

Odyssey 2008

**The Speaker and Language
Recognition Workshop**

21-24 January 2008

**Stellenbosch
South Africa**



Contents

Welcome Message	2
Scientific Committee	3
Odyssey 2008 Organizing Committee	5
Local Organizing Committee	6
General Information	7
Scholarships	9
Sponsor	9
Scientific Programme Summary	10
Scientific Programme	11
Author Index	27

Welcome Message

Welcome to Odyssey 2008: The Speaker and Language Recognition Workshop. This year hosted in Stellenbosch, South Africa from 21-24 January 2008.

This year's workshop is co-hosted by Spescom DataVoice and the Digital Signal Processing Group of Stellenbosch University.

Odyssey'08 is an ISCA Tutorial and Research Workshop held in cooperation with the ISCA Speaker and Language Characterization SIG and with technical co-sponsorship by the IEEE and the IET.

The need for fast, efficient, accurate, and robust means of recognizing people and languages is of growing importance for commercial, forensic, and government applications. The aim of this workshop is to continue to foster interactions among researchers in speaker and language recognition as the successor of the series of previous workshops:

- 1994 Automatic Speaker Recognition Workshop (Martigny, Switzerland)
- 1998 RLA2C Workshop (Avignon, France)
- 2001 A Speaker Odyssey (Crete, Greece)
- 2004 Odyssey 2004 The Speaker and Language Recognition Workshop (Toledo, Spain)
- 2006 IEEE Odyssey 2006 Workshop on Speaker and Language Recognition. (San Juan, Puerto Rico)

On behalf of the organizing committee of Odyssey 2008 and on behalf of the local support team, we wish you a productive and enjoyable conference.

- Niko Brummer and Johan du Preez



Scientific Committee

Aladdin Ariyaeenia, University of Hertfordshire, UK
Alvin Martin, NIST, USA
Lukas Burget, Brno University of Technology, Czech Republic
Claude Barras, LIMSI, France
Christian Mueller, ICSI, USA
Douglas Reynolds, MIT Lincoln Lab, USA
Daniel Ramos, Universidad Autónoma de Madrid, Spain
David van Leeuwen, TNO, The Netherlands
Johan du Preez, University of Stellenbosch, South Africa
Elizabeth Shriberg, SRI, USA
Mauro Falcone, FUB, Italy
Frederic Bimbot, IRISA, France
Sadaaki Furui, Tokyo Tech, Japan
Haizhou Li, Media Processing Department Institute for Infocomm Research, Singapore
Javier Ortega-Garcia, Universidad Autónoma de Madrid, Spain
Jean-Francois Bonastre, Université d'Avignon, France
Jiri Navratil, IBM, USA
Joaquin Gonzalez-Rodriguez, Universidad Autónoma de Madrid, Spain
Jos Bouten, Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, The Netherlands
Joe Campbell, MIT Lincoln Lab, USA
Jason Pelecanos, IBM, USA
Kay Berkling, Polytechnic University of Puerto Rico, Puerto Rico
Kevin Farrell, Nuance, USA
Laurent Besacier, CLIPS-IMAG, France
Jerome Louradour, IRIT, France
Tina Kohler, DoD, USA
Mats Blomberg, KTH, Sweden
Michael Wagner, University of Canberra, Australia
Niko Brummer, Spescom DataVoice, South Africa
Patrick Kenny, CRIM, Canada
Pietro Laface, Politecnico di Torino, Italy
Pedro Torres-Carrasquillo, MIT Lincoln Lab, USA
Robbie Vogt, Queensland University of Technology, Australia

Ran Gazit,Persay,Israel
Roland Auckenthaler,NMS Communications,USA
Sachin Kajarekar,SRI,USA
Andreas Stolcke,SRI,USA
Doug Sturim,MIT Lincoln Lab,USA
Tim Anderson,AFRL,USA
Tomoko Matsui,Institute of Statistical Mathematics,Japan
Thomas Niesler,University of Stellenbosch,South Africa
Walt Andrews,DoD,USA
Bill Campbell,MIT Lincoln Lab,USA

Odyssey 2008 Organizing Committee

Chair:

Niko Brummer, Spescom DataVoice and Johan du Preez, Digital Signal Processing Group, Stellenbosch University, South Africa.

