UK Speech

www.ukspeech.org.uk

One Day Meeting for Young Speech Researchers

Wednesday 16th July 2008

University of Surrey

Programme and Abstracts of Posters and Talks

Organiser:

Dr Philip Jackson
Centre for Vision Speech and Signal Processing
Department of Electronic Engineering (FEPS)
University of Surrey, Guildford GU2 7XH.

p.jackson@surrey.ac.uk

Programme at a glance

10:00-10:30	Coffee and badges available
10:30-11:55	Posters 1
11:55-12:00	Welcome to Surrey and introduction
12:00-13:00	Talks 1: Korin Richmond, Chandra Raut, Catherine Breslin
13:00-14:00	Lunch
14:15-15:15	Talks 2: Abeer Alwan, Eng-Jon Ong, Gregor Hofer
15:15-15:20	Closing remarks and announcements
15:20-17:00	Demos and tea
15:30-17:00	Posters 2

Contents

Acknowledgements

Thanks to the help and support of the founders of UK Speech, Prof Stephen Cox (University of East Anglia) and Dr Mark Huckwale (University College London). Thanks to Prof Joseph Kittler, the head of Centre for Vision, Speech and Signal Processing, for supporting the organisation of this meeting at the University of Surrey. Thanks to the speakers and authors for their contributions. Thanks to colleagues and research students of CVSSP who have contributed to the mobilisation and preparation for the meeting.

Programme in detail

Posters 1 — 10:30am - 11:55am

A1	Application of Weighted Finite-State Transducers to Improve Recognition Accuracy for Dysarthric Speech	Omar Caballero, Stephen Cox	University of East Anglia
A2	Covariance Modelling for Noise-Robust Speech Recognition	RC van Dalen, MJF Gales	Cambridge University Engineering Department
A3	A Shrinkage Estimator for Speech Recognition with Full Covariance HMMs	Peter Bell, Simon King	Edinburgh University
A4	Longitudinal study of ASR performance on ageing Voices	Ravichander Vipperla, Steve Renals, Joe Frankel	Edinburgh University
A5	3D video analysis of emotional speech	Nataliya Nadtoka, James Edge, Adrian Hilton, Philip Jackson	University of Surrey
A6	Audiovisual Emotion Recognition in an English Database	Sanaul Haq, Philip JB Jackson, James Edge	University of Surrey
A8	Joint Quantization Strategies for Low Bit-Rate Sinusoidal Coding	Emre Unver, Stephane Villette, Ahmet Kondoz	University of Surrey
A9	Non-acoustic Speech Interface	Sertan Kaymak, Stephane Villette, Ahmet Kondoz	University of Surrey
A10	Unsupervised Feature Vector Normalization with Combined Standard and Throat Microphones for Robust ASR	Luis Buera, Antonio Miguel, Oscar Saz, Alfonso Ortega, Eduardo Lleida	Toshiba Research Europe
A11	Audio-based alignment of pre-production video	Siripinyo Chantamunee, Yoshihiko Gotoh	University of Sheffield
A12	Tense and lax vowels as a consequence of speech embodiment	Piers Messum	
A13	The speech of our cousins and our ancestors	Antoine Serrurier, Anna Barney	University of Southampton
A14	Speech recognition using non-linear trajectories in a formant-based articulatory layer of a multiple-level segmental HMM	Hongwei Hu, Martin J Russell	University of Birmingham

Talks 1 — 12:00am - 13:00am

B1	Exploiting articulation in speech technology	Korin Richmond	CSTR, Edinburgh University
B2	Instantaneous Unsupervised Adaptation using Discriminative Mapping Transforms	Chandra Raut, Kai Yu, Mark JF Gales	Cambridge University Engineering Department
B3	Improving Automatic Phone Alignments for Speech Synthesis	Catherine Breslin	Toshiba Research Europe

	Talks 2	**2:15pm - 3:15pm**	
C1	Rapid Speaker Adaptation with Limited Data	Abeer Alwan	University of California, LA
C2	Robust Lip-Tracking using Rigid Flocks of Selected Linear Predictors	Eng-Jon Ong, Richard Bowden	University of Surrey
C3	Speech-driven Lip Motion Generation with a Trajectory HMM	Gregor Hofer, Junichi Yamagishi, Hiroshi Shimodaira	Edinburgh University

