech Recognition

Yevgen Chebotar[1], Austin Waters[2]; [1]University of Southern California, USA; [2]Google, USA

Mon-P-9-3-14, Time: 10:00

Speech recognition systems that combine multiple types of acoustic models have been shown to outperform single-model systems. However, such systems can be complex to implement and too resource-intensive to use in production. This paper describes how to use *knowledge distillation* to combine acoustic models in a way that has the best of many worlds: It improves recognition accuracy significantly, can be implemented with standard training tools, and requires no additional complexity during recognition. First, we identify a simple but particularly strong type of ensemble: a late combination of recurrent neural networks with different architectures *and* training objectives. To harness such an ensemble, we use a variant of standard cross-entropy training to distill it into a single model and then discriminatively fine-tune the result. An evaluation on 2,000-hour large vocabulary tasks in 5 languages shows that the distilled models provide up to 8.9% relative WER improvement over conventionally-trained baselines with an identical number of parameters.

NOTES

Triphone State-Tying via Deep Canonical Correlation Analysis

Weiran Wang, Hao Tang, Karen Livescu; TTIC, USA

Mon-P-9-3-15, Time: 10:00

Context-dependent phone models are used in modern speech recognition systems to account for co-articulation effects. Due to the vast number of possible context-dependent phones, state-tying is typically used to reduce the number of target classes for acoustic modeling. We propose a novel approach for state-tying which is completely data dependent and requires no domain knowledge. Our method first learns low-dimensional embeddings of context-dependent phones using deep canonical correlation analysis. The learned embeddings capture similarity between triphones and are highly predictable from the acoustics. We then cluster the embeddings and use cluster IDs as tied states. The bottleneck features of a DNN predicting the tied states achieve competitive recognition accuracy on TIMIT.

Low-Rank Representation of Nearest Neighbor Posterior Probabilities to Enhance DNN Based Acoustic Modeling

Gil Luyet, Pranay Dighe, Afsaneh Asaei, Hervé Bourlard; Idiap Research Institute, Switzerland

Mon-P-9-3-16, Time: 10:00

We hypothesize that optimal deep neural networks (DNN) class-conditional posterior probabilities live in a union of low-dimensional subspaces. In real test conditions, DNN posteriors encode uncertainties which can be regarded as a superposition of unstructured sparse noise over the optimal posteriors. We aim to investigate different ways to structure the DNN outputs by exploiting low-rank representation (LRR) techniques. Using a large number of training posterior vectors, the underlying low-dimensional subspace of a test posterior is identified through nearest neighbor analysis, and low-rank decomposition enables separation of the "optimal" posteriors from the spurious uncertainties at the DNN output. Experiments demonstrate that by processing subsets of posteriors which possess strong subspace similarity, low-rank representation enables enhancement of posterior probabilities, and leads to higher speech recognition accuracy based on the hybrid DNN-hidden Markov model (HMM) system.

Mon-P-9-4 : Robustness and Adaptation

Pacific Concourse – Poster D, 10:00–12:00, Monday, 12 Sept. 2016
Chair: Kyu Jeong Han

Improving Large Vocabulary Accented Mandarin Speech Recognition with Attribute-Based I-Vectors

Hao Zheng[1], Shanshan Zhang[1], Liwei Qiao[2], Jianping Li[2], Wenju Liu[1]; [1]Chinese Academy of Sciences, China; [2]Shanxi Electric Power, China

Mon-P-9-4-1, Time: 10:00

It has been well-recognized that the accent has a great impact on the ASR of Chinese Mandarin, therefore, how to improve the performance on the accented speech has become a critical issue in this field. The attribute feature has been proven effective on modelling accented speech, resulting in a significantly improved performance in accent recognition. In this paper, we propose an attribute-based i-vector to improve the performance of speech recognition system on large vocabulary accented Mandarin speech task. The system with proposed attribute features works well especially with sufficient training data. To further promote the performance on conditions such as resource limited condition or training data mismatched condition, we also develop Multi-Task Learning Deep Neural Networks (MTL-DNNs) with attribute classification as the secondary task to improve the discriminative ability on Mandarin speech. Experiments on the 450-hour Intel accented Mandarin speech corpus demonstrate that the system with attribute-based i-vectors achieves a significant performance improvement on sufficient training data compared with the baseline DNN-HMM system. The MTL-DNNs complement the shortage of attribute-based i-vectors on data limited and mismatched conditions and obtain obvious CER reductions.

Pitch-Adaptive Front-End Features for Robust Children's ASR

S. Shahnawazuddin, Abhishek Dey, Rohit Sinha; IIT Guwahati, India

Mon-P-9-4-2, Time: 10:00

In the presented work, we explore some of the challenges in recognizing children's speech on automatic speech recognition (ASR) systems developed using adults' speech. In such mismatched ASR tasks, a severely degraded recognition performance is observed due to the gross mismatch in the acoustic attributes between those two groups of speakers. Among the various sources of mismatch, we focus on the large differences in the average pitch values across the adult and child speakers in this work. Earlier studies have shown that the Mel-filterbank employed in the feature extraction is not able to smooth out the pitch harmonics sufficiently in particularly for the high-pitched child speakers. As a result of that, the acoustic features derived for the adult and the child speakers turn out to be significantly mismatched. For addressing this problem, we propose a simple technique based on adaptive-liftering for deriving the pitch-robust features. This enables us to reduce the sensitivity of the acoustic features to the gross variations in pitch across the speakers. The proposed features are found to result in improved performance in the context of deep neural network based ASR system. Further with the use of the existing feature normalization techniques, additional gains are noted.

ASR Confidence Estimation with Speaker-Adapted Recurrent Neural Networks

Miguel Ángel del-Agua, Santiago Piqueras, Adrià Giménez, Alberto Sanchis, Jorge Civera, Alfons Juan; Universidad Politécnica de Valencia, Spain

Mon-P-9-4-3, Time: 10:00

Confidence estimation for automatic speech recognition has been very recently improved by using Recurrent Neural Networks (RNNs), and also by speaker adaptation (on the basis of Conditional Random Fields). In this work, we explore how to obtain further improvements by combining RNNs and speaker adaptation. In particular, we explore different speaker-dependent and -independent data representations for Bidirectional Long Short Term Memory RNNs of various topologies. Empirical tests are reported on the LibriSpeech dataset showing that the best results are achieved by the proposed combination of RNNs and speaker adaptation.

NOTES

Automatic Correction of ASR Outputs by Using Machine Translation

*Luis Fernando D'Haro, Rafael E. Banchs; A*STAR, Singapore*

Mon-P-9-4-4, Time: 10:00

One of the main challenges when working with a domain-independent automatic speech recognizers (ASR) is to correctly transcribe rare or out-of-vocabulary words that are not included in the language model or whose probabilities are sub-estimated. Although the common solution would be to adapt the language models and pronunciation vocabularies, in some conditions, like when using free online recognizers, that is not possible and therefore it is necessary to apply post-recognition rectifications. In this paper, we propose an automatic correction procedure based on using a phrase-based machine translation system trained using words and phonetic encoding representations to the generated n-best lists of ASR results. Our experiments on two different datasets: human computer interfaces for robots, and human to human dialogs about tourism information show that the proposed methodology can provide a quick and robust mechanism to improve the performance of the ASR by reducing the word error rate (WER) and character error rate (CER).

A Framework for Practical Multistream ASR

Sri Harish Mallidi, Hynek Hermansky; Johns Hopkins University, USA

Mon-P-9-4-5, Time: 10:00

Robustness of automatic speech recognition (ASR) to acoustic mismatches can be improved by using multistream architecture. Past multistream approaches involve training large number of neural networks, one for each possible stream combination. During testing phase, each utterance is forward passed through all the neural networks to estimate best stream combination. In this work, we propose a new framework to reduce the complexity of multistream architecture. We show that multiple neural networks, used in the past approaches, can be replaced by a single neural network. This results in a significant decrease in the number of parameters used in the system. The test time complexity is also reduced by organizing the stream combinations in a tree structure, where each node in the tree represent a stream combination. Instead of traversing through all the nodes, we traverse through paths which resulted in a increase in the performance monitor score. Compared to state-of-the-art baseline system, the proposed approach resulted in 13.5% relative improvement word-error-rate (WER) in Aurora4 speech recognition task. We also obtained an average of 0.7% absolute decrease in WER in 5 IARPRA-BABEL Year 4 languages.

DNNs for Unsupervised Extraction of Pseudo FMLLR Features Without Explicit Adaptation Data

Neethu Mariam Joy, Murali Karthick Baskar, S. Umesh, Basil Abraham; IIT Madras, India

Mon-P-9-4-6, Time: 10:00

In this paper, we propose the use of deep neural networks (DNN) as a regression model to estimate feature-space maximum likelihood linear regression (FMLLR) features from unnormalized features. During training, the pair of unnormalized features as input and corresponding FMLLR features as target are provided and the network is optimized to reduce the mean-square error between output and target FMLLR features. During test, the unnormalized features are passed through this DNN feature extractor to obtain FMLLR-like features without any supervision or first pass decode. Further, the FMLLR-like features are generated frame-by-frame, requiring no explicit adaptation data to extract the features unlike in FMLLR or *i*-vector. Our proposed approach is therefore suitable for scenarios where there is little adaptation data. The proposed approach provides sizable improvements over basis-FMLLR and conventional FMLLR when normalization is done at utterance level on TIMIT and Switchboard-33hour data sets.

Multi-Attribute Factorized Hidden Layer Adaptation for DNN Acoustic Models

Lahiru Samarakoon, Khe Chai Sim; NUS, Singapore

Mon-P-9-4-7, Time: 10:00

Recently, the Factorized Hidden Layer (FHL) adaptation is proposed for speaker adaptation of deep neural network (DNN) based acoustic models. In addition to the standard affine transformation, an FHL contains a speaker-dependent (SD) transformation matrix using a linear combination of rank-1 matrices and an SD bias using a linear combination of vectors. In this work, we extend the FHL based adaptation to multiple variabilities of the speech signal. Experimental results on Aurora4 task show 26.0% relative improvement over the baseline when standard FHL adaptation is used for speaker adaptation. The Multi-attribute FHL adaptation shows gains over the standard FHL adaptation where improvements reach up to 29.0% relative to the baseline.

Speaker Normalization Through Feature Shifting of Linearly Transformed i-Vector

Jahyun Goo, Younggwan Kim, Hyungjun Lim, Hoirin Kim; KAIST, Korea

Mon-P-9-4-8, Time: 10:00

In this paper, we propose a simple speaker normalization for deep neural network (DNN) using i-vectors, the state-of-the-art technique for speaker recognition, for automatic speech recognition. There have been already many techniques using i-vectors for speaker adaptation or speaker variability reduction of DNN acoustic models. However, in order to add the speaker information into the acoustic feature, most of those techniques have to train a large number of parameters while dimensionality of the i-vector is quite small. We tried to apply a component-wise shift to the acoustic features by linearly transformed i-vector, and then achieved the better performance than typical approaches. On top of that, we propose to modify this structure to adapt each frame of the features, reducing the number of parameters. Experiments were conducted on the TED-LIUM release-1 corpus, and the proposed method showed some performance gains.

NOTES

Mon-SE-4 : Special Event: Computational Approaches to Linguistic Code Switching

Grand Ballroom A, 12:15–13:00, Monday, 12 Sept. 2016
Chairs: Julia Hirschberg, Abeer Alwan, Mona Diab, Thamar Solorio

Computational Approaches to Linguistic Code Switching

Mona Diab [1], Pascale Fung [2], Julia Hirschberg [3], Thamar Solorio [4]; [1] George Washington University, USA; [2] HKUST, China; [3] Columbia University, USA; [4] University of Houston, USA

Mon-SE-4, Time: 12:15

Code-switching (CS) is the phenomenon by which multilingual speakers switch back and forth between their common languages in written or spoken communication. CS may occur at the inter-utterance, intra-utterance (mixing of words from multiple languages in the same utterance) and even morphological (mixing of morphemes from different languages) levels. CS presents serious challenges for language technologies such as Automatic Speech Recognition, Language Modeling, Parsing, Machine Translation (MT), Information Retrieval (IR) and Extraction (IE), Keyword Search, and semantic processing. A prime example of this is acoustic modeling and language modeling in automatic speech recognition (ASR): techniques trained on one language quickly break down when there is mixed language input. The lack of basic tools such as language models, part-of-speech (POS) taggers and parsers trained on such mixed language data makes downstream tasks even more challenging. Even for problems that are largely considered solved for monolingual corpora, such as Language Identification, or POS Tagging, performance degrades at a rate proportional to the amount and level of mixed-language present in the data.

This special event is to bring together researchers interested in solving the CS problem, to raise community awareness of the (limited) resources available and the work currently underway for the study of CS, with particular emphasis on work in the speech community. The format will consist of a short introduction from the organizers followed by discussion. We held a workshop in CS in conjunction with EMNLP 2014, developing a shared text-based task for this purpose. We received 18 regular workshop submissions and accepted 8. The goal of this event is to engage the speech processing community now working in this area and to encourage new research by those now working primarily with monolingual corpora.

We will solicit participation from researchers working in speech processing for the analysis and/processing of CS data. Topics of relevance to the event will include the following: • Methods for improving ASR acoustic and language models in code switched data; • Domain/dialect/genre adaptation techniques applied to CS data processing; • Challenges of language identification in CS data; • Speech-to-speech translation in CS data; • Keyword search in CS data; • Cross-lingual approaches to CS; • Development of corpora to support research on CS data; • Crowdsourcing approaches for the annotation of code switched data.