Pedro A. Torres-Carrasquillo
MIT Lincoln Laboratory, USA

Frédéric Bimbot
IRISA, France

Jean-François Bonastre
University of Avignon, France

Joseph Campbell
MIT Lincoln Laboratory, USA

Joaquín González-Rodríguez
Universidad Autónoma de Madrid, Spain

John Mason
University of Wales Swansea, UK

Javier Ortega-García
Universidad Autónoma de Madrid, Spain

Renana Peres
Tel Aviv University, Israel

Douglas Reynolds
MIT Lincoln Laboratory, USA

Phil Rose
The Australian National University, Australia

Local Organizing Committee

Niko Brummer,
Spescom DataVoice

Johan du Preez and Albert Strasheim,
Digital Signal Processing Group,
Stellenbosh University

CONGRESS SECRETARIAT:

Christelle Snyman
Speaker Odyssey Conference
Unistel ConsultUS
P O Box 19063
Tygerberg
7505
TEL: 021 938 9245
FAX: 021 933 2649
E-MAIL: csnyman@sun.ac.za



General Information

VENUE

The meeting will convene in the Simonsberg Room at the Spier Estate's conference centre.

REGISTRATION AND INFORMATION DESK

The registration and information desk will be situated in the foyer of the conference centre for the duration of the congress. The desk will be open for registration at the following time:

Monday 21 January 2008

Foyer of the Spier Conference Centre

08:30-09:20

On site payment of registration fees can be done by cash, cheque or credit card.

LANGUAGE

The official language of the congress will be English. No translation service will be provided.

PRESENTATION PREVIEW ROOM

The preview room is situated in the Foyer of the Spier conference centre.

All computer presentations must be checked for viruses and reach the technicians in the preview room, at least **one hour** before your presentation.

- Computer presentations must be presentable on Windows Office XP or later. The presentations itself must be pdf or Power Point.

ADMISSION BADGES

Congress badges should be worn by all participants AT ALL TIMES during the congress, whilst visiting the exhibition area and on all congress premises. Badges should also be worn when attending the events of the social programme. Only participants or accompanying persons wearing their congress badges will be admitted to the scientific sessions, the exhibition and the social events.

CREDIT CARDS

Major credit cards (American Express, Diners Club, MasterCard and Visa) are accepted in most hotels, restaurants and shops. Conference registration fees need to be paid in advance, in cash local currency or local cheque. This conference offers credit card facilities for registration purposes only.

Social Program

MONDAY 21 JANUARY 2008

A conference dinner will be held at MOYO at Spier. The cost per person will amount to R350-00. Accompanying persons are also welcome at an additional R350 per person (please register accordingly).

Dress: Casual

WEDNESDAY 23 JANUARY 2008

Wednesday will be allocated for an enjoyable wine tour and leisurely lunch in the wine district. The following itinerary will be followed:

Tour Themes: Wine, scenery, history.

- KWV Wine Emporium
- Stellenbosch University and Historical
- Fairview Cheese & Wine Farm
- Light Lunch at the Goatshed (Fairview)
- 2 Wine Tastings
- Cellar Tour

Departure from Spier at 09h00

Arrival back at Spier by 15h00

Accompanying persons welcome at an additional R350-00 per person (please register accordingly).

Dress: Casual

Scholarships

We are grateful to receive the support of ISCA scholarships for this conference. Please obtain more information by following this link to: [ISCA grants](#).



Sponsor

We are grateful to receive financial sponsorship from IBM to support student participation.

Scientific Program Summary

Monday 21 January 2008

08:30-09:20	Registration	
09:20-09:40	Opening	
09:40-10:50	Invited Talk, by Doug Reynolds	11
10:50-11:20	Tea	
11:20-12:50	Speaker Recognition I: Forensic	11
12:50-14:20	Lunch	
14:20-15:30	Speaker Recognition II: Identification	13
15:30-16:00	Tea	
16:00-17:10	Speaker Recognition III: Intersession Variability	15

Tuesday 22 January 2008

09:00-10:10	Invited Talk, by David van Leeuwen	17
10:10-10:40	Tea	
10:40-12:40	Language Recognition	17
12:50-14:20	Lunch	
14:20-15:30	Speaker Recognition IV	19
15:30-16:00	Tea	
16:00-17:10	Speaker Recognition V	21

Thursday 24 January 2008

09:00-10:40	Speaker Classification	22
10:40-11:10	Tea	
11:10-12:20	Speaker Recognition VI	24
12:50	Lunch	

Not presented

Speaker Recognition VII	26
-------------------------	----

Scientific Program

Monday

09:40-10:50

Invited Talk

Doug Reynolds of MIT Lincoln Lab

10:50-11:20 Tea

Monday

11:20-12:50

Speaker Recognition I: Forensic

Session chair: Joaquin Gonzales-Rodriguez

11:20-11:40 *How vulnerable are prosodic features to professional imitators?*

Mireia Farrus, Michael Wagner, Jan Anguita and Javier Hernando

Voice imitation is one of the potential threats to security systems that use automatic speaker recognition. Since prosodic features have been considered for state-of-the-art recognition systems in recent years, the question arises as to how vulnerable these features are to voice mimicking. In this study, two experiments are conducted for twelve individual features in order to determine how a prosodic speaker identification system would perform against professionally imitated voices. By analysing prosodic parameters, the results show that the identification error rate increases for most of the features, except for the range of the fundamental frequency, which seems to be relatively robust against voice mimicking. When all twelve features are fused, the identification error rate increases from 5% between the target voices and the imitators' natural voices to 22% between the target voices and the imitators' impersonations.