	Demonstrations	**3:20pm - 5:00pm**	
D1	Sound Source Localisation, Acoustic Source Separation and Binaural Audio Processing	Banu Gunel	University of Surrey
D2	Quality Evaluation of Spatial Transmission and Reproduction by an Artificial Listener	Martin Dewhirst, Rob Conetta	University of Surrey
D3	3D Capture of Dynamic Speech Movements James Edge	University of Surrey	

	Posters 2	**3:30pm - 5:00pm**	
E1	Automatic Video Language Identification	Jacob L Newman, Stephen J Cox	University of East Anglia
E2	A Generalised Derivative Kernel for Speaker Verification	C Longworth, MJF Gales	Cambridge University Engineering Department
E3	Direct Posterior-Based Confidence Measurement for Spoken Term Detection	Dong Wang, Joe Frankel, Simon King	Edinburgh University
E4	HMM-based synthesis of child speech	Oliver Watts, Junichi Yamagishi, Kay Berkling, Simon King	Edinburgh University
E5	Articulatory resynthesis of EMA data	Ingmar Steiner, Korin Richmond	Edinburgh University
E6	A Novel Multi stage Approach For Blind Separation of Convolutive Speech Mixtures	Tariqullah Jan, Wenwu Wang, DeLiang Wang	University of Surrey; Ohio State University
E7	Towards deriving compact and meaningful articulatory representations: an analysis of feature extraction techniques	Veena D Singampalli, Philip JB Jackson	University of Surrey
E8	Parallel model combination and digit recognition with soccer audio	J Longton, Philip JB Jackson	University of Surrey
E9	Voice Impersonation and Speaker Recognition	Zargham Haider,Stephane Villette, Ahmet Kondoz	University of Surrey
E10	Index assignment based channel coding for speech/audio communications	Huseyin Oztoprak, Stephane Villette, Ahmet Kondoz	University of Surrey
E11	Synthesising personalised voices for individuals with progressive speech loss	Sarah Creer, Phil Green, Stuart Cunningham	University of Sheffield
E12	Cross-language Voice Conversion: Analysis from a Phonetic Perspective	Kayoko Yanagisawa, Mark Huckvale	University College London
E13	The Development of the ”TalkMaths” Speech Interface at Kingston University	Angela Wigmore, Gordon Hunter, Eckhard Pflgel, James Denholm-Price	Kingston University

A1.

Title: Application of Weighted Finite-State Transducers to Improve Recognition Accuracy for Dysarthric Speech

Authors: Omar Caballero and Stephen J Cox

Address: University of East Anglia
Norwich NR4 7TJ UK

Email: S.Caballero-morales@uea.ac.uk

Abstract

Dysarthria is a motor speech disorder characterized by weakness, paralysis, or poor coordination of the muscles responsible for speech. Although automatic speech recognition (ASR) systems have been developed for disordered speech, factors such as low intelligibility and limited vocabulary decrease speech recognition accuracy. Speaker adaptation algorithms as MLLR (maximum likelihood linear regression) have been shown to be successful for increasing word recognition accuracy for normal speakers, however they may be less successful in cases where the phoneme uttered is not the one that was intended but is substituted (or confused) by a different phoneme or phonemes, as often happens in dysarthric speech.

Previously, we proposed the use of discrete hidden Markov models (HMMs) that we termed ”metamodels” which incorporated a model of the dysarthric speaker's phonetic confusion-matrix into the ASR process in such a way as to increase word recognition accuracy. Although this technique performed better than MLLR when the number of sentences available from a speaker for confusion-matrix estimation or adaptation was low, this advantage decreased as the number of sentences increased.

Now, we report on an improved technique that makes use of a cascade of Weighted Finite-State Transducers (WFSTs) to model the speaker's phonetic confusions, the mapping from phonemes to words, and the mapping from words to a word sequence described by a grammar. This approach outperforms both standard MLLR and metamodels.

A2.

Title: Covariance Modelling for Noise-Robust Speech Recognition

Authors: Rogier C van Dalen and Mark JF Gales

Address: Cambridge University Engineering Department,
Baker Building, Room 502
Trumpington Street
Cambridge, CB2 1PZ.