Mon-O-10-1 : Neural Networks for Language Modeling

Grand Ballroom A, 13:30–15:30, Monday, 12 Sept. 2016
Chairs: Steve Renals, Ebru Arisoy

Compositional Neural Network Language Models for Agglutinative Languages

Ebru Arisoy [1], Murat Saraclar [2]; [1] MEF Üniversitesi, Turkey; [2] Boğaziçi Üniversitesi, Turkey

Mon-O-10-1-1, Time: 13:30

Continuous space language models (CSLMs) have been proven to be successful in speech recognition. With proper training of the word embeddings, words that are semantically or syntactically related are expected to be mapped to nearby locations in the continuous space. In agglutinative languages, words are made up of concatenation of stems and suffixes and, as a result, compositional modeling is important. However, when trained on word tokens, CSLMs do not explicitly consider this structure. In this paper, we explore compositional modeling of stems and suffixes in a long short-term memory neural network language model. Our proposed models jointly learn distributed representations for stems and endings (concatenation of suffixes) and predict the probability for stem and ending sequences. Experiments on the Turkish Broadcast news transcription task show that further gains on top of a state-of-the-art stem-ending-based n-gram language model can be obtained with the proposed models.

NN-Grams: Unifying Neural Network and n-Gram Language Models for Speech Recognition

Babak Damavandi, Shankar Kumar, Noam Shazeer, Antoine Bruguier; Google, USA

Mon-O-10-1-2, Time: 13:50

We present NN-grams, a novel, hybrid language model integrating n-grams and neural networks (NN) for speech recognition. The model takes as input both word histories as well as n-gram counts. Thus, it combines the memorization capacity and scalability of an n-gram model with the generalization ability of neural networks. We report experiments where the model is trained on 26B words. NN-grams are efficient at runtime since they do not include an output soft-max layer. The model is trained using noise contrastive estimation (NCE), an approach that transforms the estimation problem of neural networks into one of binary classification between data samples and noise samples. We present results with noise samples derived from either an n-gram distribution or from speech recognition lattices. NN-grams outperforms an n-gram model on an Italian speech recognition dictation task.

Recurrent Neural Network Language Model with Incremental Updated Context Information Generated Using Bag-of-Words Representation

Md. Akmal Haidar, Mikko Kurimo; Aalto University, Finland

Mon-O-10-1-3, Time: 14:10

Recurrent neural network language model (RNNLM) is becoming popular in the state-of-the-art speech recognition systems. However, it can not remember long term patterns well due to a so-called

NOTES

vanishing gradient problem. Recently, Bag-of-words (BOW) representation of a word sequence is frequently used as a context feature to improve the performance of a standard feed-forward NNLM. However, the BOW features have not been shown to benefit RNNLM. In this paper, we introduce a technique using BOW features to remember long term dependencies in RNNLM by creating a context feature vector in a separate non-linear context layer during the training of RNNLM. The context information is incrementally updated based on the BOW features and processed further in a non-linear context layer. The output of this layer is used as a context feature vector and fed into the hidden and output layers of the RNNLM. Experiments with Penn Treebank corpus indicate that our approach can provide lower perplexity with fewer parameters and faster training compared to the conventional RNNLM. Moreover, we carried out speech recognition experiments with Wall Street Journal corpus and achieved lower word error rate than RNNLM.

Sequential Recurrent Neural Networks for Language Modeling

Youssef Oualil, Clayton Greenberg, Mittul Singh, Dietrich Klakow; Universität des Saarlandes, Germany

Mon-O-10-1-4, Time: 14:30

Feedforward Neural Network (FNN)-based language models estimate the probability of the next word based on the history of the last N words, whereas Recurrent Neural Networks (RNN) perform the same task based only on the last word and some context information that cycles in the network. This paper presents a novel approach, which bridges the gap between these two categories of networks. In particular, we propose an architecture which takes advantage of the explicit, sequential enumeration of the word history in FNN structure while enhancing each word representation at the projection layer through recurrent context information that evolves in the network. The context integration is performed using an additional word-dependent weight matrix that is also learned during the training. Extensive experiments conducted on the Penn Treebank (PTB) and the Large Text Compression Benchmark (LTCB) corpus showed a significant reduction of the perplexity when compared to state-of-the-art feedforward as well as recurrent neural network architectures.

Word-Phrase-Entity Recurrent Neural Networks for Language Modeling

Michael Levit, Sarangarajan Parthasarathy, Shuangyu Chang; Microsoft, USA

Mon-O-10-1-5, Time: 14:50

The recently introduced framework of Word-Phrase-Entity language modeling is applied to Recurrent Neural Networks and leads to similar improvements as reported for n-gram language models. In the proposed architecture, RNN LMs do not operate in terms of lexical items (words), but consume sequences of tokens that could be words, word phrases or classes such as named entities, with the optimal representation for a particular input sentence determined in an iterative manner. We show how auxiliary techniques previously described for n-gram WPE language models, such as token-level interpolation and personalization, can also be realized with recurrent networks and lead to similar perplexity improvements.

LSTM, GRU, Highway and a Bit of Attention: An Empirical Overview for Language Modeling in Speech Recognition

Kazuki Irie, Zoltán Tüske, Tamer Alkhouli, Ralf Schlüter, Hermann Ney; RWTH Aachen University, Germany

Mon-O-10-1-6, Time: 15:10

Popularized by the long short-term memory (LSTM), multiplicative gates have become a standard means to design artificial neural networks with intentionally organized information flow. Notable examples of such architectures include gated recurrent units (GRU) and highway networks. In this work, we first focus on the evaluation of each of the classical gated architectures for language modeling for large vocabulary speech recognition. Namely, we evaluate the highway network, lateral network, LSTM and GRU. Furthermore, the motivation underlying the highway network also applies to LSTM and GRU. An extension specific to the LSTM has been recently proposed with an additional highway connection between the memory cells of adjacent LSTM layers. In contrast, we investigate an approach which can be used with both LSTM and GRU: a highway network in which the LSTM or GRU is used as the transformation function. We found that the highway connections enable both standalone feedforward and recurrent neural language models to benefit better from the deep structure and provide a slight improvement of recognition accuracy after interpolation with count models. To complete the overview, we include our initial investigations on the use of the attention mechanism for learning word triggers.

Mon-O-10-2 : Special Session: Sub-Saharan African Languages: From Speech Fundamentals to Applications

Grand Ballroom BC, 13:30–15:30, Monday, 12 Sept. 2016

Chairs: Martine Adda-Decker, Laurent Besacier, Marelie Davel, Larry Hyman, Martin Jansche, Francois Pellegrino, Olivier Rosec, Sebastian Stüker, Martha Tachbelie Yifiru

Automatic Speech Recognition Using Probabilistic Transcriptions in Swahili, Amharic, and Dinka

Amit Das, Preethi Jyothi, Mark Hasegawa-Johnson; University of Illinois at Urbana-Champaign, USA

Mon-O-10-2-1, Time: 13:30

In this study, we develop automatic speech recognition systems for three sub-Saharan African languages using probabilistic transcriptions collected from crowd workers who neither speak nor have any familiarity with the African languages. The three African languages in consideration are Swahili, Amharic, and Dinka. There is a language mismatch in this scenario. More specifically, utterances spoken in African languages were transcribed by crowd workers who were mostly native speakers of English. Due to this, such transcriptions are highly prone to label inaccuracies. First, we use a recently introduced technique called mismatched crowdsourcing which processes the raw crowd transcriptions to confusion networks. Next, we adapt both multilingual hidden Markov models (HMM) and deep neural network (DNN) models using the probabilistic transcriptions of the African languages. Finally, we report the results using both deterministic and probabilistic phone error rates (PER). Automatic speech recognition systems developed using this recipe are particularly useful for low resource languages where there is limited access to linguistic resources and/or transcribers in the native language.

NOTES

Speed Perturbation and Vowel Duration Modeling for ASR in Hausa and Wolof Languages

Elodie Gauthier[1], Laurent Besacier[1], Sylvie Voisin[2]; [1]LIG (UMR 5217), France; [2]DDL (UMR 5596), France

Mon-O-10-2-2, Time: 13:45

Automatic Speech Recognition (ASR) for (under-resourced) Sub-Saharan African languages faces several challenges: small amount of transcribed speech, written language normalization issues, few text resources available for language modeling, as well as specific features (tones, morphology, etc.) that need to be taken into account seriously to optimize ASR performance. This paper tries to address some of the above challenges through the development of ASR systems for two Sub-Saharan African languages: Hausa and Wolof. First, we investigate data augmentation technique (through speed perturbation) to overcome the lack of resources. Secondly, the main contribution is our attempt to model vowel length contrast existing in both languages. For reproducible experiments, the ASR systems developed for Hausa and Wolof are made available to the research community on *github*. To our knowledge, the Wolof ASR system presented in this paper is the first large vocabulary continuous speech recognition system ever developed for this language.

Improving the Lwazi ASR Baseline

Charl van Heerden, Neil Kleynhans, Marelie Davel; North-West University, South Africa

Mon-O-10-2-3, Time: 14:00

We investigate the impact of recent advances in speech recognition techniques for under-resourced languages. Specifically, we review earlier results published on the Lwazi ASR corpus of South African languages, and experiment with additional acoustic modeling approaches. We demonstrate large gains by applying current state-of-the-art techniques, even if the data itself is neither extended nor improved. We analyze the various performance improvements observed, report on comparative performance per technique — across all eleven languages in the corpus — and discuss the implications of our findings for under-resourced languages in general.

Preliminary Experiments on Unsupervised Word Discovery in Mboshi

Pierre Godard[1], Gilles Adda[1], Martine Adda-Decker[1], Alexandre Allauzen[1], Laurent Besacier[2], Hélène Bonneau-Maynard[1], Guy-Noël Kouarata[3], Kevin Löser[1], Annie Rialland[3], François Yvon[1]; [1]LIMSI, France; [2]LIG (UMR 5217), France; [3]LPP (UMR 7018), France

Mon-O-10-2-4, Time: 14:15

The necessity to document thousands of endangered languages encourages the collaboration between linguists and computer scientists in order to provide the documentary linguistics community with the support of automatic processing tools. The French-German ANR-DFG project *Breaking the Unwritten Language Barrier* (BULB) aims at developing such tools for three mostly unwritten African languages of the Bantu family. For one of them, Mboshi, a language originating from the “Cuvette” region of the Republic of Congo, we investigate unsupervised word discovery techniques from an unsegmented stream of phonemes. We compare different models and algorithms, both monolingual and bilingual, on a new corpus in Mboshi and French, and discuss various ways to represent the data with suitable granularity. An additional French-English corpus allows us to contrast the results obtained on Mboshi and to experiment with more data.

Unsupervised Phoneme Segmentation of Previously Unseen Languages

Marco Vetter[1], Markus Müller[1], Fatima Hamlaoui[2], Graham Neubig[3], Satoshi Nakamura[3], Sebastian Stüker[1], Alex Waibel[1]; [1]KIT, Germany; [2]ZAS Berlin, Germany; [3]NAIST, Japan

Mon-O-10-2-5, Time: 14:30

In this paper we investigate the automatic detection of phoneme boundaries in audio recordings of an unknown language. This work is motivated by the needs of the project BULB which aims to support linguists in documenting unwritten languages. The automatic phonemic transcription of recordings of the unwritten language is part of this. We cannot use multilingual phoneme recognizers as their phoneme inventory might not completely cover that of the new language. Thus we opted for pursuing a two step approach which is inspired by work from speech synthesis for previously unknown languages. First, we detect boundaries for phonemes, and then we classify the detected segments into phoneme units. In this paper we address the first step, i.e. the detection of the phoneme boundaries. For this we again used multilingual and crosslingual phoneme recognizers but were only interested in the phoneme boundaries detected by them and not the phoneme identities. We measured the quality of the segmentations obtained this way using precision, recall and F-measure. We compared the performance of different configurations of mono- and multilingual phoneme recognizers among each other and against a monolingual gold standard. Finally we applied the technique to Basaa, a Bantu language.

CNN-Based Phone Segmentation Experiments in a Less-Represented Language

Céline Manenti, Thomas Pellegrini, Julien Pinquier; IRIT, France

Mon-O-10-2-6, Time: 14:45

These last years, there has been a regain of interest in unsupervised sub-lexical and lexical unit discovery. Speech segmentation into phone-like units may be a first interesting step for such a task. In this article, we report speech segmentation experiments in Xitsonga, a less-represented language spoken in South Africa. We chose to use convolutional neural networks (CNN) with FBANK static coefficients as input. The models take binary decisions whether a boundary is present or not at each signal sliding frame. We compare the use of a model trained exclusively on Xitsonga data to the use of a bootstrap model trained on a larger corpus of another language, the BUCKEYE U.S. English corpus. Using a two-convolution-layer model, a 79% F-measure was obtained on BUCKEYE, with a 20 ms error tolerance. This performance is equal to the human inter-annotator agreement rate. We then used this bootstrap model to segment Xitsonga data and compared the results when adapting it with 1 to 20 minutes of Xitsonga data.

Notes

Part-of-Speech Tagging and Chunking in Text-to-Speech Synthesis for South African Languages

Georg I. Schlünz, Nkosikhona Dlamini, Rynhardt P. Kruger; CSIR, South Africa

Mon-O-10-2-7, Time: 15:00

Text-to-speech synthesis can be an empowering communication tool in the hands of the print-disabled or augmentative and alternative communication user. In an effort to improve the naturalness of synthesised speech — and thus enhance the communication experience — we apply the natural language processing tasks of part-of-speech tagging and chunking to the text in the synthesis process. We cover the South African languages of (South African) English, Afrikaans, isiXhosa, isiZulu and Sepedi. The part-of-speech tagging delivers positive results for most of the languages; however, the chunking does not give any improvement in its current form.