11:40--12:00 *Beyond the Long-term Mean: Exploring the Potential of F0 Distribution Parameters in Traditional Forensic Speaker Recognition.*

Yuko Kinoshita, Shunichi Ishihara and Phil Rose

Despite its many prima facie attractive properties for Forensic Speaker Recognition, F0 is regarded as having limited forensic value due to its large within-speaker variability. However, its forensic use to date has been limited

mostly to its long-term mean and standard deviation. This paper examines the discriminatory potential, within a Likelihood Ratio-based approach, of additional parametric features from the distribution of long-term F0: its skew, kurtosis, modal F0 and modal density. Motivated by the observation that the overall long-term F0 distribution shows less within speaker occasion-to-occasion difference, we report a forensic discrimination experiment with non-contemporaneous speech samples from 201 male experiment with non-contemporaneous speech samples from 201 male Japanese speakers. Using a multivariate LR as discriminant distance with the six LTF0 distribution parameters, an EER of 10.7% is obtained from 201 target and 80400 non-target trials. We also investigate how the EER degrades as a function of amount of voiced speech.

12:00-12:20 *Cross-entropy Analysis of the Information in Forensic Speaker Recognition*

Daniel Ramos and Joaquin Gonzalez-Rodriguez

In this work we analyze the average information supplied by a forensic speaker recognition system in an information-theoretical way. The objective is the transparent reporting of the performance of the system in terms of information, according to the needs of transparency and testability in forensic science. This analysis allows the derivation of a proper measure of goodness for forensic speaker recognition, the empirical cross-entropy (ECE), according to previous work in the literature. We also propose an intuitive representation, namely the ECE-plot, which allows forensic scientists to explain the average information given by the evidence analysis process in a clear and intuitive way. Such representation allows the forensic scientist to assess the evidence evaluation process with independence of the prior information, which is province of the court. Then, fact finders may check the average information given by the evidence analysis with the incorporation of prior information. An experimental example following NIST SRE 2006 protocol is presented in order to highlight the adequacy of the proposed framework in the forensic inferential process. An example of the presentation of the average information supplied by the forensic analysis of the speech evidence in court is also provided, simulating a real case.

12:20--12:40 *Evaluation of speech quality measures for the purpose of speaker verification*

Jonas Richiardi and Andrzej Drygajlo

Real-world deployment of speaker verification systems often have to contend with degraded signal quality and erratic statistical behaviour of the speech data being modelled. We present signal quality estimation

techniques for extraction of additional information about the speech data that can be used to improve performance of speaker verification systems in degraded conditions. We propose methods to perform objective evaluation of these quality measures for the purpose of their comparison using benchmarking databases, and show why the class must be taken into account when evaluating quality measures.

12:40-12:50 *Discussion*

12:50-14:20 *Lunch*

Monday

14:20-15:30 *Speaker Recognition II: Identification*

Session chair: Javier Ortega-Garcia

14:20-14:40 *Feature Vector Classification by Threshold for Speaker identification.*

Sang-min Yoon, Kyung-mi Park, Jae-Hyun Bae and Yung-hwan Oh

This paper describes a new feature vector classification method for speaker identification. Purpose of this paper is constructing robust speaker models which only use meaningful feature vectors and discard confusing feature vectors. To construct robust speaker model, proposed method classifies feature vectors using log-likelihood estimation. Experimental results, with various segments ranging from 0.5 to 5s, showed that our method outperforms previous method.

14:40-15:00 *Improving robustness in open set speaker identification by shallow source modeling.*

M. Zamalloa, L.J. Rodriguez, M. Penagarikano, G. Bodel, J.P. Uribe

Open set speaker identification consists of deciding whether an input utterance corresponds to a target speaker or to an impostor. The most likely among a set of target speakers is hypothesized and verified. Speaker verification is performed by comparing the likelihood score of the most likely speaker model to the likelihood score of an impostor model, and then applying a suitable threshold. The most common approach to modelling impostors is the Universal Background Model (UBM). For the UBM to be effective, it must be estimated from a large number of speakers. However, it is not always possible to gather enough data to estimate a robust UBM, and the

verification performance may degrade if impostors, or whatever sources that generate the input signals, were not suitably modelled by the UBM. In this paper, a simple approach is proposed which estimates a shallow source model (SSM) based on the input utterance, and then uses this SSM to normalize the speaker score. Though the SSM does not outperform the UBM, the combination of both models improves the recognition performance and drastically increases the robustness to signals not covered by the UBM.

