

FINAL PROGRAMME

ISCA ITRW

SPEECH ANALYSIS AND PROCESSING FOR KNOWLEDGE DISCOVERY

JUNE 4 - 6, 2008

Invited tutorials

**Professor Sarah Hawkins, Department of Linguistics,
University of Cambridge, UK**

Phonetic perspectives on modelling information in the speech signal

Abstract

This talk reassesses conventional assumptions about the informativeness of the acoustic speech signal, and shows how recent research on systematic variability in the acoustic signal is consistent with a processing model that seems biologically plausible as well as compatible with recent advances in modelling embodied visual perception and action (robotics). I will review some common assumptions about the information available from the speech signal, including the strengths and limitations of phonological features and phonemes as units for encoding information in the signal. Features and phonemes are widely regarded as essential units of human speech perception and production, and, by extension, as central to machine speech recognition and synthesis systems. Yet they are abstract theoretical constructs developed within particular schools of theoretical linguistics for quite different purposes—describing phonological systems rather than speech communication. They are not intended to describe cognitive attributes of communication, nor linguistic meaning beyond contrasts in lexical form. Not surprisingly, then, models of speech perception which map cues derived from the speech signal solely onto such phonological units can only partially account for listeners' abilities. A comprehensive model of speech communication needs a broader theoretical approach, more clearly related to cognition and the wider communicative functions of speech. Present knowledge makes this goal challenging to implement, but we now know enough to begin to discuss it.

I will give examples of variation in phonetic detail which systematically signal non-phonemic linguistic information such as the grammatical or morphological status of a stretch of sound. Other examples indicate the discourse function of the utterance. Some of these systematic differences in phonetic detail are localised in the signal, while others stretch over several syllables. Both types can make speech easier to understand, presumably by increasing the signal's perceptual coherence and/or by directly indicating linguistic structure and pragmatic functions as well as phonological form.

These data encourage the development of models that focus attention away from formal phonological contrasts in citation-form words, and towards properties of connected speech that communicate other types of meaning. A currently useful approach uses principles of declarative phonology, in particular Firthian Prosodic Analysis, which places more emphasis than other models on the communicative function of an utterance, de-emphasizes the need to identify phonemes, and uses formalisms that force us to recognise that every perceptual decision is context- and task-dependent. Prosodic and grammatical properties of utterances are formally linked in twin hierarchical structures. No one descriptive unit, or hierarchical level, is more important than any other; and properties of the physical signal can inform any unit(s) in the grammatical and prosodic hierarchies in parallel. The model advocated simultaneously accommodates detail and abstraction. It assumes parallel streams of knowledge that drive interpretation of information in the incoming signal to maximise processing efficiency. The linguistic model alone is not sufficient, but its attributes are compatible with advances in robotics which

use function-oriented, body-centred, knowledge-driven systems to do visual tasks. The combined approach may be fruitful. The approach rejects the assumption that speech processing proceeds in strictly serial order, e.g. from the lowest phonological units to higher ones. Instead, sound can be mapped onto any level of linguistic structure, in parallel or in sequence, using context-dependent probabilistic processes. The important point is that there is no predetermined or rigid sequence: the process of speech perception is governed by the properties of the particular signal in conjunction with the listener's construal of the particular situation.

While this model may not prove to be the optimal one, adopting it encourages re-evaluation of the information available from the speech signal. In particular, it allows us to focus attention onto the meaning and communicative function of speech, rather than just on phonological identity. For speech technologists, this refocusing may contribute to improved unit selection and recognition algorithms.

Chaired by: Rolf Carlson, KTH

Professor Christophe d'Alessandro, CNRS-LIMSI, Orsay, France

New paradigms for speech analysis and processing: the source-filter model revisited and gesture-controlled analysis-by-synthesis

Abstract

Knowledge discovery in speech analysis and processing is based on both static and dynamic features of the speech signals. Static features are corresponding to parameters of a model or "settings". Dynamic features are corresponding to parameter trajectories, or "gestures".