The Effect of Postlexical Deletion on Automatic Speech Recognition in Fast Spontaneously Spoken Zulu

Ewald van der Westhuizen, Thomas Niesler; Stellenbosch University, South Africa

Mon-O-10-2-8, Time: 15:15

We consider the phenomenon of postlexical deletion in fast spontaneously spoken isiZulu speech and its implication for automatic speech recognition (ASR). Analysis of hand-crafted transcripts of fast spontaneous speech recorded from broadcast media indicates that postlexical deletion, especially of vowels, is common in isiZulu. We show that ASR performance can be increased by inclusion of pronunciation variants that model such deletions. We also apply a sequence modelling approach normally used for grapheme-to-phoneme (G2P) conversion to generate orthography containing synthetic deletions. These synthetically generated contacted words are subsequently used to generate accompanying pronunciations using conventional G2P conversion. We evaluate an ASR system using these synthetically generated pronunciations, and compare it to a baseline system without such variants as well as an oracle system. Augmentation with synthetically generated pronunciations leads to an absolute improvement in word error rate (WER) of 2.36% relative to the baseline. Furthermore, the augmented system performs almost as well as the oracle system, with an absolute difference in WER of 0.38%.

Mon-O-10-3 : Speech Production Models

Bayview A, 13:30–15:30, Monday, 12 Sept. 2016
Chairs: Sofia Strömbergsson, Takayuki Arai

A New Model of Speech Motor Control Based on Task Dynamics and State Feedback

Vikram Ramanarayanan[1], Benjamin Parrell[2], Louis Goldstein[3], Srikantan Nagarajan[4], John Houde[4]; [1]Educational Testing Service, USA; [2]University of Delaware, USA; [3]University of Southern California, USA; [4]University of California at San Francisco, USA

Mon-O-10-3-1, Time: 13:30

We present a new model of speech motor control (TD-SFC) based on articulatory goals that explicitly incorporates acoustic sensory feedback using a framework for state-based control. We do this by combining two existing, complementary models of speech motor control — the Task Dynamics model [1] and the State Feedback Control model of speech [2]. We demonstrate the effectiveness of the combined model by simulating a simple formant perturbation study, and show that the model qualitatively reproduces the behavior of online compensation for unexpected perturbations reported in human subjects.

Using a Biomechanical Model and Articulatory Data for the Numerical Production of Vowels

Saeed Dabbaghchian[1], Marc Arnela[2], Olov Engwall[1], Oriol Guasch[2], Ian Stavness[3], Pierre Badin[4]; [1]KTH, Sweden; [2]Universitat Ramon Llull, Spain; [3]University of Saskatchewan, Canada; [4]GIPSA, France

Mon-O-10-3-2, Time: 13:50

We introduce a framework to study speech production using a biomechanical model of the human vocal tract, ArtiSynth. Electromagnetic articulography data was used as input to an inverse tracking simulation that estimates muscle activations to generate 3D jaw and tongue postures corresponding to the target articulator positions. For acoustic simulations, the vocal tract geometry is needed, but since the vocal tract is a cavity rather than a physical object, its geometry does not explicitly exist in a biomechanical model. A fully-automatic method to extract the 3D geometry (surface mesh) of the vocal tract by blending geometries of the relevant articulators has therefore been developed. This automatic extraction procedure is essential, since a method with manual intervention is not feasible for large numbers of simulations or for generation of dynamic sounds, such as diphthongs. We then simulated the vocal tract acoustics by using the Finite Element Method (FEM). This requires a high quality vocal tract mesh without irregular geometry or self-intersections. We demonstrate that the framework is applicable to acoustic FEM simulations of a wide range of vocal tract deformations. In particular we present results for cardinal vowel production, with muscle activations, vocal tract geometry, and acoustic simulations.

A New Model for Acoustic Wave Propagation and Scattering in the Vocal Tract

Jianguo Wei[1], Wendan Guan[1], Darcy Q. Hou[1], Dingyi Pan[2], Wenhuan Lu[1], Jianwu Dang[1]; [1]Tianjin University, China; [2]Zhejiang University, China

Mon-O-10-3-3, Time: 14:10

A new and efficient numerical model is proposed for simulating the acoustic wave propagation and scattering problems due to a complex geometry. In this model, the linearized Euler equations are solved by the finite-difference time-domain (FDTD) method on an orthogonal Eulerian grid. The complex wall boundary represented by a series of Lagrangian points is numerically treated by the immersed boundary method (IBM). To represent the interaction between these two systems, a force field is added to the momentum equation, which is calculated on the Lagrangian points and interpolated to the nearby Eulerian points. The pressure and velocity fields are then calculated alternatively using FDTD. The developed model is verified in the case of acoustic scattering by a cylinder, for which the exact solutions exist. The model is then applied to sound wave propagation in a 2D vocal tract with area function extracted from

Notes

MRI data. To show the advantage of present model, the grid points are non-aligned with the boundary. The numerical results have good agreements with solutions in literature. A FDTD calculation with boundary condition directly imposed on the grid points closest to the wall cannot give a reasonable solution.

Uncontrolled Manifolds in Vowel Production: Assessment with a Biomechanical Model of the Tongue

Andrew Szabados, Pascal Perrier; GIPSA, France

Mon-O-10-3-4, Time: 14:30

Motor equivalence is a key feature of speech motor control, since speakers must constantly adapt to various phonetic contexts and speaking conditions. The Uncontrolled Manifold (UCM) idea offers a theoretical framework for considering motor equivalence. In this framework coordination among motor control variables is separated into two subspaces, one in which changes in control variables modify the acoustic output and another one in which these changes do not influence the output. Our work develops this concept for speech production using a 2D biomechanical model of the tongue, coupled with a jaw and lip model, for vowel production. We first propose a representation of the linearized UCM based on orthogonal projection matrices. Next we characterize the UCMs of various vocal tract configurations of the 10 French oral vowels using their perturbation responses. We then investigate whether these UCMs describe phonetic classes like phonemes, front/back vowels, rounded/unrounded vowels, or whether they significantly vary across representatives of these different classes. We found they clearly differ between rounded and unrounded vowels, but are quite similar within each category. This suggests that similar motor equivalence strategies can be implemented within each of these classes and that UCMs provide a valid characterization of an equivalence strategy.

Experimental Validation of Sound Generated from Flow in Simplified Vocal Tract Model of Sibilant /s/

Tsukasa Yoshinaga[1], Kazunori Nozaki[2], Shigeo Wada[1]; [1]Osaka University, Japan; [2]Osaka University Dental Hospital, Japan

Mon-O-10-3-5, Time: 14:50

The coupled numerical simulation of flow and sound generation in a simplified vocal tract model of sibilant /s/ were validated with experimental measurements. The simplified model consists of incisors and four rectangular channels representing a throat, constriction, space behind the incisors, and lips. Velocity distribution and far-field sound were measured by a hot-wire anemometer and an acoustic microphone, respectively. Simulated amplitude of velocity fluctuation at the flow separation region was stabilized by increasing the grid resolution, and agreed with those of the measurement. Amplitude of sound pressure simulated by the low-resolution grids was larger than that of the high-resolution grids, indicating that calculation accuracy of velocity fluctuation at the separation region is required to simulate sound generation of the sibilant /s/.

Bayesian Modeling in Speech Motor Control: A Principled Structure for the Integration of Various Constraints

Jean-François Patri[1], Pascal Perrier[1], Julien Diard[2]; [1]GIPSA, France; [2]LPNC, France

Mon-O-10-3-6, Time: 15:10

Speaking involves sequences of linguistic units that can be produced under different sets of control strategies. For instance, a given phoneme can be achieved with different acoustic properties, and a sequence of phonemes can be performed at different speech rates and with different prosodies. How does the Central Nervous System select a specific control strategy among all the available ones? In a previously published article we proposed a Bayesian model that addressed this question with respect to the multiplicity of acoustic realizations of a sequence of phonemes. One of the strengths of Bayesian modeling is that it is well adapted to the combination of multiple constraints. In the present paper we illustrate this feature by defining an extension of our previous model that includes force constraints related to the level of effort for the production of phoneme sequences, as it could be the case in clear versus casual speech. The integration of this additional constraint is used to model the control of articulation clarity. Pertinence of the results is illustrated by controlling a biomechanical model of the vocal tract for speech production.

Mon-O-10-4 : Speaker States and Traits

Bayview B, 13:30–15:30, Monday, 12 Sept. 2016

Chairs: Agustin Gravano, Mark Huckvale

Facing Realism in Spontaneous Emotion Recognition from Speech: Feature Enhancement by Autoencoder with LSTM Neural Networks

Zixing Zhang[1], Fabien Ringeval[2], Jing Han[1], Jun Deng[1], Erik Marchi[1], Björn Schuller[1]; [1]Universität Passau, Germany; [2]LIG (UMR 5217), France

Mon-O-10-4-1, Time: 13:30

During the last decade, speech emotion recognition technology has matured well enough to be used in some real-life scenarios. However, these scenarios require an almost silent environment to not compromise the performance of the system. Emotion recognition technology from speech thus needs to evolve and face more challenging conditions, such as environmental additive and convolutional noises, in order to broaden its applicability to real-life conditions. This contribution evaluates the impact of a front-end feature enhancement method based on an autoencoder with long short-term memory neural networks, for robust emotion recognition from speech. Support Vector Regression is then used as a back-end for time- and value-continuous emotion prediction from enhanced features. We perform extensive evaluations on both non-stationary additive noise and convolutional noise, on a database of spontaneous and natural emotions. Results show that the proposed method significantly outperforms a system trained on raw features, for both arousal and valence dimensions, while having almost no degradation when applied to clean speech.

NOTES

Defining Emotionally Salient Regions Using Qualitative Agreement Method

Srinivas Parthasarathy, Carlos Busso; University of Texas at Dallas, USA

Mon-O-10-4-2, Time: 13:50

Conventional emotion classification methods focus on predefined segments such as sentences or speaking turns that are labeled and classified at the segment level. However, the emotional state dynamically fluctuates during human interactions, so not all the segments have the same relevance. We are interested in detecting regions within the interaction where the emotions are particularly salient, which we refer to as *emotional hotspots*. A system with this capability can have real applications in many domains. A key step towards building such a system is to define reliable hotspot labels, which will dictate the performance of machine learning algorithms. Creating ground-truth labels from scratch is both expensive and time consuming. This paper also demonstrates that defining those emotionally salient segments using perceptual evaluation is a hard problem resulting in low inter-evaluator agreement. Instead, we propose to define emotionally salient regions leveraging existing time-continuous emotional labels. The proposed approach relies on the *qualitative agreement* (QA) method, which dynamically captures increasing or decreasing trends across emotional traces provided by multiple evaluators. The proposed method is more reliable than just averaging traces across evaluators, providing the flexibility to define hotspots at various reliability levels without having to recollect new perceptual evaluations.

Representation Learning for Speech Emotion Recognition

Sayan Ghosh[1], Eugene Laksana[1], Louis-Philippe Morency[2], Stefan Scherer[1]; [1]University of Southern California, USA; [2]Carnegie Mellon University, USA

Mon-O-10-4-3, Time: 14:10

Speech emotion recognition is an important problem with applications as varied as human-computer interfaces and affective computing. Previous approaches to emotion recognition have mostly focused on extraction of carefully engineered features and have trained simple classifiers for the emotion task. There has been limited effort at representation learning for affect recognition, where features are learnt directly from the signal waveform or spectrum. Prior work also does not investigate the effect of transfer learning from affective attributes such as valence and activation to categorical emotions. In this paper, we investigate emotion recognition from spectrogram features extracted from the speech and glottal flow signals; spectrogram encoding is performed by a stacked autoencoder and an RNN (Recurrent Neural Network) is used for classification of four primary emotions. We perform two experiments to improve RNN training : (1) Representation Learning — Model training on the glottal flow signal to investigate the effect of speaker and phonetic invariant features on classification performance (2) Transfer Learning — RNN training on valence and activation, which is adapted to a four emotion classification task. On the USC-IEMOCAP dataset, our proposed approach achieves a performance comparable to the state of the art speech emotion recognition systems.

Multilingual Speech Emotion Recognition System Based on a Three-Layer Model

Xingfeng Li, Masato Akagi; JAIST, Japan

Mon-O-10-4-4, Time: 14:30

Speech Emotion Recognition (SER) systems currently are focusing on classifying emotions on each single language. Since optimal acoustic sets are strongly language dependent, to achieve a generalized SER system working for multiple languages, issues of selection of common features and retraining are still challenging. In this paper, we therefore present a SER system in a multilingual scenario from perspective of human perceptual processing. The goal is twofold. Firstly, to predict multilingual emotion dimensions accurately such as human annotations. To this end, a three layered model consist of acoustic features, semantic primitives, emotion dimensions, along with Fuzzy Inference System (FIS) were studied. Secondly, by knowledge of human perception of emotion among languages in dimensional space, we adopt direction and distance as common features to detect multilingual emotions. Results of estimation performance of emotion dimensions comparable to human evaluation is furnished, and classification rates that are close to monolingual SER system performed are achieved.

Analysis of Multi-Lingual Emotion Recognition Using Auditory Attention Features

Ozlem Kalinli; Sony, USA

Mon-O-10-4-5, Time: 14:50

In this paper, we build mono-lingual and cross-lingual emotion recognition systems and report performance on English and German databases. The emotion recognition system uses biologically inspired auditory attention features together with a neural network for learning the mapping between features and emotion classes. We first build mono-lingual systems for both Berlin Database of Emotional Speech (EMO-DB) and LDC's Emotional Prosody (Emo-Prosody) and achieve 82.7% and 56.7% accuracy for five class emotion classification (neutral, sad, angry, happy, and boredom) using leave-one-speaker-out cross validation. When tested with cross-lingual systems, the five-class emotion recognition accuracy drops to 55.1% and 41.4% accuracy for EMO-DB and Emo-Prosody, respectively. Finally, we build a bilingual emotion recognition system and report experimental results and their analysis. Bilingual system performs close to the performance of individual mono-lingual systems.