15:00-15:20 *Comparison of a Joint Iterative Method for Multiple Speaker Identification with Sequential Blind Source Separation and Speaker Identification.*

Youngmoo E. Kim, John MacLaren Walsh and Travis M. Doll

An individual's voice is hardly ever heard in complete isolation. More commonly, it occurs simultaneously along with other interfering sounds, including those of other overlapping voices. Though there has been a great deal of progress in automatic speaker identification, the majority of past work has focused on the case of non-overlapping speakers. Many of these systems are easily confounded by more realistic scenarios where multiple talkers may be overlapping or speaking simultaneously.

Furthermore, the variations due to different acoustic environments in real-world settings are detrimental to well-known systems that aim to separate the features or the acoustic signal of a mixture of talkers. We propose a system that, given multiple acoustic observations, attempts to jointly identify and separate the acoustic features of multiple simultaneous talkers that fall within a library of known individuals. This system uses the probabilistic framework of expectation propagation (EP) to iteratively determine model-based statistics of both individual acoustic features and speaker identity. In our initial study, we demonstrate that this framework exhibits performance that in the upper-bound significantly exceeds that of a sequential method employing blind source separation followed by speaker identification on the estimated source signals.

15:20-15:30 *Discussion*

15:30-16:00 *Tea*

Monday

16:00-17:10 **Speaker Recognition III: Intersession Variability**

Session chair: Patrick Kenny

16:00-16:20 *Comparison Between Factor Analysis and GMM Support Vector Machines for Speaker Verification*
Najim Dehak, Reda Dehak, Patrick Kenny and Pierre Dumouchel

We present a comparison between speaker verification systems based on factor analysis modeling and support vector machines using GMM supervectors as features. All systems used the same acoustic features and they were trained and tested on the same data sets. We test two types of kernel (one linear, the other non-linear) for the GMM support vector machines. The results show that factor analysis using speaker factors gives the best results on the core condition of the NIST 2006 speaker recognition evaluation. The difference is particularly marked on the English language subset. Fusion of all systems gave an equal error rate of 4.2% (all trials) and 3.2% (English trials only).

16:20-16:40 *Discriminant NAP for SVM Speaker Recognition.*
Robbie Vogt, Sachin Kajarekar, Sridha Sridharan

Nuisance Attribute Projection (NAP) provides an effective method of removing the unwanted session variability in a Support Vector Machine (SVM) based speaker recognition system by removing the principal components of this variability. There is no guarantee with the methods proposed, however, that desired speaker variability is retained. This paper investigates the possibility of training NAP discriminatively to remove session variability while maintaining desirable speaker variability through an approach which is a variation on Scatter Difference Analysis (SDA). Experiments on NIST SRE tasks with a GMM mean supervector SVM system demonstrate a modest improvement by using SDA for NAP training by adding some speaker scatter.

16:40-17:00 *The Role of Speaker Factors in the NIST Extended Data Task.*
Patrick Kenny, Najim Dehak, Reda Dehak, Vishwa Gupta and Pierre Dumouchel

We tested factor analysis models having various numbers of speaker factors on the core condition and the extended data condition of the 2006 NIST speaker recognition evaluation. In order to ensure strict disjointness between training and test sets, the factor analysis models were trained without using

any of the data made available for the 2005 evaluation. The factor analysis training set consisted primarily of Switchboard data and so was to some degree mismatched with the 2006 test data (drawn from the Mixer collection). Consequently, our initial results were not as good as those submitted for the 2006 evaluation. However we found that we could compensate for this by a simple modification to our score normalization strategy, namely by using 1000 z-norm utterances in zt-norm. Our purpose in varying the number of speaker factors was to evaluate the eigenvoice MAP and classical MAP components of the inter-speaker variability model in factor analysis. We found that on the core condition (i.e. 23 minutes of enrollment data), only the eigenvoice MAP component plays a useful role. On the other hand, on the extended data condition (i.e. 1520 minutes of enrollment data) both the classical MAP component and the eigenvoice component proved to be useful provided that the number of speaker factors was limited. Our best result on the extended data condition (all trials) was an equal error rate of 2.2% and a detection cost of 0.011.

17:00-17:10 *Discussion*

Tuesday

09:00-10:10 Invited Talk

David A. van Leeuwen of TNO Human Factors

A Human Benchmark for the NIST Language Recognition Evaluation 2005

co-authored by Michael de Boer and Rosemary Orr

In this paper we describe a human benchmark experiment for language recognition. We used the same task, data and evaluation measure as in the NIST Language Recognition Evaluation (LRE) 2005. For the primary condition of interest all 10-second trials were used in the experiment. The experiment was conducted by 38 subjects, who each processed part of the trials. For the seven-language closed set condition the human subjects obtained an average CDET of 23.1%. This result can be compared to machine results of the 2005 submission, for instance that of Brno University of Technology, whose system scored 7.15% at this task. A detailed statistical analysis is given of the human benchmark results. We argue that the result can best be expressed as the performance of 'naive subjects.'