Email: rcv25@cam.ac.uk

Abstract

Model compensation is a standard way of improving speech recognisers' robustness to noise. Most model compensation techniques produce diagonal covariances. However, this fails to handle changes in the feature correlations due to the noise. A scheme will be presented that allows full covariance matrices to be estimated. One problem is that full covariance matrix estimation will be more sensitive to approximations, like those for dynamic parameters which are known to be crude. Here, a linear transformation of a window of consecutive frames is used as the basis for dynamic parameter compensation. A second problem is that the resulting full covariance matrices slow down decoding. This is addressed by using predictive linear transforms that decorrelate the feature space, so that the decoder can then use diagonal covariance matrices. On a noise-corrupted Resource Management task, the proposed scheme outperformed the standard VTS compensation scheme.

A3.

Title: A Shrinkage Estimator for Speech Recognition with Full Covariance HMMs

Authors: Peter Bell and Simon King

Address: The Centre for Speech Technology Research
University of Edinburgh
Informatics Forum
10 Crichton Street
Edinburgh EH8 9AB
United Kingdom

Email: Peter.Bell@ed.ac.uk

Abstract

We consider the problem of parameter estimation in full-covariance Gaussian mixture systems for automatic speech recognition. Due to the high dimensionality of the acoustic feature vector, the usual sample covariance matrix (obtained via the EM algorithm) has a high variance and is often poorly-conditioned when the amount of training data is limited.

In this paper we propose the use of a shrinkage estimator to solve these problems. The shrinkage estimator combines two estimators of differing dimensionality so as to optimise a bias-variance trade-off.

We explain how the shrinkage estimator can be obtained in the context of an HMM-GMM system, and derive a formula for the optimal shrinkage intensity. We present results of experiments on a phone recognition task, controlling the amount of training data available. Our results show that the shrinkage estimator gives a performance improvement over a standard full-covariance system. We also compare the performance with a system using Semi-tied Covariance Matrices, a common technique for covariance matrix dimensionality reduction.

A4.

Title: Longitudinal study of ASR performance on ageing Voices

Authors: Ravichander Vipperla, Steve Renals and Joe Frankel

Address: The Centre for Speech Technology Research
University of Edinburgh
Informatics Forum
10 Crichton Street
Edinburgh EH8 9AB
United Kingdom

Email: r.c.vipperla@sms.ed.ac.uk

Abstract

With ageing, human speech production mechanism undergoes several anatomical changes. This leads to notable changes in the elderly voices. Typical characteristics of ageing voices include changes in the fundamental frequency, increased jitter and shimmer, increased breath in speech, and slower speaker rate. It is of interest to see if these changes significantly effect the performance of Automatic Speech Recognition (ASR) systems.

This study presents the results of longitudinal performance of automatic speech recognition on ageing voices. Experiments were conducted on the audio recordings of the proceedings of the Supreme Court Of The United States (SCOTUS) corpus. This corpus has a good representation of elderly voices over 60 years of age and data for some speakers is available over a period of few years to allow a longitudinal study.

Results of the experiments show that the ASR Word Error Rates (WER) for elderly voices are significantly higher than those of adult voices. The WER increases gradually as the age of the elderly speakers increases. Use of MLLR based speaker adaptation on ageing voices improves the WER but the performance is considerably lower compared to adult voices. Speaker adaptation however reduces the rate of WER increase with age during old age.

A5.

Title: 3D video analysis of emotional speech

Authors: Nataliya Nadtoka, James Edge, Adrian Hilton and Philip JB Jackson

Address: Centre for Vision Speech and Signal Processing
University of Surrey
Guildford
Surrey GU2 7XH
United Kingdom

Email: N.Nadtoka@surrey.ac.uk

Abstract

The aim of the study is to learn the relationship between facial movements and the acoustics of speech sounds. We recorded A database of 3D video of the face, including markers, and corresponding synchronized audio of a single speaker. The database consists of 110 English sentences. These sentences were selected for strong expressive content in the fundamental emotions: Anger, Surprise, Sadness, Happiness, Fear and Disgust. Comparisons are made with the same sentences with neutral expression. Principal component analysis of the marker movements was performed to identify significant modes of variation. The results of this analysis show that there are various characteristic difference between visual features of emotional versus neutral speech. The findings of the current research provide a basis for generating realistic animations of emotional speech for applications such as computer games and films.

A6.