On the Correlation and Transferability of Features Between Automatic Speech Recognition and Speech Emotion Recognition

Haytham M. Fayek, Margaret Lech, Lawrence Cavedon; RMIT University, Australia

Mon-O-10-4-6, Time: 15:10

The correlation between Automatic Speech Recognition (ASR) and Speech Emotion Recognition (SER) is poorly understood. Studying such correlation may pave the way for integrating both tasks into a single system or may provide insights that can aid in advancing both systems such as improving ASR in dealing with emotional speech or embedding linguistic input into SER. In this paper, we quantify the relation between ASR and SER by studying the relevance of features learned between both tasks in deep convolutional neural networks using transfer learning. Experiments are conducted using the TIMIT and IEMOCAP databases. Results reveal an intriguing

NOTES

correlation between both tasks, where features learned in some layers particularly towards initial layers of the network for either task were found to be applicable to the other task with varying degree.

Mon-O-10-5 : Speaker Recognition

Seacliff BCD, 13:30–15:30, Monday, 12 Sept. 2016
Chairs: John Hansen, Koichi Shinoda

On the Influence of Text Content on Pass-Phrase Strength for Short-Duration Text-Dependent Automatic Speaker Authentication

Giacomo Valenti[1], Adrien Daniel[1], Nicholas Evans[2]; [1]NXP Software, France; [2]EURECOM, France

Mon-O-10-5-1, Time: 13:30

In the context of automatic speaker verification it is well known that different speech units offer different levels of speaker discrimination. For short-duration, text-dependent automatic speaker recognition, a user's pass-phrase bears influence on how reliably they can be recognized; just as is the case with text passwords, some spoken pass-phrases are more secure than others. This paper investigates the influence of text or phone content on recognition performance. This work is performed using the shortest duration subset of the standard RSR2015 database. With a thorough statistical analysis, the work shows how significant reductions in error rates can be achieved by preventing the use of weak passwords and that improvements in performance are consistent across disjoint speaker subsets. The ultimate goal of this work is to develop an automated means of enforcing the use of stronger or more discriminant spoken pass-phrases.

Articulation Rate Filtering of CQCC Features for Automatic Speaker Verification

Massimiliano Todisco, Héctor Delgado, Nicholas Evans; EURECOM, France

Mon-O-10-5-2, Time: 13:50

This paper introduces a new articulation rate filter and reports its combination with recently proposed constant Q cepstral coefficients (CQCCs) in their first application to automatic speaker verification (ASV). CQCC features are extracted with the constant Q transform (CQT), a perceptually-inspired alternative to Fourier-based approaches to time-frequency analysis. The CQT offers greater frequency resolution at lower frequencies and greater time resolution at higher frequencies. When coupled with cepstral analysis and the new articulation rate filter, the resulting CQCC features are readily modelled using conventional techniques. A comparative assessment of CQCCs and mel frequency cepstral coefficients (MFCC) for a short-duration speaker verification scenario shows that CQCCs generally outperform MFCCs and that the two feature representations are highly complementary; fusion experiments with the RSR2015 and RedDots databases show relative reductions in equal error rates of as much as 60% compared to an MFCC baseline.

The IBM Speaker Recognition System: Recent Advances and Error Analysis

Seyed Omid Sadjadi[1], Jason W. Pelecanos[1], Sriram Ganapathy[2]; [1]IBM, USA; [2]Indian Institute of Science, India

Mon-O-10-5-3, Time: 14:10

We present the recent advances along with an error analysis of the IBM speaker recognition system for conversational speech. Some of the key advancements that contribute to our system include: a nearest-neighbor discriminant analysis (NDA) approach (as opposed to LDA) for intersession variability compensation in the i-vector space, the application of speaker and channel-adapted features derived from an automatic speech recognition (ASR) system for speaker recognition, and the use of a DNN acoustic model with a very large number of output units (~10k senones) to compute the frame-level soft alignments required in the i-vector estimation process. We evaluate these techniques on the NIST 2010 SRE extended core conditions (C1–C9), as well as the *10sec–10sec* condition. To our knowledge, results achieved by our system represent the best performances published to date on these conditions. For example, on the extended *tel-tel* condition (C5) the system achieves an EER of 0.59%. To garner further understanding of the remaining errors (on C5), we examine the recordings associated with the low scoring target trials, where various issues are identified for the problematic recordings/trials. Interestingly, it is observed that correcting the pathological recordings not only improves the scores for the target trials but also for the non-target trials.

Probabilistic Approach Using Joint Clean and Noisy i-Vectors Modeling for Speaker Recognition

Waad Ben Kheder, Driss Matrouf, Moez Ajili, Jean-François Bonastre; LIA, France

Mon-O-10-5-4, Time: 14:30

Additive noise is one of the main challenges for automatic speaker recognition and several compensation techniques have been proposed to deal with this problem. In this paper, we present a new "data-driven" denoising technique operating in the i-vector space based on a joint modeling of clean and noisy i-vectors. The joint distribution is estimated using a large set of i-vectors pairs (clean i-vectors and their noisy versions generated artificially) then integrated in an MMSE estimator in the test phase to compute a "cleaned-up" version of noisy test i-vectors. We show that this algorithm achieves up to 80% of relative improvement in EER. We also present a version of the proposed algorithm that can be used to compensate multiple "unseen" noises. We test this technique on the recently published SITW database and show a significant gain compared to the baseline system performance.

Generalized Discriminant Analysis (GDA) for Improved i-Vector Based Speaker Recognition

Fahimeh Bahmaninezhad, John H.L. Hansen; University of Texas at Dallas, USA

Mon-O-10-5-5, Time: 14:50

In general, the majority of recent speaker recognition systems employ an i-Vector configuration as their front-end. Post-processing of i-Vectors usually requires a Linear Discriminant Analysis (LDA) phase to reduce the dimensions of the i-Vectors as well as improve discrimination of speaker classes based on the Fisher criterion.

NOTES

Given that channel, noise, and other types of mismatch are generally present in the data, it is better to discriminate the speaker's data non-linearly. Generalized Discriminant Analysis (GDA) uses kernel functions to map the data into a high dimensional feature-space which leads to non-linear discriminant analysis. In this study, we replace LDA with GDA in an i-Vector based speaker recognition system and study the effectiveness of various kernel functions. It is shown, based on equal error rate (EER) and minimum of detection cost function, that GDA not only improves performance for regular test utterances, but is also useful for short duration test segments. NIST2010 Speaker Recognition Evaluation (SRE) core and extended-core (coreext) conditions are employed for experiments; in addition, we evaluate the system for short duration segments on the 10-sec test condition and truncated coreext test data. The relative improvement in EER is 20% for the cosine kernel employed here with GDA processing.

Noise and Metadata Sensitive Bottleneck Features for Improving Speaker Recognition with Non-Native Speech Input

Yao Qian, Jidong Tao, David Suendermann-Oeft, Keelan Evanini, Alexei V. Ivanov, Vikram Ramanarayanan; Educational Testing Service, USA

Mon-O-10-5-6, Time: 15:10

Recently, text independent speaker recognition systems with phonetically-aware DNNs, which allow the comparison among different speakers with "soft-aligned" phonetic content, have significantly outperformed standard i-vector based systems [9–12]. However, when applied to speaker recognition on a non-native spontaneous corpus, DNN-based speaker recognition does not show its superior performance due to the relatively lower accuracy of phonetic content recognition. In this paper, noise-aware features and multi-task learning are investigated to improve the alignment of speech feature frames into the sub-phonemic "senone" space and to "distill" the L1 (native language) information of the test takers into bottleneck features (BNFs), which we refer to as metadata sensitive BNFs. Experimental results show that the system with metadata sensitive BNFs can improve speaker recognition performance by a 23.9% relative reduction in equal error rate (EER) compared to the baseline i-vector system. In addition, L1 info is just used to train the BNFs extractor, so it is not necessary to be used as input for BNFs extraction, i-vector extraction and scoring for the enrollment and evaluation sets, which can avoid the use of erroneous L1s claimed by imposters.

Mon-O-10-6 : VAD and Audio Events

Seacliff A, 13:30–15:30, Monday, 12 Sept. 2016
Chair: Malcolm Slaney

Robust Audio Event Recognition with 1-Max Pooling Convolutional Neural Networks

Huy Phan, Lars Hertel, Marco Maass, Alfred Mertins; Universität zu Lübeck, Germany

Mon-O-10-6-1, Time: 13:30

We present in this paper a simple, yet efficient convolutional neural network (CNN) architecture for robust audio event recognition. Opposing to deep CNN architectures with multiple convolutional and pooling layers topped up with multiple fully connected layers, the proposed network consists of only three layers: convolutional, pooling, and softmax layer. Two further features distinguish it from the deep architectures that have been proposed for the task: varying-size convolutional filters at the convolutional layer and 1-max pooling scheme at the pooling layer. In intuition, the network tends to select the most discriminative features from the whole audio signals for recognition. Our proposed CNN not only shows state-of-the-art performance on the standard task of robust audio event recognition but also outperforms other deep architectures up to 4.5% in terms of recognition accuracy, which is equivalent to 76.3% relative error reduction.

Audio-Based Distributional Representations of Meaning Using a Fusion of Feature Encodings

Giannis Karamanolakis [1], Elias Iosif [1], Athanasia Zlatintsi [1], Aggelos Pikrakis [2], Alexandros Potamianos [1]; [1] NTUA, Greece; [2] University of Piraeus, Greece

Mon-O-10-6-2, Time: 13:50

Recently a "Bag-of-Audio-Words" approach was proposed [1] for the combination of lexical features with audio clips in a multimodal semantic representation, i.e., an Audio Distributional Semantic Model (ADSM). An important step towards the creation of ADSMs is the estimation of the semantic distance between clips in the acoustic space, which is especially challenging given the diversity of audio collections. In this work, we investigate the use of different feature encodings in order to address this challenge following a two-step approach. First, an audio clip is categorized with respect to three classes, namely, music, speech and other. Next, the feature encodings are fused according to the posterior probabilities estimated in the previous step. Using a collection of audio clips annotated with tags we derive a mapping between words and audio clips. Based on this mapping and the proposed audio semantic distance, we construct an ADSM model in order to compute the distance between words (lexical semantic similarity task). The proposed model is shown to significantly outperform (23.6% relative improvement in correlation coefficient) the state-of-the-art results reported in the literature.

Robust DNN-Based VAD Augmented with Phone Entropy Based Rejection of Background Speech

Yuya Fujita, Ken-ichi Iso; Yahoo! JAPAN, Japan

Mon-O-10-6-3, Time: 14:10

We propose a DNN-based voice activity detector augmented by entropy based frame rejection. DNN-based VAD classifies a frame into speech or non-speech and achieves significantly higher VAD performance compared to conventional statistical model-based VAD. We observed that many of the remaining errors are false alarms caused by background human speech, such as TV/radio or surrounding peoples' conversations. In order to reject such background speech frames, we introduce an entropy-based confidence measure using the phone posterior probability output by a DNN-based acoustic model. Compared to the target speaker's voice background speech tends to have relatively unclear pronunciation or is contaminated by other types of noises so its entropy becomes larger than audio signals with only the target speaker's voice. Combining DNN-based VAD and the entropy criterion, we reject speech frames classified by the DNN-based VAD as having an entropy larger than a threshold value. We have evaluated the proposed approach and confirmed greater than 10% reduction in Sentence Error Rate.

Notes

Feature Learning with Raw-Waveform CLDNNs for Voice Activity Detection

Ruben Zazo[1], *Tara N. Sainath*[2], *Gabor Simko*[2], *Carolina Parada*[2]; [1]*Universidad Autónoma de Madrid, Spain;* [2]*Google, USA*

Mon-O-10-6-4, Time: 14:30

Voice Activity Detection (VAD) is an important preprocessing step in any state-of-the-art speech recognition system. Choosing the right set of features and model architecture can be challenging and is an active area of research. In this paper we propose a novel approach to VAD to tackle both feature and model selection jointly. The proposed method is based on a CLDNN (Convolutional, Long Short-Term Memory, Deep Neural Networks) architecture fed directly with the raw waveform. We show that using the raw waveform allows the neural network to learn features directly for the task at hand, which is more powerful than using log-mel features, specially for noisy environments. In addition, using a CLDNN, which takes advantage of both frequency modeling with the CNN and temporal modeling with LSTM, is a much better model for VAD compared to the DNN. The proposed system achieves over 78% relative improvement in False Alarms (FA) at the operating point of 2% False Rejects (FR) on both clean and noisy conditions compared to a DNN of comparable size trained with log-mel features. In addition, we study the impact of the model size and the learned features to provide a better understanding of the proposed architecture.

The SRI System for the NIST OpenSAD 2015 Speech Activity Detection Evaluation

Martin Graciarena[1], *Luciana Ferrer*[2], *Vikramjit Mitra*[1]; [1]*SRI International, USA;* [2]*Universidad de Buenos Aires, Argentina*

Mon-O-10-6-5, Time: 14:50

In this paper, we present the SRI system submission to the NIST OpenSAD 2015 speech activity detection (SAD) evaluation. We present results on three different development databases that we created from the provided data. We present system-development results for feature normalization; for feature fusion with acoustic, voicing, and channel bottleneck features; and finally for SAD bottleneck-feature fusion. We present a novel technique called test adaptive calibration, which is designed to improve decision-threshold selection for each test waveform. We present unsupervised test adaptation of the fusion component and describe its tight synergy to the test adaptive calibration component. Finally, we present results on the evaluation test data and show how the proposed techniques lead to significant gains on channels unseen during training.

Model Adaptation and Active Learning in the BBN Speech Activity Detection System for the DARPA RATS Program

Damianos Karakos, Scott Novotney, Le Zhang, Richard Schwartz; Raytheon BBN Technologies, USA

Mon-O-10-6-6, Time: 15:10

Model adaptation is an important task in many human language technology fields, as it allows one to reduce differences that arise due to various forms of variability. Here, we focus on the speech activity detection (SAD) task, in the context of the DARPA RATS program, where the training data do not cover all channels (transmitter/receiver characteristics) that are encountered at test time. For supervised adaptation, limited manually labeled data from the (novel) channel of interest are used to adapt the model; for unsupervised adaptation, the labels are automatically generated with a baseline model. The modeling is done with long short-term memory neural networks, and we make the case that strong regularization is of paramount importance when adapting such models. Results on two different datasets show that adaptation gives rise to large gains (at least 27related task, that of active learning, is also considered. In active learning, data to be annotated for supervised adaptation are selected automatically, with the ultimate goal of maximizing performance. We investigate an algorithm for active learning that utilizes the output of a SAD decoder and show that it performs significantly better (by 10% relative) than random selection.