10:10-10:40 *Tea*

Tuesday

10:40-12:50 Language Recognition

Session chair: Haizhou Li

10:40-11:00 *The ICSI 2007 Language Recognition System.*

Christian Muller and Joan-Isaac Biel

In this paper, we describe the ICSI 2007 language recognition system. The system constitutes a variant of the classic PPRLM (parallel phone recognizer followed by language modeling) approach. We used a combination of frame-by-frame multilayer perceptron (MLP) phone classifiers for English, Arabic, and Mandarin and one open loop hidden Markov Model (HMM) phone recognizer (trained on English data). The maximum likelihood language modeling is substituted by support-vector-machines (SVMs) as a more powerful, discriminative classification method. Rank normalization is used as a normalization method superior to mean-variance normalization. Results are presented on the NIST 2005 language recognition evaluation (LRE05) set and a test set taken from the LRE07 training corpus. The average NIST cost of the system on the LRE05 set is 0.0886.

11:00-11:20 *Language Identification: Insights from the Classification of Hand Annotated Phone Transcripts.*
Timothy Kempton and Roger K. Moore

Language Identification (LID) of speech can be split into two processes; phone recognition and language modelling. This two stage approach underlies some of the most successful LID systems. As phone recognizers become more accurate it is useful to simulate a very accurate phone recognizer to determine the effect on the overall LID accuracy. This can be done by using phone transcripts. In this paper LID is performed on phone transcripts from six different languages in the OGI multi-language telephone speech corpus. By simulating a phone recognizer that classifies phones into ten broad classes, a simple n-gram model gives low LID equal error rates (EER) of <1% on 30 seconds of test data. Language models based on these accurate phone transcripts can reveal insights into the phonology of different languages.

11:20-11:40 *Building language detectors using small amounts of training data.*
David A. van Leeuwen and Niko Brummer

In this paper we present language detectors built using relatively small amounts of training data. This is carried out using the modelling power of a Linear Discriminant Analysis back-end for the languages which have a small amount of training data. We present experiments on NIST 2005 Language Recognition Evaluation data, where we use a jackknifing technique to remove well-trained language knowledge from the LDA back-end, using only sparse trials for training the LDA. We investigate three systems, which show different levels of loss of language detection capability. We validate the technique on an independent collection of 21 languages, where we show that with less than one hour training we obtain an error rate for 'new' languages that is only slightly over twice the error rate for languages for which the full 60 hours of CallFriend data is available.

11:40-12:00 *NIST 2007 Language Recognition Evaluation.*
Alvin F. Martin and Audrey N. Le

This paper discusses NIST's 2007 evaluation of language recognition. Some history of earlier NIST language evaluations is covered, and the test procedures and protocols, evaluation data used, and planned measures of performance for the 2007 evaluation are described. The participants and submissions of the 2007 evaluation are described, and preliminary information

is included on the evaluation performance results after brief initial analysis.

12:00-12:20 *Pruned Universal Symbol Sequences for LZW based Language Identification*
S.V. Basavaraja and T.V. Sreenivas

We present a improved language modeling technique for Lempel-Ziv-Welch (LZW) based LID scheme. The previous approach to LID using LZW algorithm prepares the language pattern table using LZW algorithm. Because of the sequential nature of the LZW algorithm, several language specific patterns of the language were missing in the pattern table. To overcome this, we build a universal pattern table, which contains all patterns of different length. For each language it's corresponding language specific pattern table is constructed by retaining the patterns of the universal table whose frequency of appearance in the training data is above the threshold. This approach reduces the classification score (Compression Ratio [LZW-CR] or the weighted discriminant score [LZW-WDS]) for non native languages and increases the LID performance considerably.

12:20-12:40 *Discussion*
12:50-14:20 *Lunch*

Tuesday

14:20-15:30 *Speaker Recognition IV*

Session chair: John Mason

14:20-14:40 *Improving the performance of text-independent short duration SVM- and GMM-based speaker verification*
Benoit Fauve, Nicholas Evans and John Mason

In the task of automatic speaker verification (ASV) it is well known that the duration of the speech signals is an important factor in the ultimate accuracy of the system. This paper deals with some of the aspects of adapting systems to work with limited amounts of data. First we highlight the importance of a well-tuned speech detection front-end when working with short durations. We consider a well-established technique (GMM) as well as a recent development (SVM on GMM mean supervectors), showing their limitations and alternatives. In particular the benefit of eigenvoice modelling in the context of short duration tasks is highlighted. Finally experiments on standard NIST databases demonstrate fusion potential between the presented techniques and significant gains when compared to a single GMM.