In a first part, the source filter-model of speech production is revisited. Although the voiced source component is usually described by non-linear time-domain glottal flow models, spectral modelling suggests that it can be considered as a mixed-phase filter, with an anticausal component corresponding to glottal open phase and a causal component corresponding to glottal closure. Identification of this causal-anticausal model can be achieved exploiting the phase properties of the glottal flow. Two signal representations taking advantage of this description have recently been investigated at LIMSI (Orsay) and FPMs (Mons): Zero of the Z Transform representation and lines of maximum phase of the wavelet transform. The performances of these representations for source-filter separation and estimation of various parameters (glottal closure instants, open quotient, glottal flow asymmetry and spectral richness) demonstrate their viability as alternatives to inverse filtering and ElectroGlottographic analysis.

Speech synthesis has been for a long time one of the most fruitful tools for knowledge discovering in speech analysis and processing. In a second part, it is argued that this paradigm can be extended to dynamic features analysis, using real-time gesture-controlled instruments for analysis-by-synthesis. Experiments are reported, showing that such instruments allow for real-time manual control of glottal flow parameters, voice source aperiodicities and vocal tract formants. This could bring new insights into the dynamics of voice and speech in tasks such as expression of attitudes and prosody mimicking.

Chaired by: Paavo Alku, HUT

Regular papers

Oral Session 1

Estimating Speech Production Parameters

Wednesday June 4th, 13.30 - 15.00

Chaired by: Torbjörn Svendsen, NTNU

A unified approach for F0 extraction and aperiodicity estimation based on a temporally stable power spectral representation

Hideki Kawahara 1, Masanori Morise 1, Toru Takahashi 2, Ryuichi Nisimura 1, Hideki Banno 3, Toshio Irino 1

1 Faculty of Systems Engineering, Wakayama University, Wakayama, Japan

2 Graduate School of Informatics, Kyoto University, Kyoto, Japan

3 Faculty of Science and Technology, Meijo University, Nagoya, Japan

A power spectrum estimation method for periodic signals was proposed to provide temporally stable representation and has been applied to reformulate STRAIGHT, a system for speech analysis modification and synthesis based on stable spectral envelope estimation. This article proposes a specialized F0 detector based on a ratio between this stable spectrum and corresponding spectral envelope. By allocating multiple specialized F0 detectors and integrating individual clues, the proposed method selectively detects only fundamental components and yields a probability measure for each estimate. It also provides a method to estimate aperiodicity in each frequency band by making use of estimated fundamental frequency information to design a quadrature signal on the frequency axis for filtering periodic spectral component due to the signal periodicity. The proposed method shed new lights on source filter representation/decomposition of speech signals.

On the Estimation of the Speech Harmonic Model

Yannis Pantazis 1, Olivier Rosenc 2 and Yannis Stylianou 1

1 Institute of Computer Science, FORTH, and Multimedia Informatics Lab, CSD, UoC, Greece

2 Orange Labs TECH/SSTP/VMI, Lannion, France

In this paper we present and compare four time-domain approaches for estimating the parameters of a harmonic speech model. The classic approach of Least Squares is directly compared with a Total Least Squares approach trying to overcome errors in the estimation of the fundamental frequency of the model. Both of these approaches are suboptimal since they split the estimation problem into two subproblems; to the estimation of amplitudes and phases and to the estimation of fundamental frequency. To improve the accuracy of the parameters estimation of the harmonic model two iterative non linear approaches are then presented, based on the Steepest Descent and Newton-Gauss optimization algorithms, where all parameters of the harmonic model are estimated simultaneously. The approach based on the Newton-Gauss optimization algorithm provided the best accuracy as this is measured by the Signal-to-Noise Ratio criterion.