Mon-P-10-1 : Spoken Term Detection

Pacific Concourse – Poster A, 13:30–15:30, Monday, 12 Sept. 2016
Chair: Nancy Chen

Fusion Strategies for Robust Speech Recognition and Keyword Spotting for Channel- and Noise-Degraded Speech

Vikramjit Mitra[1], *Julien VanHout*[1], *Wen Wang*[1], *Chris Bartels*[1], *Horacio Franco*[1], *Dimitra Vergyri*[1], *Abeer Alwan*[2], *Adam Janin*[3], *John H.L. Hansen*[4], *Richard M. Stern*[5], *Abhijeet Sangwan*[4], *Nelson Morgan*[3]; [1]*SRI International, USA;* [2]*University of California at Los Angeles, USA;* [3]*ICSI, USA;* [4]*University of Texas at Dallas, USA;* [5]*Carnegie Mellon University, USA*

Mon-P-10-1-1, Time: 13:30

Recognizing speech under high levels of channel and/or noise degradation is challenging. Current state-of-the-art automatic speech recognition systems are sensitive to changing acoustic conditions, which can cause significant performance degradation. Noise-robust acoustic features can improve speech recognition performance under varying background conditions, where it is usually observed that robust modeling techniques and multiple system fusion can help to improve the performance even further. This work investigates a wide array of robust acoustic features that have been previously used to successfully improve speech recognition robustness. We use these features to train individual acoustic models, and we analyze their individual performance. We investigate and report results for simple feature combination, feature-map combination at the output of convolutional layers, and fusion of deep neural nets at the senone posterior level. We report results for speech recognition on a large-vocabulary, noise- and channel-degraded Levantine Arabic speech corpus distributed through the Defense Advance Research Projects Agency (DARPA) Robust Automatic Speech Transcription (RATS) program. In addition, we report keyword spotting results to demonstrate the effect of robust features and multiple levels of information fusion.

NOTES

Recurrent Neural Network-Based Phoneme Sequence Estimation Using Multiple ASR Systems' Outputs for Spoken Term Detection

Naoki Sawada, Hiromitsu Nishizaki; University of Yamanashi, Japan

Mon-P-10-1-2, Time: 13:30

This paper describes a novel correct phoneme sequence estimation method that uses a recurrent neural network (RNN)-based framework for spoken term detection (STD). In an automatic speech recognition (ASR)-based STD framework, ASR performance (word or subword error rate) affects STD performance. Therefore, it is important to reduce ASR errors to obtain good STD results. In this study, we use an RNN-based phoneme estimator, which estimates a correct phoneme sequence of an utterance from some sorts of phoneme-based transcriptions produced by multiple ASR systems in post-processing, to reduce phoneme errors. With two types of test speech corpora, the proposed phoneme estimator obtained phoneme-based N-best transcriptions with fewer phoneme recognition errors than the N-best transcriptions from the best ASR system we prepared. In addition, the STD system with the RNN-based phoneme estimator drastically improved STD performance with two test collections for STD compared to our previously proposed STD system with a conditional random fields-based phoneme estimator.

Enhancing Data-Driven Phone Confusions Using Restricted Recognition

Mark Kane[1], Julie Carson-Berndsen[2]; [1]Daon, Ireland; [2]University College Dublin, Ireland

Mon-P-10-1-3, Time: 13:30

This paper presents a novel approach to address data sparseness in standard confusion matrices and demonstrates how enhanced matrices, which capture additional similarities, can impact the performance of spoken term detection. Using the same training data as for the standard phone confusion matrix, an enhanced confusion matrix is created by iteratively restricting the recognition process to exclude one acoustic model per iteration. Since this results in a greater amount of confusion data for each phone, the enhanced confusion matrix encodes more similarities. The enhanced phone confusion matrices perform demonstrably better than standard confusion matrices on a spoken term detection task which uses both HMMs and DNNs.

Rapid Update of Multilingual Deep Neural Network for Low-Resource Keyword Search

*Chongjia Ni[1], Lei Wang[1], Cheung-Chi Leung[1], Feng Rao[2], Li Lu[2], Bin Ma[1], Haizhou Li[1]; [1]A*STAR, Singapore; [2]Tencent, China*

Mon-P-10-1-4, Time: 13:30

This paper proposes an approach to rapidly update a multilingual deep neural network (DNN) acoustic model for low-resource keyword search (KWS). We use submodular data selection to select a small amount of multilingual data which covers diverse acoustic conditions and is acoustically close to a low-resource target language. The selected multilingual data together with a small amount of the target language data are then used to rapidly update the readily available multilingual DNN. Moreover, the weighted cross-entropy criterion is applied to update the multilingual DNN to obtain the acoustic model for the target language. To verify the proposed approach, experiments were conducted based on four speech corpora (including Cantonese, Pashto, Turkish, and Tagalog) provided by the IARPA Babel program and the OpenKWS14 Tamil corpus. The 3-hour very limited language pack (VLLP) of the Tamil corpus is considered as the target language, while the other four speech corpora are viewed as multilingual sources. Comparing with the traditional cross-lingual transfer approach, the proposed approach achieved a 19% relative improvement in actual term weighted value on the 15-hour evaluation set in the VLLP condition, when a word-based or word-morph mixed language model was used. Furthermore, the proposed approach was observed to have similar performance as the KWS system based on the acoustic model built using the target language and all multilingual data from scratch, but with shorter training time.

Toward High-Performance Language-Independent Query-by-Example Spoken Term Detection for MediaEval 2015: Post-Evaluation Analysis

*Cheung-Chi Leung[1], Lei Wang[1], Haihua Xu[2], Jingyong Hou[3], Van Tung Pham[2], Hang Lv[3], Lei Xie[3], Xiong Xiao[2], Chongjia Ni[1], Bin Ma[1], Eng Siong Chng[2], Haizhou Li[1]; [1]A*STAR, Singapore; [2]NTU, Singapore; [3]Northwestern Polytechnical University, China*

Mon-P-10-1-5, Time: 13:30

This paper documents the significant components of a state-of-the-art language-independent query-by-example spoken term detection system designed for the Query by Example Search on Speech Task (QUESST) in MediaEval 2015. We developed exact and partial matching DTW systems, and WFST based symbolic search systems to handle different types of search queries. To handle the noisy and reverberant speech in the task, we trained tokenizers using data augmented with different noise and reverberation conditions. Our post-evaluation analysis showed that the phone boundary label provided by the improved tokenizers brings more accurate speech activity detection in DTW systems. We argue that acoustic condition mismatch is possibly a more important factor than language mismatch for obtaining consistent gain from stacked bottleneck features. Our post-evaluation system, involving a smaller number of component systems, can outperform our submitted systems, which performed the best for the task.

Mon-P-10-2 : Speech Enhancement and Noise Reduction

Pacific Concourse – Poster B, 13:30–15:30, Monday, 12 Sept. 2016
Chairs: Hynek Hermansky, Hemant Patil

Novel Subband Autoencoder Features for Non-Intrusive Quality Assessment of Noise Suppressed Speech

Meet H. Soni, Hemant A. Patil; DA-IICT, India

Mon-P-10-2-1, Time: 13:30

In this paper, we propose a novel feature extraction architecture of Deep Neural Network (DNN), namely, subband autoencoder (SBAE). The proposed architecture is inspired by the Human Auditory System (HAS) and extracts features from speech spectrum in an

Notes

unsupervised manner. We have used features extracted by this architecture for non-intrusive objective quality assessment of noise suppressed speech signal. The quality assessment problem is posed as a *regression* problem in which mapping between the acoustic features of speech signal and the corresponding subjective score is found using single layer Artificial Neural Network (ANN). We have shown experimentally that proposed features give more powerful mapping than Mel filterbank energies, which are state-of-the-art acoustic features for various speech technology applications. Moreover, proposed method gives more accurate and correlated objective scores than current standard objective quality assessment metric ITU-T P.563. Experiments performed on NOIZEUS database for different test conditions also suggest that objective scores predicted using proposed method are more robust to different amount and types of noise.

SNR-Based Progressive Learning of Deep Neural Network for Speech Enhancement

Tian Gao[1]*, Jun Du*[1]*, Li-Rong Dai*[1]*, Chin-Hui Lee*[2]*;*
[1]*USTC, China;* [2]*Georgia Institute of Technology, USA*

Mon-P-10-2-2, Time: 13:30

In this paper, we propose a novel progressive learning (PL) framework for deep neural network (DNN) based speech enhancement. It aims at decomposing the complicated regression problem of mapping noisy to clean speech into a series of subproblems for enhancing system performances and reducing model complexities. As an illustration, we design a signal-to-noise ratio (SNR) based PL architecture by guiding each hidden layer of the DNN to learn an intermediate target with gradual SNR gains explicitly. Furthermore, post-processing, with the rich set of information from the multiple learning targets, can further be conducted. Experimental results demonstrate that SNR-based progressive learning can effectively improve perceptual evaluation of speech quality and short-time objective intelligibility in low SNR environments, and reduce the model parameters by 50% when compared with the DNN baseline system. Moreover, when combined with post-processing, the proposed approach can be further improved.

A Novel Risk-Estimation-Theoretic Framework for Speech Enhancement in Nonstationary and Non-Gaussian Noise Conditions

Jishnu Sadasivan, Chandra Sekhar Seelamantula;
Indian Institute of Science, India

Mon-P-10-2-3, Time: 13:30

We address the problem of suppressing background noise from noisy speech within a risk estimation framework, where the clean signal is estimated from the noisy observations by minimizing an unbiased estimate of a chosen risk function. For Gaussian noise, such a risk estimate was derived by Stein, which eventually went on to be called Stein's unbiased risk estimate (SURE). Stein's formalism is restricted to Gaussian noise and exclusive risk estimators have been developed for each noise type. On the other hand, we consider linear denoising functions and derive an unbiased risk estimate without making any assumption about the noise distribution. The proposed unbiased estimate depends only on the second-order statistics of noise and makes the proposed framework applicable to many practical denoising problems where the noise distribution is not known a priori, but one has access only to the samples of noise. We demonstrate the usefulness of the proposed methodology for speech enhancement using subband shrinkage, where the shrinkage parameters are obtained by minimizing the newly developed risk estimator. The proposed methodology is also applicable to non-stationary noise conditions. We show that the proposed denoising algorithm outperforms the state-of-the art algorithms in terms of standard speech-quality evaluation metrics.

Two-Stage Temporal Processing for Single-Channel Speech Enhancement

Suman Samui, Indrajit Chakrabarti, Soumya Kanti Ghosh; IIT Kharagpur, India

Mon-P-10-2-4, Time: 13:30

Most of the conventional speech enhancement methods operating in the spectral domain often suffer from spurious artifact called *musical noise*. Moreover, these methods also incur an extra overhead time for noise power spectral density estimation. In this paper, a speech enhancement framework is proposed by cascading two temporal processing stages. The first stage performs excitation source based temporal processing that involves identifying and boosting the excitation source based speech-specific features present at the gross and fine temporal levels, whereas the second stage provides noise reduction by estimating standard deviation of noise in time-domain by using a robust estimator. The proposed noise reduction stage is quite simply implementable and computationally less complex as it does not require noise estimation in spectral domain as a preprocessing phase. The experimental results have established that the proposed scheme produces on an average 60–65% improvement in the speech quality (PESQ scores) and intelligibility (STOI scores) at 0 and -5 dB input SNR when compared to existing standard approaches.

A Class-Specific Speech Enhancement for Phoneme Recognition: A Dictionary Learning Approach

Nazreen P.M., A.G. Ramakrishnan, Prasanta Kumar Ghosh; Indian Institute of Science, India

Mon-P-10-2-5, Time: 13:30

We study the influence of using class-specific dictionaries for enhancement over class-independent dictionary in phoneme recognition of noisy speech. We hypothesize that, using class-specific dictionaries would remove the noise more compared to a class-independent dictionary, thereby resulting in better phoneme recognition. Experiments are performed with speech data from TIMIT corpus and noise samples from NOISEX-92 database. Using KSVD, four types of dictionaries have been learned: class-independent, manner-of-articulation-class, place-of-articulation-class and 39 phoneme-class. Initially, a set of labels are obtained by recognizing the speech, enhanced using a class-independent dictionary. Using these approximate labels, the corresponding class-specific dictionaries are used to enhance each frame of the original noisy speech, and this enhanced speech is then recognized. Compared to the results obtained using the class-independent dictionary, the 39 phoneme-class based dictionaries provide a relative phoneme recognition accuracy improvement of 5.5%, 3.7%, 2.4% and 2.2%, respectively for factory2, m109, leopard and babble noises, when averaged over 0, 5 and 10 dB SNRs.

NOTES

Robust Example Search Using Bottleneck Features for Example-Based Speech Enhancement

Atsunori Ogawa [1], Shogo Seki [1], Keisuke Kinoshita [1], Marc Delcroix [1], Takuya Yoshioka [1], Tomohiro Nakatani [1], Kazuya Takeda [2]; [1] NTT, Japan; [2] Nagoya University, Japan

Mon-P-10-2-6, Time: 13:30

Example-based speech enhancement is a promising approach for coping with highly non-stationary noise. Given a noisy speech input, it first searches in noisy speech corpora for the noisy speech examples that best match the input. Then, it concatenates the clean speech examples that are paired with the matched noisy examples to obtain an estimate of the underlying clean speech component in the input. This framework works well if the noisy speech corpora contain the noise included in the input. However, it is impossible to prepare corpora that cover all types of noisy environments. Moreover, the example search is usually performed using noise sensitive mel-frequency cepstral coefficient features (MFCCs). Consequently, a mismatch between an input and the corpora is inevitable. This paper proposes using bottleneck features (BNFs) extracted from a deep neural network (DNN) acoustic model for the example search. Since BNFs have good noise robustness (invariance), the mismatch is mitigated and thus a more accurate example search can be performed. Experimental results on the Aurora4 corpus show that the example-based approach using BNFs greatly improves the enhanced speech quality compared with that using MFCCs. It also consistently outperforms a conventional DNN-based approach, i.e. a denoising autoencoder.