14:40-15:00 *Factor Analysis Modelling for Speaker Verification with Short Utterances.*
Robbie Vogt, Chris Lustri and Sridha Sridharan

This paper examines combining both relevance MAP and subspace speaker adaptation processes to train GMM speaker models for use in speaker verification systems with a particular focus on short utterance lengths. The subspace speaker adaptation method involves developing a speaker GMM mean supervector as the sum of a speaker-independent prior distribution and a speaker dependent offset constrained to lie within a low-rank subspace, and has been shown to provide improvements in accuracy over ordinary relevance MAP when the amount of training data is limited. It is shown through testing on NIST SRE data that combining the two processes provides speaker models which lead to modest improvements in verification accuracy for limited data situations, in addition to improving the performance of the speaker verification system when a larger amount of available training data is available.

15:00-15:20 *ALIZE/SpkDet: a state-of-the-art open source software for speaker recognition.*
Jean-Francois Bonastre, Nicolas Scheffer, Driss Matrouf, Corinne Fredouille, Anthony Larcher, Alexandre Preti, Gilles Pouchoulin, Nicholas Evans, Benoît Fauve and John Mason

This paper presents the ALIZE/SpkDet open source software packages for text independent speaker recognition. This software is based on the well-known UBM/GMM approach. It includes also the latest speaker recognition developments such as Latent Factor Analysis (LFA) and unsupervised adaptation. Discriminant classifiers such as SVM supervectors are also provided, linked with the Nuisance Attribute Projection (NAP). The software performance is demonstrated within the framework of the NIST'06 SRE evaluation campaign. Several other applications like speaker diarization, embedded speaker recognition, password dependent speaker recognition and pathological voice assessment are also presented.

15:20-15:30 *Discussion*
15:30-16:00 *Tea*

Tuesday

16:00-17:10 Speaker Recognition V

Session chair: Jason Pelecanos

16:00-16:20 *Kernel Combination for SVM Speaker Verification.*
Reda Dehak, Najim Dehak, Patrick Kenny and Pierre Dumouchel

We present a new approach to construct kernels used on support vector machines for speaker verification. The idea is to learn new kernels by taking linear combination of many kernels such as the Generalized Linear Discriminant Sequence kernels (GLDS) and Gaussian Mixture Models (GMM) supervector kernels. In this new linear kernel combination, the weights are speaker dependent rather than universal weights on score level fusion and there is no need to extra-data to estimate them. An experiment on the NIST 2006 speaker recognition evaluation dataset (all trials) was done using three different kernel functions (GLDS kernel, Gaussian and linear GMM supervector kernels). We compared our kernel combination to the optimal linear score fusion obtained using logistic regression. The optimal weights was trained on all 1conv4w-1conv4w trials of NIST-SRE 2005. Testing on NIST-SRE 2006 database, we had an equal error rate of 5.9% using the kernel combination method which is better than the optimal score fusion system (6.1%).

16:20-16:40 *An Anticorrelation Kernel for Improved System Combination in Speaker Verification.*
Luciana Ferrer, Kemal Sonmez and Elizabeth Shriberg

This paper presents a method for training SVM-based classification systems for combination with other existing classification systems designed for the same task. Ideally, a new system should be designed such that, when combined with the existing systems, the resulting performance is optimized. To achieve this goal, we include a regularization term in the SVM objective function that aims to reduce the within-class correlation between the resulting scores and the scores produced by one of the existing systems, introducing a trade-off between such correlation and the system's individual performance. That is, the SVM system "takes one for the team", falling somewhat short of its best possible performance in order to be more complementary to the existing system. We report results on the NIST 2005 and 2006 speaker recognition evaluations (SRE) using three component systems: a standard UBM-GMM system, an MLLR-based system, and a prosodic system, and show that the proposed technique results in performance gains of 16% inEER and 23% in DCF.

16:40-17:00 *MLLR Techniques for Speaker Recognition.*
Marc Ferras, Cheung Chi Leung, Claude Barras and Jean-Luc Gauvain

Maximum-Likelihood Linear Regression (MLLR) and Constrained MLLR (CMLLR) have been recently used for feature extraction in speaker recognition. These systems use (C)MLLR transforms as features that are modeled with Support Vector Machines (SVM). This paper evaluates and compares several of these approaches for the NIST Speaker Recognition task. Single CMLLR and up to 4-phonetic-class MLLR transforms are explored using Gaussian Mixture Models (GMM) and large-vocabulary speech recognition Hidden Markov Models (HMM), using both speaker recognition and speech recognition cepstral front-ends and normalizations. Results for the individual systems as well as in combination with two standard cepstral systems are provided. Relative gains of 3% and 12% were obtained when combining the best performing CMLLR-based and MLLR-based systems with two standard cepstral systems, respectively.