Analysis of Stop Consonants in Indian Languages Using Excitation Source Information in Speech Signal

B. Yegnanarayana 1, K. Sri Rama Murthy 2 and S. Rajendran 1

1 International Institute of Information Technology, Hyderabad-500032, India

2 Dept. of Computer Science & Engineering, IIT Madras, Chennai-600036, India

In this paper we propose excitation-based features for extracting information about the manner of articulation for stop consonants.

The excitation-based features are derived from very low frequency information in the signal and also from the normalized error computed from the linear prediction residual. The proposed zero-frequency filtered signal brings out the region of glottal activity during excitation. Likewise, the normalized error helps to distinguish regions of noise and pure voicing. These nonspectral methods of analysis of stop consonants seem to provide additional and some better features over the features derived from the traditional methods based on short-time spectrum analysis.

Oral Session 2 *Attribute Detection and Knowledge Discovery*
Wednesday June 4th, 15.30 - 17.00

Chaired by: Jim Baker, Carnegie Mellon University and Johns Hopkins University

Integration of Asynchronous Knowledge Sources in a Novel Speech Recognition Framework.
Hugo Van hamme
Katholieke Universiteit Leuven, dept. ESAT, Belgium

Hidden Markov Models have been essential in obtaining today's successes in speech recognition. However, some limitations of HMMs become clear: for example it is difficult to successfully exploit features that are measured at different time scales than the centisecond scale at which the spectral features are measured. Little success has been achieved in integrating utterance level information such as prosody, segmental information and finer detail such as voice onset times. In this paper, we apply latent semantic analysis (LSA) techniques known from the text processing field to histograms of acoustic event co-occurrence (HAC) to propose a novel speech recognition framework. We show that the HAC-method can deal with correlated information and exploit knowledge sources that are asynchronous.

Incorporating Suprasegmental Knowledge for Phone Recognition with Conditional Random Fields

Prateeti Mohapatra, Eric Fosler-Lussier
Ohio State University

In this paper, we investigate the integration of lexical stress and syllabic position of segments in a Multi-Layer Perceptron (MLP) classification system that is part of our Conditional Random Fields (CRF) phone recognizer. CRFs are used to integrate MLP posterior estimates, particularly of phonological features or phonetic classes, which stand in as representations of the acoustics; we show that incorporating suprasegmental information as part of the MLP classification system augments the acoustic space in a beneficial way for phonological feature based CRF models. TIMIT phone recognition experiments show a small but statistically significant improvement.

An Experimental Study on Continuous Phone Recognition with Little or No Language-Specific Training Data

Dau-Cheng Lyu 1, Sabato Marco Siniscalchi 2 and Chin-Hui Lee 3
1 Department of Electrical Engineering, Chang Gung University, Tao-Yuan, Taiwan
2 Department of Electronics and Telecommunications Norwegian, University of Science and Technology, Trondheim, Norway
3 School of ECE, Georgia Institute of Technology, Atlanta, GA 30332 USA

We study continuous phone recognition with little or no language-specific speech training data. The phone recognizer integrates three levels of information from: (1) frame based speech attribute detectors, (2) artificial neural network based phone event mergers, and (3) decoding based

evidence verifiers. With a set of acoustic phonetic attributes defined over a number of available languages, a collection of attribute-to-phone mapping rules can either be specified in a language-dependent way, one for each language, or even independently for all languages if the attribute specification is complete to cover all phones and the phone definition is universal to cover all spoken languages. We report on experimental results on Japanese phone recognition with the OGI Multilingual Speech Corpus. It is interesting that a good performance can be achieved without using any Japanese speech training data, and the phone accuracy rates vary depending on how the attribute detectors and phone mergers are configured. Further improvement is observed by adding little Japanese data to train the attribute-to-phone mergers.