Title: Audiovisual Emotion Recognition in an English Database

Authors: Sanaul Haq, Philip JB Jackson and James Edge

Address: Centre for Vision Speech and Signal Processing
University of Surrey
Guildford
Surrey GU2 7XH
United Kingdom

Email: S.Haq@surrey.ac.uk

Abstract

Human communication is based on verbal and nonverbal information, e.g., facial expressions and intonation cue the speaker's emotional state. Important speech features for emotion recognition are prosody (pitch, energy and duration) and voice quality (spectral energy, formants, MFCCs, jitter/shimmer). For facial expressions, features related to forehead, eye region, cheek and lip are important. Both audio and visual modalities provide relevant cues. Thus, audio and visual features were extracted and combined to evaluate emotion recognition on a British English corpus. The database of 120 utterances was recorded from an actor with 60 markers painted on his face, reading sentences in seven emotions (N=7): anger, disgust, fear, happiness, neutral, sadness and surprise. Recordings consisted of 15 phonetically-balanced TIMIT sentences per emotion, and video of the face captured by a 3dMD system. A total of 106 utterance-level audio features (prosodic and spectral) and 240 visual features (2D marker coordinates) were extracted. Experiments were performed with audio, visual and audiovisual features. The top 40 features were selected by sequential forward backward search using Bhattacharyya distance criterion. PCA and LDA transformations, calculated on the training data, were applied. Gaussian classifiers were trained with PCA and LDA features. Data was jack-knifed with 5 sets for training and 1 set for testing. Results were averaged over 6 tests.

The emotion recognition accuracy was higher for visual features than audio features, for both PCA and LDA. Audiovisual results were close to those with visual features. Higher performance was achieved with LDA compared to PCA. The best recognition rate, 98%, was achieved for 6 LDA features (N-1) with audiovisual and visual features, whereas audio LDA scored 53%. Maximum PCA results for audio, visual and audiovisual features were 41%, 97% and 88% respectively. Future work involves experiments with more subjects and investigating the correlation between vocal and facial expressions of emotion.

A8.

Title: Joint Quantization Strategies for Low Bit-Rate Sinusoidal Coding

Authors: Emre Unver, Stephane Villette and Ahmet Kondoz

Address: Centre for Communication Systems Research
University of Surrey
Guildford
Surrey GU2 7XH
United Kingdom

Email: E.Unver@surrey.ac.uk

Abstract

Although there exist speech coding standards producing high quality speech above 4 kbps there is still room for improvement at lower bit rates, especially at 2.4 kbps and below. Such low bit rate speech coders, which are usually based on a sinusoidal model, are of particular interest for military and security applications. The saved bandwidth may be reallocated for channel coding for example. Alternatively, having a very low bit rate speech coder may allow more discrete and less power-demanding transmission, or be useful in alternative applications, such as watermarking or steganography.

High quality speech coding at very low bit rates can be achieved by relaxing the delay constraint and combining several frames together in a metaframe. This allows correlation between consecutive speech frames to be exploited more efficiently. In this work, experiments have been carried out to optimize the metaframe size as a trade-off between quantization gains and buffering delay, and it has been found that metaframes consisting of 5 frames of 20 ms speech yield optimum performance. Moreover, it has been observed that metaframes can contain different types of speech , which have different characteristics and therefore coding requirements. In this work, a number of metaframe classes have been defined, and quantization schemes as well as bit allocation according to the needs of each metaframe class have been devised. A new method for voicing determination from the spectral envelope shape is also presented which can be used to lower the bit rate even further by not transmitting any voicing parameter.

The proposed techniques have been applied to the SB-LPC vocoder to produce speech at 1200 and 800 bps, and it has been found that the quality is comparable to the standard versions at 2400 and 1200 bps. A listening test comparing the SB-LPC vocoder with the proposed techniques to some standard low bit rate coders, such as MELP, has been carried out.

A9.

Title: Non-acoustic Speech Interface

Authors: Sertan Kaymak, Stephane Villette and Ahmet Kondoz

Address: Centre for Communication Systems Research
University of Surrey
Guildford
Surrey GU2 7XH
United Kingdom

Email: S.Kaymak@surrey.ac.uk

Abstract

Several speech interfaces used to synthesize speech from articulatory measurements in real time have been proposed in the literature. These measurements can be obtained from EMG/EPG measurements, a ”non audible murmur microphone” signal (NAM) or from ultrasound tongue and optical lip images. These interfaces are known as silent speech interfaces and are used for voice communication in noisy environments, in situations where silence must be maintained, or as an alternative to tracheo-oesophageal speech.