17:00-17:10 *Discussion*

Thursday

09:00-10:40 *Speaker Classification*

Session chair: Jean-Francois Bonastre

09:00-09:20 *Recognizing Arabic Speakers with English Phones*
Andreas Stolcke and Sachin Kajarekar

We investigate the question of whether phone recognition models trained on large English databases can be used for speaker recognition in another language. Such a cross-language use of recognition models is an attractive option when a speaker recognition system is to be ported to a new language without the necessary data resources, while retaining some of the advantages of phone modeling and ASR-based feature extraction. We compare the performance of such systems to a baseline cepstral GMM system (which is inherently language independent), and to a phone recognition-based system trained exclusively on Arabic data. Our results indicate that cross-language models are highly competitive, and, at least in our case, have a performance advantage over within-language training and the language-independent baseline. We also examine the effect of coverage of colloquial Arabic dialects in the training data.

09:20-09:40 *Age and Gender Classification using Modulation Cepstrum.*
Jitendra Ajmera and Felix Burkhardt

This paper proposes using modulation cepstrum coefficients instead of cepstral coefficients for extracting metadata information such as age and gender. These coefficients are extracted by applying discrete cosine transform to a time-sequence of cepstral coefficients. Lower order coefficients of this transformation represent smooth cepstral trajectories over time. Results presented in this paper show that cepstral trajectories corresponding to lower (3-14 Hz) modulation frequencies provide best discrimination. The proposed system achieves 50.2% overall accuracy for this 7-class task while accuracy of human labelers on a subset of evaluation material used in this work is 54.7%.

09:40-10:00 *Detecting Nonnative Speech Using Speaker Recognition Approaches.*
Elizabeth Shriberg, Luciana Ferrer, Sachin Kajarekar, Nicolas Scheffer, Andreas Stolcke and Murat Akbacak

Detecting whether a talker is speaking his native language is useful for speaker recognition, speech recognition, and intelligence applications. We study the problem of detecting nonnative speakers of American English, using two standard speech corpora. We apply approaches effective in speaker verification to this task, including systems based on MLLR, phone N-gram, prosodic, and word N-gram features. Results show equal error rates between 12% and 20%, depending on the system, test data, and choice of training data. Asymmetries in performance are most likely explained by differences in native language distributions in the corpora. Model combination yields substantial improvements over individual models, with the best result being around 8.6% EER. While phone N-grams are widely used in related tasks (e.g., language and dialect ID), we find that it is the least effective model in combination; MLLR, prosody, and word N-gram systems play stronger roles. Overall, results suggest that individual systems and system combinations found useful for speaker ID also offer promise for nonnativeness detection, and that further efforts are warranted in this area.

10:00-10:20 *Accent identification in the presence of code-mixing.*
Thomas Niesler and Febe de Wet

We investigate whether automatic accent identification is more effective for English utterances embedded in a different language as part of a mixed code than for English utterances that are part of a monolingual dialogue. Our focus is on Xhosa and Zulu, two South African languages for which code mixing with English is very common. In order to carry out our investigation, we extract English utterances from mixed-code Xhosa and Zulu speech corpora, as well as comparable utterances from an English-only corpus by Xhosa and Zulu mother-tongue speakers. Experiments show that accent identification is substantially more accurate for the utterances originating from the mixed-code speech. We conclude that accent identification is more successful for these utterances because accents are more pronounced for English embedded in mother-tongue speech than for English spoken as part of a monolingual dialogue by non-native speakers.

10:20-10:40 *Discussion*
10:40-11:10 *Tea*

Thursday

11:10-12:20 *Speaker Recognition VI*

Session chair: Doug Reynolds

11:10-11:30 *Comparisons of Recent Speaker Recognition Approaches based on Word-Conditioning.*
Howard Lei and Nikki Mirghafori

We examine the effectiveness of various speaker recognition approaches based on word-conditioning. Subsets of 62 keywords (used for word-conditioning) are examined for their individual and combined effectiveness for a keyword HMM approach, a supervector keyword HMM approach, a keyword phone N-grams approach, and a keyword phone HMM approach. Our results demonstrate the effectiveness of acoustic features and importance of keyword frequency in individual keyword results, where the keywords *yeah* and *you know* outperform others. We also demonstrate the power of SVMs, in conjunction with acoustic features, in keyword combination experiments, in which the supervector keyword HMM approach (4.3% EER) outperforms other keyword-based approaches, and achieves a 6.5% improvement over the GMM baseline (4.6% EER) on the SRE06 8-conversation-side task.