Oral Session 3

Features for Speaker Recognition

Thursday June 5th, 09.00 - 10.30

Chaired by: Yegnanarayana Bayya, IIT, India

On the Relative Importance of the Short-Time Magnitude and Phase Spectra Towards Speaker Dependent Information

Kamil K. Wojcicki, Kuldip K. Paliwal, Signal Processing Laboratory, Griffith University, Nathan, Australia

In this work, we investigate the relative contribution of the short-time magnitude and phase spectra towards speaker dependent information. The effect of the analysis window function type is also examined. For this purpose we conduct a human speaker verification experiment that uses phase-only and magnitude-only stimuli. The stimuli are constructed using the analysis-modification-synthesis procedure. The results of our pilot experiment show that the short-time magnitude spectrum contains little speaker information for a low dynamic range analysis window and high amount of speaker information for a large dynamic range window. On the other hand, the short-time phase spectrum contains speaker information predominantly for the low dynamic range analysis window. These suggestive results show that the short-time phase spectrum, commonly discarded in feature extraction for speaker recognition, contains useful speaker information. This suggests that further research into feature extraction from the short-time phase spectrum is warranted.

Spectral Slope Measurements in Emotionally Expressive Speech

Lucas Tamarit 1, Martijn Goudbeek 2 & Klaus Scherer 1,2

1 Swiss Center for Affective Sciences, Geneva, Switzerland

2 Geneva Emotion Research Group, University of Geneva, Geneva, Switzerland

Characterizations of the voice spectrum in emotional utterances like the Hammarberg index or the amount of energy below a certain pivot point often ignore speaker dependent pitch information. Furthermore, these approximations of spectral slope are not very sensitive to details of the spectral shape. Here, we present three approaches to incorporate speaker specific information in determining the appropriate pivot value for distinguishing between low and high frequencies. Additionally, three different approaches of characterizing the spectral slope of the LTAS (Long-Term Average Spectrum) are presented. The validity of these approaches is tested on a corpus of emotional speech developed at the University of Geneva. The results show the feasibility and usefulness of taking speaker specific information into account as well as more extensive characterizations of the slope of the LTAS.

Automated Speaker Recognition Using Compressed Temporal-Spectral Dynamics Information of Password Spectrograms

Amitava Das 1 and *Gokul Chittaranjan* 2

1 Microsoft Research Lab – India; 196/36 2nd Main; Sadashivnagar; Bangalore India 560 080.

2 Student intern at MSR-India

Prevalent speaker recognition methods use only spectral-envelope based features such as MFCC, ignoring the rich speaker identity information contained in the temporal-spectral dynamics of the entire speech signal. We propose a new feature for speaker recognition called compressed spectral dynamics (CSD) which effectively captures such spectral dynamics and the inherent speaker identity. The discriminative power of CSD allows the classification part to remain simple. The proposed method, a simple nearest neighbor classifier using CSD, delivers performance competitive to conventional MFCC+DTW based text-dependent speaker recognition methods at significantly reduced complexity.

Oral Session 4

Acoustic Event Detection

Thursday June 5th, 11.00 - 12.30

Chaired by:

Ove Andersen, AAU

Innovative acoustic probes to test predictions of wider utterance context

D. R. L. Davies 1, *J. B. Millar* 2

1 Information Sciences and Engineering, University of Canberra, Australia

2 Research School of Information Sciences and Engineering, ANU, Australia

Innovative measures that are targeted to specific regions of the acoustic stream of speech are described as part of a predictive speech recognition system comprising multiple dimensions. Each dimension generates its own constraint on the next stage of interpretation of an unknown utterance and together they suggest targeted questions to be asked of the acoustic stream. Acoustic probes that address distinctions between vocalic nuclei and between stop consonants are presented as illustrations of the technique. A novel parametric stability level measure providing segmentation of the acoustic stream is applied alongside more conventional measures and their relative performance is noted.

Time-Varying Cepstral Coefficients

Trond Skogstad, Torbjørn Svendsen

NTNU

This paper introduces a new set of cepstral features, based on a slightly modified version of the time-varying linear predictive models pioneered by Subba Rao and Liporace. In these models, the non-stationarity of the speech signal is accommodated by expressing the filter coefficients as a weighted combination of known basis functions. By running the parameterized filter coefficients through the recursive link between all-pole models and cepstral coefficients, we obtain a time-varying cepstral representation in analytical form. In a preliminary recognition experiment this representation is shown to give a satisfactory performance. It is argued that the introduced features are well suited for tasks such as detection of landmarks and stationary segments.