Speech synthesis from ultrasound tongue and optical lip images is the focus of this work. Synthesizing speech is based on creating audiovisual databases which include features extracted from tongue and lip images as well as features extracted from acoustic signal. Classification methods are trained to map features of images to features of the acoustic signal. After training, these classification methods are used to determine acoustic units from features extracted from images. A small audiovisual database was built and some measurements were done on both images and the corresponding acoustic signal. The tongue surface on ultrasound images was represented by the deformable model. Similarly, the Line Spectral Frequencies of speech signal segments were calculated on the acoustic speech signal. In future work, a larger database will be constructed and HMM-based stochastic models will be trained on this database. These trained models will be applied to acquired ultrasound and lip images to synthesize speech in real time.

A10.

Title: Unsupervised Feature Vector Normalization with Combined Standard and Throat Microphones for Robust ASR

Authors: Luis Buera, Antonio Miguel, Oscar Saz, Alfonso Ortega and Eduardo Lleida

Address: Toshiba Research Europe Ltd.
St. George House
1st Guildhall Street
Cambridge CB2 3NH

Email: buera@unizar.es

Abstract

When training and testing acoustic conditions differ, the accuracy of speech recognition systems rapidly degrades. To compensate this mismatch, classical solutions have been developed, such as feature vector normalization/adaptation methods or acoustic model adaptation methods. However, another possible approach consists on complementing the standard microphone signal with robust additional signals.

In this work, we propose an on-line unsupervised compensation technique for robust speech recognition that combines standard and throat microphone feature vectors. The solution, called Multi-Environment Model-based LInear Normalization with Throat microphone information, MEMLINT, is an extension of MEMLIN formulation. Hence, standard microphone noisy space and throat microphone space are modelled as GMMs and a set of linear transformations are learnt for each pair of Gaussians (one for each GMM) using training stereo data. To compensate some kinds of degradation which are not considered in MEMLINT we propose to use jointly an on-line unsupervised acoustic model adaptation method based on rotation transformations over an expanded HMM-state space. To include the new expanded acoustic models in the search algorithm, augMented stAte space acousTic dEcoder, MATE, is used.

Some results with an own recorded database show the effective performance of the proposed technique with respect to single microphone robustness techniques, obtaining very competitive results even in unseen conditions: 6.24% average WER for all SNRs (0dB, 5dB, 10dB, 15dB, 20dB and clean) with respect to the baseline (46.33% average WER).

A11.

Title: Audio-based alignment of pre-production video

Authors: Siripinyo Chantamunee and Yoshihiko Gotoh

Address: ICOSS, 219 Portobello Street
Sheffield, S1 4DP

Email: S.Chantamunee@dcs.shef.ac.uk

Abstract

Video is a multimodal data that combines audio and visual information. Recent applications, such as video retrieval, copy detection, and video summarisation, may require comparison of contents to find any similarities between multiple streams of video. This work is concerned with one approach to similarity detection. In particular, we aim to develop a method that aligns multiple streams of video based on their audio similarities.

Pre-production video is used for development. It is a raw material that is unedited and used to produce a video stream such as a movie or a television programme [1]. It often contains unedited sound and many repetitive visual frame sequences. They are not copies but retakes of the same scene; thus there exists some differences between shots such as a camera angle, person's speech and action. Most recent works were based on the similarity in visual information [2]. They may not always handle this problem well – for example, a camera angle can be different, or a person can face in a different direction. On the other hand, retake shots may contain a similar, if not identical, sequence of audio sounds.

In this work we investigate a statistical approach to audio-based video shot alignment. The method utilises Gaussian mixture model to synchronise multiple audio streams of repetitive shots. It requires to maintain inter-frame correlation, however spatial and temporal information is eliminated. Instead the method makes the maximum likelihood estimation of synchonised frames locally. We have evaluated the algorithm on the pre-production video summarisation task [1]. Experiment indicates that, based solely on audio, the shot alignment method achieved high scores for pleasant summary creation and redundancy removal.