11:30-11:50 *Phoneme and Sub-Phoneme T-Normalization for Text-Dependent Speaker Recognition.*
Doroteo T. Toledano, Cristina Esteve-Elizalde, Joaquin Gonzalez-Rodriguez, Ruben Fernandez Pozo and Luis Hernandez Gomez.

Test normalization (T-Norm) is a score normalization technique that is regularly and successfully applied in the context of text-independent speaker recognition. It is less frequently applied, however, to text-dependent or text-prompted speaker recognition, mainly because its improvement in this context is more modest. In this paper we present a novel way to improve the performance of T-Norm for text-dependent systems. It consists in applying score T-Normalization at the phoneme or sub-phoneme level instead of at the sentence level. Experiments on the YOHO corpus show that, while using standard sentence-level T-Norm does not improve equal error rate (EER), phoneme and sub-phoneme level T-Norm produce a relative EER reduction of 18.9% and 20.1% respectively on a state-of-the-art HMM based text-dependent speaker recognition system. Results are even better for working points with low false acceptance rates.

11:50-12:10 *Dimension Reduction of the Modulation Spectrogram for Speaker Verification.*
Tomi Kinnunen, Kong-Aik Lee and Haizhou Li

A so-called modulation spectrogram is obtained from the conventional speech spectrogram by short-term spectral analysis along the temporal trajectories of the frequency bins. In its original definition, the modulation spectrogram is a high-dimensional representation and it is not clear how to extract features from it. In this paper, we define a low-dimensional feature which captures the shape of the modulation spectra. The recognition accuracy of the modulation spectrogram based classifier is improved from our previous result of EER=25.1% to EER=17.4% on the NIST 2001 speaker recognition task.

12:10-12:20 *Discussion*

Thursday

12:2012:50

12:50

Closing

Lunch

Speaker Recognition VII (not presented)

Support Vector Machines Based Text Dependent Speaker Verification Using HMM Supervectors

Chengyu Dong, Yuan Dong, Jing Li and Haila Wang

Conventional subword based hidden Markov models (HMMs) have proven to be an effective approach for text-dependent speaker verification. The standard training method works by modeling the MAP adapted means of subword HMMs. In this paper, we propose the use of HMM supervectors from the speaker models as features in support vector machines (SVMs) classifier. An HMM supervector is constructed by stacking means of adapted mixture components from all states within HMMs. We present two SVM kernels: linear kernel and dynamic time alignment kernel (DTAK) based on the KL divergence to evaluate the system. In addition, another effective method is proposed to normalize SVM output scores using speaker independent HMM supervectors. Experimental results show that the SVM system with HMM supervectors achieves lower performance than conventional HMM verification system, but their fusion can give a significant improvement.

Component Score Weighting for GMM based Text-Independent Speaker Verification

Liang Lu, Yuan Dong, Xianyu Zhao, Hao Yang, Jian Zhao and Haila Wang

GMM/UBM framework is widely used in Automatic Speaker Verification (ASV), however, due to the insufficiency of the training data, both the hypothesized speaker and impostors are not well modeled, especially to some of the Gaussian component mixtures. Thus, the Gaussian mixtures in each GMM model have different discriminative capabilities, and the mismatch between testing and training data will also aggravate this situation. In this paper, we propose a novel approach, namely, Component Score Weighing (CSW), to reweight the Gaussian mixtures and highlight those which have high discriminative capability by post-processing the log-likelihood ratio (LLR). The original log-likelihood in GMM systems is assigned to each Gaussian component mixture, deriving two component score serials, which we called the dominant score serial and the residual score serial. A nonlinear score weighting function is then applied to reweigh those scores, respectively. Experiments on NIST 2006 SRE corpus show that, this approach achieves notable performance gains over our previous baseline system (about 12% relative improvement in minimum detection cost function (DCF) value).

Author Index

Ajmera, Jitendra	23
Basavaraja, SV	19
Bonastre, Jean-Francois	20
Dehak, Najim	15
Dehak, Reda	21
Dong, Chengyu	26
Fauve, Benoit	19
Ferras, Marc	22
Ferrer, Luciana	21
Gonzales-Rodrigues, Joaquin	11
Kempton, Timothy	18
Kenny, Patrick	15
Kim, Youngmoo E	14
Kinnunen, Tomi	25
Kinoshita, Yuko	11
Lei, Howard	24
Lu, Liang	26
Martin, Alvin F	18
Muller, Christian	17
Niesler, Thomas	24
Ramos, Daniel	12
Richiardi, Jonas	12
Shriberg, Elizabeth	23
Stolcke, Andreas	22
Toledano, Doroteo T	25
Van Leeuwen, David A	17, 18
Vogt, Robbie	15, 20
Yoon, Sang-min	13
Zamalloa, M	13