Effective Segmentation based on Vocal Effort Change Point Detection

Chi Zhang and John H.L. Hansen

Center of Robust Speech Systems (CRSS)

Erik Jonsson School of Engineering & Computer Science

University of Texas at Dallas, Richardson, Texas 75083, USA

Non-neutral speech data has a strong negative impact on speech processing systems such as Automatic Speech Recognition (ASR) or speaker ID systems [1]. It is therefore necessary to detect and segment non-neutral speech data before further processing steps. Alternatively, the detection and segmentation of non-neutral speech segments from an input speech stream can be used in speech analysis and understanding, or in speech file retrieval systems to detect speech files containing whispered speech representing sensitive information, or shouted speech denoting strong emotion. This study addresses the segmentation problem for vocal effort change by deploying an improved feature based T2-BIC algorithm. Several features are considered as input to the T2-BIC algorithm in this study. A new fused evaluation criterion, Multi-Error Score (MES), is proposed to explore which feature conveys the most information on vocal effort. Results show that the lowest mean MES (56.49) occurs for the energy ratio feature for segmentation of different vocal effort speech segments based on vocal effort change point detection. Finally, recommendations are made for integrating this framework to advance knowledge processing for subsequent speech systems.

Poster Session

Speech Analysis and Modelling for Production and Recognition

Thursday June 5th, 13.30 - 17.00

Chaired by:

Chin-Hui Lee, Georgia Tech

Fitting Mass-Spring Models to Glottal Flow Estimates

Tom Bäckström

Department of Signal Processing and Acoustics, TKK (Helsinki University of Technology), POBox 3000, FI-02015 TKK, Finland.

We present a straightforward method for estimating the parameters of a physical mass-spring model of the vocal folds using a glottal flow signal. Initial experiments show that the method can successfully be used to model the glottal flow waveform. Parameters obtained can be used to characterise the underlying physical system in, for example, when used as features for speech recognition applications.

Investigating Explicit Model Transformations for Speaker Normalization in Speech Recognition

Mats Blomberg, Daniel Elenius

Dept Speech, Music and Hearing, CSC/KTH, Stockholm, Sweden

In this work we extend the test utterance adaptation technique used in vocal tract length normalization to a larger number of speaker characteristic features. We perform partially joint estimation of four features: the VTLN warping factor, the corner position of the piece-wise linear warping function, spectral tilt in voiced segments, and model variance ratio. In experiments on the Swedish PF-Star children database, the parameters contribute in lowering the recognition error rate.

Noise robust digit recognition using sparse representations

J. F. Gemmeke, B. Cranen

Centre for Language and Speech Technology (CLST), Radboud University, P.O. Box 9103, NL-6500 HD Nijmegen, The Netherlands

Despite the use of noise robustness techniques, automatic speech recognition (ASR) systems make many more recognition errors than humans, especially in very noisy circumstances. We argue that this inferior recognition performance is largely due to the fact that in ASR, speech is typically processed on a frame-by-frame basis preventing the redundancy in the speech signal to be optimally exploited.

We present a novel non-parametric classification method that can handle missing data while simultaneously exploiting the dependencies between the reliable features in an entire word.

We compare the new method with a state-of-the-art HMM-based speech decoder in which missing data are imputed on a frame-by-frame basis. Both methods are tested on a single digit recognition task (based on AURORA-2 data) using an oracle and an estimated harmonicity mask. We show that at an SNR of -5 dB using the reliable features of an entire word allows an accuracy of 91\% (using mel-log-energy features in combination with an oracle mask), while a conventional frame-based approach achieves only 61\%. Results obtained with the harmonicity mask suggest that this specific mask estimation technique is simply unable to deliver sufficient reliable features for acceptable recognition rates at these low SNRs.