Speech Enhancement in Multiple-Noise Conditions Using Deep Neural Networks

Anurag Kumar [1], Dinei Florencio [2]; [1] Carnegie Mellon University, USA; [2] Microsoft, USA

Mon-P-10-2-7, Time: 13:30

In this paper we consider the problem of speech enhancement in real-world like conditions where multiple noises can simultaneously corrupt speech. Most of the current literature on speech enhancement focus primarily on presence of single noise in corrupted speech which is far from real-world environments. Specifically, we deal with improving speech quality in office environment where multiple stationary as well as non-stationary noises can be simultaneously present in speech. We propose several strategies based on Deep Neural Networks (DNN) for speech enhancement in these scenarios. We also investigate a DNN training strategy based on psychoacoustic models from speech coding for enhancement of noisy speech.

Perception Optimized Deep Denoising AutoEncoders for Speech Enhancement

Prashanth Gurunath Shivakumar, Panayiotis Georgiou; University of Southern California, USA

Mon-P-10-2-8, Time: 13:30

Speech Enhancement is a challenging and important area of research due to the many applications that depend on improved signal quality. It is a pre-processing step of speech processing systems and used for perceptually improving quality of speech for humans. With recent advances in Deep Neural Networks (DNN), deep Denoising Auto-Encoders have proved to be very successful for speech enhancement. In this paper, we propose a novel objective loss function, which takes into account the perceptual quality of speech. We use that to train Perceptually-Optimized Speech Denoising Auto-Encoders (POS-DAE). We demonstrate the effectiveness of POS-DAE in a speech enhancement task. Further we introduce a two level DNN architecture for denoising and enhancement. We show the effectiveness of the proposed methods for a high noise subset of the QUT-NOISE-TIMIT database under mismatched noise conditions. Experiments are conducted comparing the POS-DAE against the Mean Square Error loss function using speech distortion, noise reduction and Perceptual Evaluation of Speech Quality. We find that the proposed loss function and the new 2-stage architecture give significant improvements in perceptual speech quality measures and the improvements become more significant for higher noise conditions.

HMM-Based Speech Enhancement Using Sub-Word Models and Noise Adaptation

Akihiro Kato, Ben Milner; University of East Anglia, UK

Mon-P-10-2-9, Time: 13:30

This work proposes a method of speech enhancement that uses a network of HMMs to first decode noisy speech and to then synthesise a set of features that enables a clean speech signal to be reconstructed. Different choices of acoustic model (whole-word, monophone and triphone) and grammars (highly constrained to no constraints) are considered and the effects of introducing or relaxing acoustic and grammar constraints investigated. For robust operation in noisy conditions it is necessary for the HMMs to model noisy speech and consequently noise adaptation is investigated along with its effect on the reconstructed speech. Speech quality and intelligibility analysis find triphone models with no grammar, combined with noise adaptation, gives highest performance that outperforms conventional methods of enhancement at low signal-to-noise ratios.

Semi-Supervised Joint Enhancement of Spectral and Cepstral Sequences of Noisy Speech

Li Li [1], Hirokazu Kameoka [1], Takuya Higuchi [2], Hiroshi Saruwatari [1]; [1] University of Tokyo, Japan; [2] NTT, Japan

Mon-P-10-2-10, Time: 13:30

While spectral domain speech enhancement algorithms using non-negative matrix factorization (NMF) are powerful in terms of signal recovery accuracy (e.g., signal-to-noise ratio), they do not necessarily lead to an improvement in the quality of the enhanced speech in the feature domain. This implies that naively using these algorithms as front-end processing for e.g., speech recognition and speech conversion does not always lead to satisfactory results. To address this problem, this paper proposes a novel method that aims to jointly enhance the spectral and cepstral sequences of noisy speech, by optimizing a combined objective function consisting of an NMF-based model-fitting criterion defined in the spectral domain and a Gaussian mixture model (GMM)-based probability distribution defined in the cepstral domain. We derive a novel majorizer for this objective function, which allows us to derive a convergence-guaranteed iterative algorithm based on a majorization-minimization scheme for the optimization. Experimental results revealed that the proposed method outperformed the conventional NMF approach in terms of both signal-to-distortion ratio and cepstral distance.

NOTES

A priori SNR Estimation Using a Generalized Decision Directed Approach

Aleksej Chinaev, Reinhold Haeb-Umbach; Universität Paderborn, Germany

Mon-P-10-2-11, Time: 13:30

In this contribution we investigate *a priori* signal-to-noise ratio (SNR) estimation, a crucial component of a single-channel speech enhancement system based on spectral subtraction. The majority of the state-of-the art *a priori* SNR estimators work in the power spectral domain, which is, however, not confirmed to be the optimal domain for the estimation. Motivated by the generalized spectral subtraction rule, we show how the estimation of the *a priori* SNR can be formulated in the so called generalized SNR domain. This formulation allows to generalize the widely used decision directed (DD) approach. An experimental investigation with different noise types reveals the superiority of the generalized DD approach over the conventional DD approach in terms of both the mean opinion score — listening quality objective measure and the output global SNR in the medium to high input SNR regime, while we show that the power spectrum is the optimal domain for low SNR. We further develop a parameterization which adjusts the domain of estimation automatically according to the estimated input global SNR.

A DNN-HMM Approach to Non-Negative Matrix Factorization Based Speech Enhancement

Ziteng Wang, Xu Li, Xiaofei Wang, Qiang Fu, Yonghong Yan; Chinese Academy of Sciences, China

Mon-P-10-2-12, Time: 13:30

General speaker-independent models have been used in non-negative matrix factorization (NMF) based speech enhancement algorithms for the practical applicability. And additional regulation is necessary when choosing the optimal models for speech reconstruction. In this paper, we propose a novel utilization of deep neural network (DNN) to select the models used for separating speech from noise. Specifically, multiple local dictionaries are learned, whereas only one is activated for each block in the separation step. Besides, the temporal dependencies between blocks are represented by hidden Markov model (HMM), with which it turns out a hybrid DNN-HMM framework. The most probable activation sequence is then solved by the Viterbi algorithm. Experimental evaluations which focus on a speech denoising application are carried out. The results confirm that our proposed approach achieves better performance when compared with some existing methods.

SNR-Aware Convolutional Neural Network Modeling for Speech Enhancement

Szu-Wei Fu[1], Yu Tsao[1], Xugang Lu[2]; [1]Academia Sinica, Taiwan; [2]NICT, Japan

Mon-P-10-2-13, Time: 13:30

This paper proposes a signal-to-noise-ratio (SNR) aware convolutional neural network (CNN) model for speech enhancement (SE). Because the CNN model can deal with local temporal-spectral structures of speech signals, it can effectively disentangle the speech and noise signals given the noisy speech signals. In order to enhance the generalization capability and accuracy, we propose two SNR-aware algorithms for CNN modeling. The first algorithm employs a multi-task learning (MTL) framework, in which restoring clean speech and estimating SNR level are formulated as the main and the secondary tasks, respectively, given the noisy speech input. The second algorithm is an SNR adaptive denoising, in which the SNR level is explicitly predicted in the first step, and then an SNR-dependent CNN model is selected for denoising. Experiments were carried out to test the two SNR-aware algorithms for CNN modeling. Results demonstrate that CNN with the two proposed SNR-aware algorithms outperform the deep neural network counterpart in terms of standardized objective evaluations when using the same number of layers and nodes. Moreover, the SNR-aware algorithms can improve the denoising performance with unseen SNR levels, suggesting their promising generalization capability for real-world applications.

An Iterative Phase Recovery Framework with Phase Mask for Spectral Mapping with an Application to Speech Enhancement

Kehuang Li[1], Bo Wu[2], Chin-Hui Lee[1]; [1]Georgia Institute of Technology, USA; [2]Xidian University, China

Mon-P-10-2-14, Time: 13:30

We propose an iterative phase recovery framework to improve spectral mapping with an application to improving the performance of state-of-the-art speech enhancement systems using magnitude-based spectral mapping with deep neural networks (DNNs). We further propose to use an estimated time-frequency mask to reduce sign uncertainty in the overlap-add waveform reconstruction algorithm. In a series of enhancement experiments using a DNN baseline system, by directly replacing the original phase of noisy speech with the estimated phase obtained with a classical phase recovery algorithm, the proposed iterative technique reduces the log-spectral distortion (LSD) by 0.41 dB from the DNN baseline, and increases the perceptual evaluation speech quality (PESQ) by 0.05 over the DNN baseline, averaging over a wide range of signal and noise conditions. The proposed phase mask mechanism further increases the segmental signal-to-noise ratio (SegSNR) by 0.44 dB at an expense of a slight degradation in LSD and PESQ comparing with the algorithm without using any phase mask.

A Novel Research to Artificial Bandwidth Extension Based on Deep BLSTM Recurrent Neural Networks and Exemplar-Based Sparse Representation

Bin Liu, Jianhua Tao; Chinese Academy of Sciences, China

Mon-P-10-2-15, Time: 13:30

This paper presents a two stages artificial bandwidth extension (ABE) framework which combine deep bidirectional Long Short Term Memory (BLSTM) recurrent neural network with exemplar-based sparse representation to estimate missing frequency band. It demonstrates the suitability of proposed method for modeling log power spectra of speech signals in ABE. The BLSTM-RNN which can capture information from anywhere in the feature sequence is used to estimate the log power spectra in the high-band firstly and the exemplar-based sparse representation which could alleviate the over-smoothing problem is applied to generated log power spectra in the second stage. In addition, rich acoustic features in the low-band are considered to reduce the reconstruction error. Experimental results demonstrate that the proposed framework can achieve significant improvements in both objective and subjective measures over the different baseline methods.

NOTES

Mon-P-10-3 : Far-Field, Robustness and Adaptation

Pacific Concourse – Poster C, 13:30–15:30, Monday, 12 Sept. 2016
Chair: Shinji Watanabe

Coping with Unseen Data Conditions: Investigating Neural Net Architectures, Robust Features, and Information Fusion for Robust Speech Recognition

Vikramjit Mitra, Horacio Franco; SRI International, USA

Mon-P-10-3-1, Time: 13:30

The introduction of deep neural networks has significantly improved automatic speech recognition performance. For real-world use, automatic speech recognition systems must cope with varying background conditions and unseen acoustic data. This work investigates the performance of traditional deep neural networks under varying acoustic conditions and evaluates their performance with speech recorded under realistic background conditions that are mismatched with respect to the training data. We explore using robust acoustic features, articulatory features, and traditional baseline features against both in-domain microphone channel-matched and channel-mismatched conditions as well as out-of-domain data recorded using far- and near-microphone setups containing both background noise and reverberation distortions. We investigate feature-combination techniques, both outside and inside the neural network, and explore neural-network-level combination at the output decision level. Results from this study indicate that robust features can significantly improve deep neural network performance under mismatched, noisy conditions, and that using multiple features reduces speech recognition error rates. Further, we observed that fusing multiple feature sets at the convolutional layer feature-map level was more effective than performing fusion at the input feature level or at the neural-network output decision level.

On the Use of Gaussian Mixture Model Framework to Improve Speaker Adaptation of Deep Neural Network Acoustic Models

Natalia Tomashenko [1], Yuri Khokhlov [2], Yannick Estève [1]; [1] LIUM, France; [2] STC-innovations, Russia

Mon-P-10-3-2, Time: 13:30

In this paper we investigate the Gaussian Mixture Model (GMM) framework for adaptation of context-dependent deep neural network HMM (CD-DNN-HMM) acoustic models. In the previous work an initial attempt was introduced for efficient transfer of adaptation algorithms from the GMM framework to DNN models. In this work we present an extension, further detailed exploration and analysis of the method with respect to state-of-the-art speech recognition DNN setup and propose various novel ways for adaptation performance improvement, such as, using bottleneck features for GMM-derived feature extraction, combination of GMM-derived with conventional features at different levels of DNN architecture, moving from monophones to triphones in the auxiliary GMM model in order to extend the number of adapted classes, and finally, using lattice-based information and confidence scores in maximum a posteriori adaptation of the auxiliary GMM model. Experimental results on the TED-LIUM corpus show that the proposed adaptation technique can be effectively integrated into DNN setup at different levels and provide additional gain in recognition performance.

Analytical Assessment of Dual-Stream Merging for Noise-Robust ASR

Louis ten Bosch [1], Bert Cranen [1], Yang Sun [2]; [1] Radboud Universiteit Nijmegen, The Netherlands; [2] Nuance Communications, Germany

Mon-P-10-3-3, Time: 13:30

In previous studies (on Aurora2), it was found that merging a posteriori probability streams from different classifiers (GMM, MLP, Sparse Coding) can improve the noise robustness of ASR. Maximizing word accuracy required the stream weights to be systematically dependent on the specific input streams and SNR. The tuning of the weights, however, was largely a matter of trial and error and typically involved a laborious grid search. In this paper, we propose two fundamental, analytical methods to better understand these empirical findings. To that end, we maximize the trustworthiness of merged streams as function of the stream weights. Trustworthiness is defined as the probability that the winning state in a probability vector correctly predicts a golden reference state obtained by a forced alignment. Even though our approach is not directly equivalent to optimizing word accuracy, both methods appear highly useful to obtain insight in stream properties that determine the success of a given merge (or the lack thereof). Furthermore, both methods clearly support the trends that exist in the grid-search based empirical observations.