1. Over et al. (2007) ACM Workshop on Video Summarization, Augusburg.
2. Shrestha et al. (2006) ACM Multimedia, Santa Barbara.

A12.

Title: Tense and lax vowels as a consequence of speech embodiment

Authors: Piers Messum

Email: p.messum@gmail.com

Abstract

When a healthy adult is speaking conversationally his speech breathing provides a stable platform of subglottal pressure and a flexible source of airflow. It appears to be an ideal power supply and we are able to largely ignore it in theoretical work and modelling.

However, the aerodynamics and respiratory mechanics of child speech are very different from the adult model. Aerodynamic variables do not scale linearly, but in opposing directions: the pressures in child speech are higher, but flows are similar, and airways are smaller. Mechanically, a child's respiratory drive is pulsatile rather than smooth. This is heightened in stress-accent languages - like English - where a young speaker must reinforce pulses for greater loudness on stressed syllables.

Unsurprisingly, a pulsatile power supply conditions events downstream of the respiratory system. One result is the appearance of the tense and lax vowel classes of English.

Lax vowels contrast with tense vowels as follows: (1) they are always ”checked” by a following consonant, (2) they require only moderate displacement of the tongue from its resting position, and (3) they are ”short”. If these characteristics were arbitrary and independent - as widely believed - then it would be a remarkable coincidence that each divides the vowel inventory into classes with the same membership.

Instead, these characteristics emerge together under constraints in child speech that do not appear in the adult model. These require a child to check vowels made with open articulations and lengthen those made with close ones. The former behaviour protects the subglottal pressure head. The latter is the indirect result of limiting airflow to avoid unwanted turbulent noise at the point of maximum oral constriction. Increasing laryngeal resistance to do this prolongs the time it takes to dissipate a pulse.

A13.

Title: The speech of our cousins and our ancestors

Authors: Antoine Serrurier and Anna Barney

Address: Institute of Sound and Vibration Research
University Road, Highfield
Southampton S017 1BJ

Email: as4@isvr.soton.ac.uk

Abstract

For 40 years the speech production capabilities of our ancestors have been subject to many studies and debates in the speech research community. The speech production with a human-like articulated vocal tract is dependant on several factors: (1) the anatomy of the vocal tract, (2) the mobility of the various articulators of the vocal tract and (3) the conscious control of these various articulators. Human beings are nowadays the only species able to speak, i.e. the only species having at the same time the adapted anatomy, a wide mobility of the articulators and a conscious control of them. Chimpanzees, the closest primates to humans in terms of evolution, lack this ability. After decades of study, however, researchers are still divided on the causes of this incapacity: while some ascribe it to a lack of mobility of the articulators, others argue in favour of a lack of control only. As part of a wider EU project called HandToMouth, this work intends to bring some new articulatory modelling inputs in order to contribute to this debate. More particularly, articulatory models of the vocal tract of human and nonhuman primates based on feeding tasks are being developed. The objective of this work is to determine whether the lack of mobility or control is the reason for the lack of speech capability of our closest cousins the chimpanzees. In the future, the developed articulatory models will be deformed and fitted to early hominids like Neanderthals to determine more precisely the range of sounds potentially produced by the human ancestors.

A14.

Title: Speech recognition using non-linear trajectories in a formant-based articulatory layer of a multiple-level segmental HMM

Authors: Hongwei Hu and Martin J Russell

Address: Multi-modal Interaction Laboratory
School of Electronic, Electrical and Computer Engineering
The University Of Birmingham
Edgbaston, Birmingham B15 2TT
United Kingdom