The Hartley Phase Spectrum as a noise-robust feature in speech analysis

Ioannis Paraskevas, Maria Rangoussi

Department of Electronic Engineering, University of Surrey, Guildford, UK, Department of Electronics, Technological Education Institute of Piraeus, Athens, Greece

The importance of information contained in the signal phase spectrum is recognized by research dealing with speech or audio signals. Accurate phase extraction is a significant preprocessing step for speech applications such as coding, synchronization, synthesis or recognition. The Hartley Phase Spectrum (HPS) has been introduced as an advantageous alternative to the Fourier Phase Spectrum (FPS), for the analysis of speech and audio signals in order to extract phase information retaining features. In particular, it has been shown that the HPS suffers fewer discontinuities as compared to the FPS, while it requires no unwrapping. Successful exploitation of the HPS in preliminary audio classification and speech modeling experiments prompt us to further investigate in this work the noise robustness properties of the HPS, aiming towards the analysis of noisy speech. The noise robustness of the HPS is proved here via its probability density function for Gaussian noise and extended to the case of speech signals with additive noise.

Complex Wavelet Modulation Sub-Bands and Speech

Jean-Marc Luneau, Jérôme Lebrun, Søren Holdt Jensen

ES-MISP, Aalborg University, I3S-CNRS UMR-6070

A new class of signal transforms called Modulation Transforms has recently been introduced. They add a new dimension to the classical time/frequency representations, namely the modulation spectrum. Although very efficient to deal with different applications like feature extraction, speech recognition and also analysis for audio coding, these transforms show their limits e.g. when used to remove non-trivial noise from speech signals. Modulation sub-band decompositions based on the computation of the Hilbert envelope have been proved to create disturbing artifacts. We detail here a new way to deal properly with the phase and the magnitude of the modulation spectrum in a linear and analytic framework based on a complex wavelet transform. This Complex Wavelet Modulation Sub-Band transform gives

some interesting results in speech denoising and proposes a new approach for analytic signal processing in general.

Comparing Human and Machine Recognition Performance on a VCV Corpus

Odette Scharenborg 1 and *Martin Cooke* 2

1 Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands

2 Speech and Hearing Research Group, Dept. of Computer Science, University of Sheffield, UK

Listeners outperform ASR systems in every speech recognition task. However, what is not clear is where this human advantage originates. This paper investigates the role of acoustic feature representations. We test four (MFCCs, PLPs, Mel Filterbanks, Rate Maps) acoustic representations, with and without 'pitch' information, using the same back-end. The results are compared with listener results at the level of articulatory feature classification. While no acoustic feature representation reached the levels of human performance, both MFCCs and Rate maps achieved good scores, with Rate maps nearing human performance on the classification of voicing. Comparing the results on the most difficult articulatory features to classify showed similarities between the humans and the SVMs: e.g., 'dental' was by far the least well identified by both groups. Overall, adding pitch information seemed to hamper classification performance.

Acoustic profiles in emotion – the GEMEP corpus

Martijn Goudbeek 1, *Klaus R. Scherer* 1, 2

1 Geneva Emotion Research Group, University of Geneva, Switzerland

2 Swiss Center for Affective Science, Geneva, Switzerland

The first basic acoustic measurements on a new corpus of acted emotional expressions are presented. The corpus has been constructed in an attempt to bridge the divide between naturalistic emotional expressions, laboratory inductions, and acted expressions. The corpus contains a large number of emotions (including some rarely studied ones) that have been elicited from professional actors in interactive sessions. Acoustic measurements of duration, fundamental frequency, intensity and voice quality parameters are presented. The results show reliable effects for emotional expression, mostly confirming predictions. The structure in the data and the future applications of the corpus are discussed.