Use of Generalised Nonlinearity in Vector Taylor Series Noise Compensation for Robust Speech Recognition

Erfan Loweimi, Jon Barker, Thomas Hain; University of Sheffield, UK

Mon-P-10-3-4, Time: 13:30

Designing good normalisation to counter the effect of environmental distortions is one of the major challenges for automatic speech recognition (ASR). The Vector Taylor series (VTS) method is a powerful and mathematically well principled technique that can be applied to both the feature and model domains to compensate for both additive and convolutional noises. One of the limitations of this approach, however, is that it is tied to MFCC (and log-filterbank) features and does not extend to other representations such as PLP, PNCC and phase-based front-ends that use power transformation rather than log compression. This paper aims at broadening the scope of the VTS method by deriving a new formulation that assumes a power transformation is used as the non-linearity during feature extraction. It is shown that the conventional VTS, in the log domain, is a special case of the new extended framework. In addition, the new formulation introduces one more degree of freedom which makes it possible to tune the algorithm to better fit the data to the statistical requirements of the ASR back-end. Compared with MFCC and conventional VTS, the proposed approach provides up to 12.2% and 2.0% absolute performance improvements on average, in Aurora-4 tasks, respectively.

Joint Optimization of Denoising Autoencoder and DNN Acoustic Model Based on Multi-Target Learning for Noisy Speech Recognition

Masato Mimura, Shinsuke Sakai, Tatsuya Kawahara; Kyoto University, Japan

Mon-P-10-3-5, Time: 13:30

Denoising autoencoders (DAEs) have been investigated for enhancing noisy speech before feeding it to the back-end deep neural network

NOTES

(DNN) acoustic model, but there may be a mismatch between the DAE output and the expected input of the back-end DNN, and also inconsistency between the training objective functions of the two networks. In this paper, a joint optimization method of the front-end DAE and the back-end DNN is proposed based on a multi-target learning scheme. In the first step, the front-end DAE is trained with an additional target of minimizing the errors propagated by the back-end DNN. Then, the unified network of DAE and DNN is fine-tuned for the phone state classification target, with an extra target of input speech enhancement imposed to the DAE part. The proposed method has been evaluated with the CHiME3 ASR task, and demonstrated to improve the baseline DNN as well as the simple coupling of DAE with DNN. The method is also effective as a post-filter of a beamformer.

Optimization of Speech Enhancement Front-End with Speech Recognition-Level Criterion

Takuya Higuchi, Takuya Yoshioka, Tomohiro Nakatani; NTT, Japan

Mon-P-10-3-6, Time: 13:30

This paper concerns the use of speech enhancement to improve automatic speech recognition (ASR) performance in noisy environments. Speech enhancement systems are usually designed separately from a back-end recognizer by optimizing the front-end parameters with signal-level criteria. Such a disjoint processing approach is not always useful for ASR. Indeed, time-frequency masking, which is widely used in the speech enhancement community, sometimes degrades the ASR performance because of the artifacts created by masking. This paper proposes a speech recognition-oriented front-end approach that optimizes the front-end parameters with an ASR-level criterion, where we use a complex Gaussian mixture model (CGMM) for mask estimation. First, the process of CGMM-based time-frequency masking is reformulated as a computation network. By connecting this CGMM network to the input layer of the acoustic model, the CGMM parameters can be optimized for each test utterance by back propagation using an unsupervised acoustic model adaptation scheme. Experimental results show that the proposed method achieves a relative improvement of 7.7% on the CHiME-3 evaluation set in terms of word error rate.

Factorized Linear Input Network for Acoustic Model Adaptation in Noisy Conditions

Dung T. Tran, Marc Delroix, Atsunori Ogawa, Tomohiro Nakatani; NTT, Japan

Mon-P-10-3-7, Time: 13:30

Deep neural network (DNN) based acoustic models have obtained remarkable performance for many speech recognition tasks. However, recognition performance still remains too low in noisy conditions. To address this issue, a speech enhancement front-end is often used before recognition. Such a front-end can reduce noise but there may remain a mismatch due to the difference in training and testing conditions and the imperfectness of the enhancement front-end. Acoustic model adaptation can be used to mitigate such a mismatch. In this paper, we investigate an extension of the linear input network (LIN) adaptation framework, where the feature transformation is realized as a weighted combination of affine transforms of the enhanced input features. The weights are derived from a vector characterizing the noise conditions. We tested our approach on the real data set of CHiME3 challenge task, confirming the effectiveness of our approach.

Data Augmentation Using Multi-Input Multi-Output Source Separation for Deep Neural Network Based Acoustic Modeling

Yusuke Fujita, Ryoich Takashima, Takeshi Homma, Masahito Togami; Hitachi, Japan

Mon-P-10-3-8, Time: 13:30

We investigate the use of local Gaussian modeling (LGM) based source separation to improve speech recognition accuracy. Previous studies have shown that the LGM based source separation technique has been successfully applied to the runtime speech enhancement and the speech enhancement of training data for deep neural network (DNN) based acoustic modeling. In this paper, we propose a data augmentation method utilizing the multi-input multi-output (MIMO) characteristic of LGM based source separation. We first investigate the difference between unprocessed multi-microphone signals and multi-channel output signals from LGM based source separation as augmented training data for DNN based acoustic modeling. Experimental results using the third CHiME challenge dataset show that the proposed data augmentation outperforms the conventional data augmentation. In addition, we experiment the beamforming applied to the source separated signals as runtime speech enhancement. The results show that the proposed runtime beamforming further improves the speech recognition accuracy.

Microphone Distance Adaptation Using Cluster Adaptive Training for Robust Far Field Speech Recognition

Animesh Prasad, Khe Chai Sim; NUS, Singapore

Mon-P-10-3-9, Time: 13:30

Microphone distance adaptation is an important and challenging problem for far field speech recognition using a single distant microphone. This paper investigates the use of Cluster Adaptive Training (CAT) to learn a structured Deep Neural Network (DNN) that can be quickly adapted to cope with changes in the distance between the microphone and speaker at test time. A speech corpus was created by re-recording the Wall Street Journal (WSJ0) audio using far-field microphones with 8 different distances from the source. Experimental results show that unsupervised adaptation of the CAT-DNN model achieved up to 0.9% absolute word error rate reduction compared to the canonical model trained on multi-style data.

An Investigation on the Use of i-Vectors for Robust ASR

Dimitrios Dimitriadis[1], Samuel Thomas[1], Sriram Ganapathy[2]; [1]IBM, USA; [2]Indian Institute of Science, India

Mon-P-10-3-10, Time: 13:30

In this paper we propose two different i-vector representations that improve the noise robustness of automatic speech recognition (ASR). The first kind of i-vectors is derived from “noise only” components of speech provided by an adaptive denoising algorithm, the second variant is extracted from mel filterbank energies containing both speech and noise. The effectiveness of both these representations is shown by combining them with two different kinds of spectral features — the commonly used log-mel filterbank energies and Teager energy spectral coefficients (TESCs). Using two different DNN

NOTES

architectures for acoustic modeling — a standard state-of-the-art sigmoid-based DNN and an advanced architecture using leaky ReLUs, dropout and rescaling, we demonstrate the benefit of the proposed representations. On the Aurora-4 multi-condition training task the proposed front-end improves ASR performance by 4%.

The Sheffield Wargame Corpus — Day Two and Day Three

Yulan Liu[1], Charles Fox[2], Madina Hasan[1], Thomas Hain[1]; [1]University of Sheffield, UK; [2]University of Leeds, UK

Mon-P-10-3-11, Time: 13:30

Improving the performance of distant speech recognition is of considerable current interest, driven by a desire to bring speech recognition into people's homes. Standard approaches to this task aim to enhance the signal prior to recognition, typically using beamforming techniques on multiple channels. Only few real-world recordings are available that allow experimentation with such techniques. This has become even more pertinent with recent works with deep neural networks aiming to learn beamforming from data. Such approaches require large multi-channel training sets, ideally with location annotation for moving speakers, which is scarce in existing corpora. This paper presents a freely available and new extended corpus of English speech recordings in a natural setting, with moving speakers. The data is recorded with diverse microphone arrays, and uniquely, with ground truth location tracking. It extends the 8.0 hour Sheffield Wargames Corpus released in Interspeech 2013, with a further 16.6 hours of fully annotated data, including 6.1 hours of female speech to improve gender bias. Additional blog-based language model data is provided alongside, as well as a Kaldi baseline system. Results are reported with a standard Kaldi configuration, and a baseline meeting recognition system.

Recurrent Models for Auditory Attention in Multi-Microphone Distant Speech Recognition

Suyoun Kim, Ian Lane; Carnegie Mellon University, USA

Mon-P-10-3-12, Time: 13:30

Integration of multiple microphone data is one of the key ways to achieve robust speech recognition in noisy environments or when the speaker is located at some distance from the input device. Signal processing techniques such as beamforming are widely used to extract a speech signal of interest from background noise. These techniques, however, are highly dependent on prior spatial information about the microphones and the environment in which the system is being used. In this work, we present a neural attention network that directly combines multi-channel audio to generate phonetic states without requiring any prior knowledge of the microphone layout or any explicit signal preprocessing for speech enhancement. We embed an attention mechanism within a Recurrent Neural Network based acoustic model to automatically tune its attention to a more reliable input source. Unlike traditional multi-channel preprocessing, our system can be optimized towards the desired output in one step. Although attention-based models have recently achieved impressive results on sequence-to-sequence learning, no attention mechanisms have previously been applied to learn potentially asynchronous and non-stationary multiple inputs. We evaluate our neural attention model on the CHiME-3 task, and show that the model achieves comparable performance to beamforming using a purely data-driven method.

Semi-Supervised Speaker Adaptation for In-Vehicle Speech Recognition with Deep Neural Networks

Wonkyum Lee[1], Kyu J. Han[2], Ian Lane[1]; [1]Carnegie Mellon University, USA; [2]Ford, USA

Mon-P-10-3-13, Time: 13:30

In this paper, we present a new i-vector based speaker adaptation method for automatic speech recognition with deep neural networks, focusing on in-vehicle scenarios. Our proposed method is, rather than augmenting i-vectors to acoustic feature vectors to form concatenated input vectors for adapting neural network acoustic model parameters, is to perform feature-space transformation with smaller *transformation neural networks* dedicated to acoustic feature vectors and i-vectors, respectively, followed by a layer of *linear combination* of the network outputs. This feature-space transformation is learned via semi-supervised learning without any parameter change in the original deep neural network acoustic model. Experimental results show that our proposed method achieves 18.3% relative improvement in terms of word error rate compared to the speaker independent performance, and verify that it has a potential to replace well-known feature-space Maximum Likelihood Linear Regression (fMLLR) in in-vehicle speech recognition with deep neural networks.

Mon-P-10-4 : Low Resource Speech Recognition

Pacific Concourse – Poster D, 13:30–15:30, Monday, 12 Sept. 2016
Chair: Hung-Yi Lee

Semi-Supervised Training in Deep Learning Acoustic Model

Yan Huang, Yongqiang Wang, Yifan Gong; Microsoft, USA

Mon-P-10-4-1, Time: 13:30

We studied semi-supervised training in a fully connected deep neural network (DNN), unfolded recurrent neural network (RNN), and long short-term memory recurrent neural network (LSTM-RNN) with respect to transcription quality, importance data sampling, and training data amount. We found that DNN, unfolded RNN, and LSTM-RNN exhibit increased sensitivity to labeling errors. One point relative WER increase in the training transcription translates to *a half point* WER increase in DNN and slightly more in unfolded RNN; while in LSTM-RNN it translates to *one full point* WER increase. LSTM-RNN is notably more sensitive to transcription errors. We further found that the importance sampling has similar impact on all three models. In supervised training, importance sampling yields 2∼3% relative WER reduction against random sampling. The gain is reduced in semi-supervised training. Lastly, we compared the model capacity with increased training data. Experimental results suggest that LSTM-RNN can benefit more from enlarged training data comparing to unfolded RNN and DNN.

We trained a semi-supervised LSTM-RNN using 2600 hours of transcribed and 10000 hours of untranscribed data on a mobile speech task. The semi-supervised LSTM-RNN yields 6.56% relative WER reduction against the supervised baseline trained from 2600 hours of transcribed speech.

NOTES

Multilingual Data Selection for Low Resource Speech Recognition

Samuel Thomas, Kartik Audhkhasi, Jia Cui, Brian Kingsbury, Bhuvana Ramabhadran; IBM, USA

Mon-P-10-4-2, Time: 13:30

Feature representations extracted from deep neural network-based multilingual frontends provide significant improvements to speech recognition systems in low resource settings. To effectively train these frontends, we introduce a data selection technique that discovers language groups from an available set of training languages. This data selection method reduces the required amount of training data and training time by approximately 40%, with minimal performance degradation. We present speech recognition results on 7 very limited language pack (VLLP) languages from the second option period of the IARPA Babel program using multilingual features trained on up to 10 languages. The proposed multilingual features provide up to 15% relative improvement over baseline acoustic features on the VLLP languages.

An Investigation on Training Deep Neural Networks Using Probabilistic Transcriptions

Amit Das, Mark Hasegawa-Johnson; University of Illinois at Urbana-Champaign, USA

Mon-P-10-4-3, Time: 13:30

In this study, a transfer learning technique is presented for cross-lingual speech recognition in an adverse scenario where there are no natively transcribed transcriptions in the target language. The transcriptions that are available during training are transcribed by crowd workers who neither speak nor have any familiarity with the target language. Hence, such transcriptions are likely to be inaccurate. Training a deep neural network (DNN) in such a scenario is challenging; previously reported results have described DNN error rates exceeding the error rate of an adapted Gaussian Mixture Model (GMM). This paper investigates multi-task learning techniques using deep neural networks which are suitable for this scenario. We report, for the first time, absolute improvement in phone error rates (PER) in the range 1.3–6.2% over GMMs adapted to probabilistic transcriptions. Results are reported for Swahili, Hungarian, and Mandarin.