Speech Analysis by Time-Varying Lattice Filters

Karl Schnell

Institute of Applied Physics, Goethe-University Frankfurt, Max-von-Laue-Str. 1, D-60438 Frankfurt am Main, Germany

In this contribution, for speech analysis an estimation procedure considering time-varying reflection coefficients is proposed. The time-varying reflection coefficients of the lattice filter are estimated by minimizing the output powers of each section of the FIR lattice filter. For that purpose, the coefficient trajectory of each speech segment is parameterized by linear basis functions. To ensure continuous trajectories at the frame boundaries, the trajectory of each frame is estimated with respect to the coefficients of the left-side frame resulting in a dependent analysis of the frames. The frames can be analyzed successively from left to right in an adaptive way yielding a continuous piece-wise linear parameter trajectory in terms of reflection coefficients.

Using Zeros of the z-transform in the Analysis of Speech Signals

Paul Dalsgaard, Christian F Pedersen, Ove Andersen 1, Yegnanarayana Bayya 2

1 Department of Electronic Systems, Aalborg University, Denmark

2 International Institute of Information Technology, Hyderabad, India

The objective of this study is to analyse speech signals using the zeros of the z-transform of the signal. Trajectories of the zeros are used to study the characteristics of speech production. The trajectories are obtained by varying the parameters of the window function used on the signal segment. A skew Poisson function is defined with three parameters to control the window function. The proposed method does not assume any model for the analysis. Both synthetic and natural speech signals are analyzed. The goal of this study is to demonstrate, and eventually separate the information of the source part from the vocal tract system part of the speech production process from the signal. This may provide new and additional insights into the speech production process over and above the existing methods such as the spectral and group delay methods. The results from experiments with varying Poisson window parameters and speech signals are presented and discussed.

Oral Session 5

Speech Attributes and Knowledge Discovery

Friday June 6th, 09.00 - 10.30

Chaired by:

Rolf Carlson, KTH

An acoustic investigation of the [ATR] feature effect on vowel-to-vowel coarticulation

Christina Orphanidou, Greg Kochanski and John Coleman

Phonetics Laboratory, University of Oxford, UK

We report quantitative measurements of vowel-to-vowel coarticulation for neighbouring and non-neighbouring vowels that result from changes in the Advanced Tongue Root (ATR) feature. Native speakers of Southern British English produced utterances; we matched pairs that were identical except for a contrast of [+ATR] vs. [-ATR] on one transmitter vowel. From recordings, we computed an acoustic representation at the centres of the vowels, measuring the change in pronunciation associated with the change in [ATR] value.

We observed a consistent coarticulatory effect two syllables after the transmitter when it was a low vowel (i.e. the /aa/ vs. /uh and /oo/ vs. /o/ pairs) but little or no effect when the transmitter was a high vowel (i.e. the /i/ vs. /ii/ and /u/ vs. /uu/ pairs.)

Interestingly, we observed more dramatic coarticulation on the next-nearest neighbour than on the vowel following the transmitter.

Joint Optimization of Event Detectors and Evidence Merger for Continuous Phone Recognition

Sabato Marco Siniscalchi, Øystein Birkenes, Magne H. Johnsen, and Torbjørn Svendsen

Department of Electronics and Telecommunications, NTNU, Trondheim, Norway

In the recent years, different data-driven methods have been proposed to detect articulatory features (AF) from short-term spectral representation. The main motivations for the AF based approach are as follows. First, the AFs in general can more accurately and parsimoniously characterize the acoustic variability associated with conversational speech. Further, while not explored in this work, AFs are more language universal than phones, and therefore they can generalize better and are easier to adapt to new languages. For use in phone based systems the AF scores are input to an evidence merger which produces phone posteriors as outputs. Several classifiers are usually built, and each classifier is trained for detecting a single articulatory feature (describing manner and/or place).

We believe that joint optimization of all the classifiers and the subsequent phone evidence merger may be beneficial for the classification performance.

This work is a preliminary study on this direction, and it is validated on the continuous phone recognition task. A bank of articulatory detectors, designed using hidden Markov models (HMMs), learns the mapping from the MFCC space to the articulatory space. The detectors outputs are then combined by the evidence merger. The AF based phone posteriors is integrated into an existing ASR engine and applied to N-best rescoring. Experimental results show promising performance on the TIMIT corpus.

Unsupervised detection of words – questioning the relevance of segmentation

Louis ten Bosch 1, *Hugo Van hamme* 2, *Lou Boves* 1

1 Language and Speech, Radboud University Nijmegen, the Netherlands

2 ESAT, Katholieke Universiteit Leuven, Belgium

In this paper, we discuss a computational model of language acquisition that is able to detect and build word-like representations on the basis of multimodal input data. Experiments carried out on three European languages (Finnish, Swedish, and Dutch) show that internal word representations can be learned without a predefined lexicon. The computational model is inspired by the memory structure that is assumed to underlie human cognitive processing. The model does not use any prior segmentation, nor does it use the concept of segmentation later in the processing. This calls into question the importance that is conventionally attributed to the segmentation of the speech signal in terms of symbolic units for the purpose of detecting structure in speech.

Oral Session 6

Speech Recognition and Classification

Friday June 6th, 11.00 - 12.30

Chaired by:

Paavo Alku, HUT

Enhancing Noise Robustness in Automatic Speech Recognition Using Stabilized Weighted Linear Prediction (SWLP)

Jouni Pohjalainen, Carlo Magi, Paavo Alku

Department of Signal Processing and Acoustics Helsinki University of Technology, P.O. Box 3000, FI-02015 TKK, Finland

Stabilized weighted linear prediction (SWLP) is a recently developed method to compute stable all-pole models of speech by applying temporal weighting of the residual energy. In this study, SWLP is used for spectrum estimation in the first stage of the MFCC computation. The resulting acoustic feature representation is tested in a speech recognition front-end in simulated noisy conditions. When compared to other spectrum estimation methods as a part of the MFCC framework, the proposed spectrum estimation method clearly outperforms the FFT (periodogram), linear prediction and minimum variance distortionless response (MVDR) methods in terms of noise robustness.

Feature selection algorithms for the creation of multistream speech recognizers

Yotaro Kubo 1, *Shigeki Okawa* 2, *Akira Kurematsu* 1, *Katsuhiko Shirai* 1

1 Department of Computer Science and Engineering, Waseda University, Tokyo, Japan

2 Chiba Institute of Technology, Narashino, Japan

In this paper, we present a method to split a feature stream into multiple feature streams. The efficiency of ensemble classifiers for speech recognition is confirmed by several experiments.

The conventional methods for constructing multiple classifiers are done by

splitting the feature stream by type of features or subbands where the features are associated. The splitting approach is well suited for obtaining high-dimensional features because it naturally leads to dimension reduction of features.

In order to take advantage of ensemble classifiers, each classifier should compensate for the errors due to the other classifiers. Therefore, each classifier should be independent from others. We proposed a method to split a feature stream using stream independency criteria in order to constructing independent classifiers.

We evaluated several stream splitting methods and compare word error rate by conducting continuous digit recognition experiments on noisy speech. Our method can reduce 30.9% of the word error when compared with the single classifier method, while it reduces 3.2% of the word error when compared with conventional multistream approach.

Discrimination of Speech from Nonspeech in Broadcast News Based on Modulation Frequency Features

Maria Markaki 1, *Yannis Stylianou* 1,2

1 Computer Science Department, University of Crete, Greece

2 Institute of Computer Science, FORTH, Greece

We describe a content based speech discrimination algorithm in broadcast news based on the time-varying information provided by the modulation spectrum. Due to the varying degrees of redundancy and discriminative power of the acoustic and modulation frequency subspaces, we first employ a generalization of SVD to tensors (Higher Order SVD) to reduce dimensions.

We further select the optimal principal axes in each subspace based on mutual information. Projection of modulation spectral features in these axes results in a compact feature set at a very low cost for subsequent classification with SVMs. We present experimental comparison between our algorithm and MFCCs using the same classifier and dataset