Analysis of Mismatched Transcriptions Generated by Humans and Machines for Under-Resourced Languages

*Van Hai Do[1], Nancy F. Chen[2], Boon Pang Lim[2], Mark Hasegawa-Johnson[1]; [1]ADSC, Singapore; [2]A*STAR, Singapore*

Mon-P-10-4-4, Time: 13:30

When speech data with native transcriptions are scarce in an under-resourced language, automatic speech recognition (ASR) must be trained using other methods. Semi-supervised learning first labels the speech using ASR from other languages, then re-trains the ASR using the generated labels. Mismatched crowdsourcing asks crowd-workers unfamiliar with the language to transcribe it. In this paper, self-training and mismatched crowdsourcing are compared under exactly matched conditions. Specifically, speech data of the target language are decoded by the source language ASR systems into source language phone/word sequences. We find that (1) human mismatched crowdsourcing and cross-lingual ASR have similar error patterns, but different specific errors. (2) These two sources of information can be usefully combined in order to train a better target-language ASR. (3) The differences between the error patterns of non-native human listeners and non-native ASR are small, but when differences are observed, they provide information about the relationship between the phoneme systems of the annotator/source language (Mandarin) and the target language (Vietnamese).

ASR for South Slavic Languages Developed in Almost Automated Way

Jan Nouza, Radek Safarik, Petr Cerva; Technical University of Liberec, Czech Republic

Mon-P-10-4-5, Time: 13:30

Slavic languages pose several specific challenges that need to be addressed in an ASR system design. Since we have already built an engine suited for highly-inflected languages, we focus on adopting it for new languages, now. In this case, we present an efficient way to adapt the system to all (seven) South Slavic languages, using methods and tools that benefit from language similarities, easily adjustable G2P rules or common phonetic subsets. We show that it is possible to build accurate language and acoustic models in an almost automated way, entirely from resources found on the web. The AMs are trained via cross-lingual bootstrapping followed by lightly supervised retraining from public data, like broadcast and parliament archives. Tests done on a set of main broadcast news in each language show WER values in range 16.8 to 21.5%, which includes also errors caused by OOL (out-of-language) utterances often occurring in this type of spoken programs.

Improving Under-Resourced Language ASR Through Latent Subword Unit Space Discovery

Marzieh Razavi, Mathew Magimai-Doss; Idiap Research Institute, Switzerland

Mon-P-10-4-6, Time: 13:30

Development of state-of-the-art automatic speech recognition (ASR) systems requires acoustic resources (i.e., transcribed speech) as well as lexical resources (i.e., phonetic lexicons). It has been shown that acoustic and lexical resource constraints can be overcome by first training an acoustic model that captures acoustic-to-multilingual phone relationships on language-independent data; and then training a lexical model that captures grapheme-to-multilingual phone relationships on the target language data. In this paper, we show that such an approach can be employed to discover a latent space of subword units for under-resourced languages, and subsequently improve the performance of the ASR system through both acoustic and lexical model adaptation. Specifically, we present two approaches to discover the latent space: (1) inference of a subset of the multilingual phone set based on the learned grapheme-to-multilingual phone relationships, and (2) derivation of automatic subword unit space based on clustering of the grapheme-to-multilingual phone relationships. Experimental studies on Scottish Gaelic, a truly under-resourced language, show that both approaches lead to significant performance improvements, with the latter approach yielding the best system.

NOTES

Language Adaptive DNNs for Improved Low Resource Speech Recognition

Markus Müller, Sebastian Stüker, Alex Waibel; KIT, Germany

Mon-P-10-4-7, Time: 13:30

Deep Neural Network (DNN) acoustic models are commonly used in today's state-of-the-art speech recognition systems. As neural networks are a data driven method, the amount of available training data directly impacts the performance. In the past, several studies have shown that multilingual training of DNNs leads to improvements, especially in resource constrained tasks in which only limited training data in the target language is available.

Previous studies have shown speaker adaptation to be successfully performed on DNNs. This is achieved by adding speaker information (e.g. i-Vectors) as additional input features. Based on the idea of adding additional features, we here present a method for adding language information to the input features of the network. Preliminary experiments have shown improvements by providing supervised information about language identity to the network.

In this work, we extended this approach by training a neural network to encode language specific features. We extracted those features unsupervised and used them to provide additional cues to the DNN acoustic model during training. Our results show that augmenting acoustic input features with this language code enabled the network to better capture language specific peculiarities. This improved the performance of systems trained using data from multiple languages.

Improved Multilingual Training of Stacked Neural Network Acoustic Models for Low Resource Languages

Tanel Alumäe, Stavros Tsakalidis, Richard Schwartz; Raytheon BBN Technologies, USA

Mon-P-10-4-8, Time: 13:30

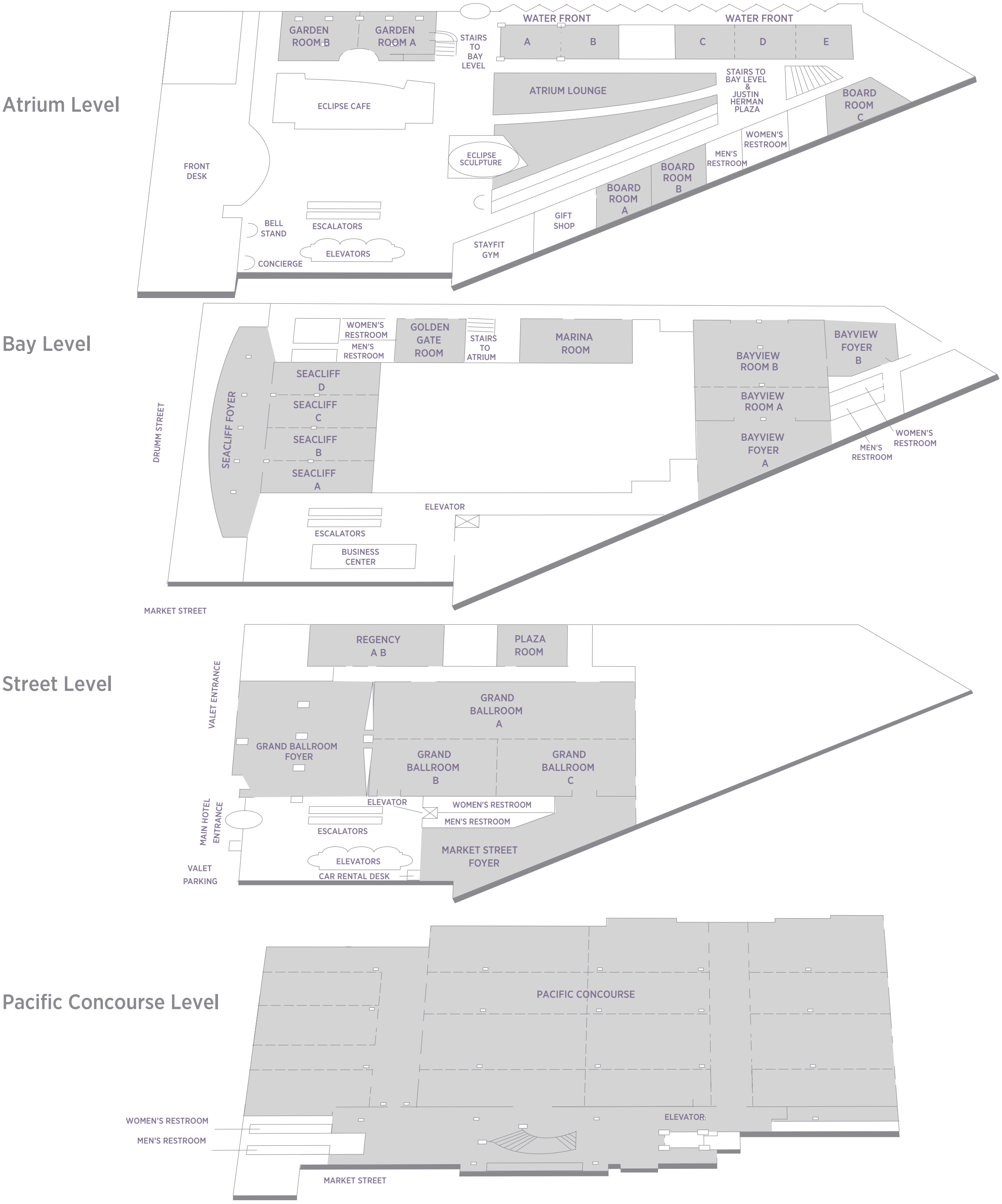
This paper proposes several improvements to multilingual training of neural network acoustic models for speech recognition and keyword spotting in the context of low-resource languages. We concentrate on the stacked architecture where the first network is used as a bottleneck feature extractor and the second network as the acoustic model. We propose to improve multilingual training when the amount of data from different languages is very different by applying balancing scalers to the training examples. We also explore how to exploit multilingual data to train the second neural network of the stacked architecture. An ensemble training method that can take advantage of both unsupervised pretraining as well as multilingual training is found to give the best speech recognition performance across a wide variety of languages, while system combination of differently trained multilingual models results in further improvements in keyword search performance.

Notes

Author Index

NOTES

HYATT REGENCY SAN FRANCISCO VENUE MAP



Atrium Level

Bay Level

Street Level

Pacific Concourse Level

Atrium Lounge	Atrium Level
Bayview A	Bay Level
Bayview B	Bay Level
Board Room ABC	Atrium Level
Garden Room AB	Atrium Level
Golden Gate Room	Bay Level
Grand Ballroom ABC	Street Level
Marina Room	Bay Level
Market Street Foyer	Street Level
Pacific Concourse	Pacific Concourse Level
Plaza Room	Street Level
Regency AB	Street Level
Seacliff ABCD	Bay Level
Waterfront ABCDE	Atrium Level

INTERSPEECH 2016 AGENDA AT A GLANCE

Thursday, 8 September	Friday, 9 September	Saturday, 10 September	Sunday, 11 September	Monday, 12 September
Registration 08:00 - 19:00 | Grand Ballroom Foyer	**Registration 08:00 - 19:30 | Grand Ballroom Foyer**	**Registration | 08:00 - 18:00 | Grand Ballroom Foyer** **Speaker Check-In | 07:30 - 17:00 | Regency AB**	**Registration 08:00 - 18:00 | Grand Ballroom Foyer**	**Registration 08:00 - 17:30 | Grand Ballroom Foyer**
Speaker Check-In 07:30 - 17:00 | Regency AB	**Speaker Check-In 07:30 - 17:00 | Regency AB**	Special Event: Mindfulness 08:00 - 08:30 | Grand Ballroom ABC	**Speaker Check-In 07:30 - 17:00 | Regency AB**	**Speaker Check-In 07:30 - 13:30 | Regency AB**
Morning Tutorials 08:30 - 10:00	Opening Session 08:30 - 09:30 Grand Ballroom ABC	08:30 - 09:30 Keynote 2: Edward Chang Grand Ballroom ABC	08:30 - 09:30 Keynote 3: Anne Fernald Grand Ballroom ABC	08:30 - 09:30 Keynote 4: Dan Jurafsky Grand Ballroom ABC
	Keynote 1: ISCA Medalist: John Makhoul 09:30 - 10:30 Grand Ballroom ABC	Refreshment Break 09:30 - 10:00 | Pacific Concourse	Refreshment Break 09:30 - 10:00 | Pacific Concourse	Refreshment Break 09:30 - 10:00 | Pacific Concourse
Refreshment Break 10:00 - 10:30 | Seacliff Foyer		CONCURRENT SESSIONS | 10:00 - 12:00 Special Event: Speaker Comparison for Forensic and Investigative Applications II Grand Ballroom A Oral 4 Poster 4 Show & Tell 4	CONCURRENT SESSIONS | 10:00 - 12:00 Oral 6 Poster 6 Show & Tell 6	CONCURRENT SESSIONS | 10:00 - 12:00 Special Event: Speech Ventures Grand Ballroom A Oral 9 Poster 9
Morning Tutorials Continued 10:30 - 12:00	Refreshment Break 10:30 - 11:00 | Pacific Concourse CONCURRENT SESSIONS | 11:00 - 13:00 Oral 1 Poster 1 Show & Tell 1			
Lunch Break 12:00 - 13:00		Lunch Break 12:00 - 13:30	Lunch Break 12:00 - 13:30	Special Event: Computational Approaches to Linguistic Code Switching 12:15 - 13:00 Grand Ballroom A Lunch Break 12:00 - 13:30
Afternoon Tutorials 13:00 - 14:30	Lunch Break 13:00 - 14:30	CONCURRENT SESSIONS | 13:30 - 15:30 Oral 5 Poster 5 Show & Tell 5	CONCURRENT SESSIONS | 13:30 - 15:30 Oral 7 Poster 7	CONCURRENT SESSIONS | 13:30 - 15:30 Oral 10 Poster 10
Refreshment Break 14:30 - 15:00 | Seacliff Foyer	CONCURRENT SESSIONS | 14:30 - 16:30 Oral 2 Poster 2 Show & Tell 2	Refreshment Break 15:30 - 16:00 | Pacific Concourse	Refreshment Break 15:30 - 16:00 | Pacific Concourse	Refreshment Break 15:30 - 16:00 | Pacific Concourse
Afternoon Tutorials Continued 15:00 - 16:30	Refreshment Break 16:30 - 17:00 | Pacific Concourse	ISCA General Assembly 16:00 - 17:30 Grand Ballroom A	CONCURRENT SESSIONS | 16:00 - 18:00 Oral 8 Poster 8	Closing Session 16:00 - 17:00 Grand Ballroom ABC
	CONCURRENT SESSIONS | 17:00 - 19:00 Oral 3 Poster 3 Show & Tell 3	Reviewer's Reception 18:30 - 20:00 The City Club		
	Welcome Reception 19:00 - 21:00 Hyatt Regency San Francisco Atrium Lounge	Student Reception 19:00 - 21:00 Jones	Banquet 19:00 - 22:00 California Academy of Sciences (Advance purchase required) *18:30* *Buses depart from Hyatt Regency San Francisco*	