Email: HWH400@bham.ac.uk

Abstract

This paper is concerned with a multiple-level segmental HMM (MSHMM), in which the relationship between the symbolic (phonetic) and surface (acoustic, e.g. MFCC) representations of a speech signal is regulated by an intermediate articulatory-based representation. Up until now, the articulatory layer has been purely based on formant frequencies and speech dynamics are modeled as piecewise constant linear trajectories in the articulatory space. These trajectories are transformed into the acoustic space using a set of one or more linear articulatory-to-acoustic mappings. The resultant model is referred to as a linear/linear MSHMM. However, it has been shown that a linear/linear system is inadequate for speech pattern modeling. The purpose of this research is to determine whether the use of smooth, non-linear formant trajectories in a multiple-level segmental HMM can result in improved speech recognition performance compared with a MSHMM using linear formant trajectories. In this paper, the non-linear formant trajectories are generated based on the 'trajectory HMM' method proposed by Tokuda et al. For consistency with the previous linear/linear MSHMM, the articulatory-to-acoustic mappings are also linear, but the mapping parameters are re-estimated based on the non-linear trajectories data. The new model is thus referred to as a non-linear/linear MSHMM. As an appropriate decoder for non-linear/linear MSHMMs is not yet available, the N-best rescoring paradigm is used to evaluate the performance of the non-linear formant trajectories. A 1,000 N-best list is generated using a set of standard decision-tree based triphone HMMs with HTK. This N-best list is rescored using both monophone MSHMMs (49 models) and triphone MSHMMs (926 models, including some biphones and monophones), combined with five different articulatory-to-acoustic mapping schemes. The rescoring results on TIMIT corpus show that the introduction of non-linear formant trajectories results in improvement on recognition phone accuracy compared with linear trajectories.

B1.

Title: Exploiting articulation in speech technology

Authors: Korin Richmond

Address: The Centre for Speech Technology Research
University of Edinburgh
Informatics Forum
10 Crichton Street
Edinburgh EH8 9AB
United Kingdom

Email: korin@cstr.ed.ac.uk

Abstract

Mainstream speech technology deals largely with an acoustic representation of speech in one form or another. This is natural, since the acoustic domain is where the speech signal exists in transmission between human speakers, and moreover we can conveniently record and generate acoustic signals. The acoustic speech signal, though, is generated in humans by the physical articulatory system, which offers an alternative to the acoustic representation normally dealt with in speech technology.

An articulatory representation of speech has certain attractive properties which may be exploited in modelling speech. For example, the articulators move relatively slowly and their movements are continuous; the mouth cannot "jump" from one configuration to a completely different one instantaneously. It has been widely proposed that taking into account the properties of the speech production system could improve speech processing methods by providing useful constraints. Many potential applications have been suggested: for example, low bit-rate speech coding, speech analysis and synthesis, automatic speech recognition and animated talking heads.

In this talk, I will briefly expand upon the potential for using articulatory representations in speech technology. I will then introduce two concrete examples of work on exploiting articulatory data and representations. First, I will present recent work on incorporating articulation into HMM-based speech synthesis, highlighting the advantages this can bring. Second, I will discuss the acoustic-articulatory inversion mapping. Here, for a given acoustic speech signal we aim to estimate the underlying sequence of articulatory configurations which produced it. This could break reliance on the availability of measured articulatory data for every speech signal. This is relevant because, ultimately, for an articulatory approach to be practical, it is likely we will need convenient and non-invasive access to an articulatory representation of speech.

B2.

Title: Instantaneous Unsupervised Adaptation using Discriminative Mapping Transforms

Authors: Chandra K Raut, Kai Yu and Mark JF Gale

Address: Cambridge University Engineering Department,
Baker Building, Room 502
Trumpington Street
Cambridge, CB2 1PZ.

Email: ckr21@cam.ac.uk

Abstract

Speaker or environmental adaptation is an important stage for HMM-based speech recognition systems. The most common approach for rapidly adapting model parameters to a particular speaker is to use linear transforms estimated with maximum likelihood (ML). Though discriminative criteria, such as minimum phone error, are commonly used to train the HMM parameters, their use for estimating adaptation transforms has been limited. This is because for unsupervised adaptation they are highly sensitive to errors in the supervision hypothesis. Another issue with unsupervised adaptation is that it is not normally possible to start adapting the models straightaway. Adaptation is delayed until robust parameter estimation is achieved. This prevents any adaptation gains for single utterances where there is limited data.

This work describes initial attempts to address both of these problems. To enable robust discriminative parameter estimation discriminative mapping transforms (DMTs) have been proposed. DMTs transform ML-based linear transforms into discriminative-like transforms. In contrast to standard discriminative approaches, DMTs are constrained to be the same for all speakers. Thus during test-set adaptation only the more robust ML speaker-specific transforms are required to be estimated. These are then transformed using the speaker-independent DMTs. Instantaneous adaptation can be achieved by using a Bayesian framework. This allows robust transforms, or distributions over transforms, to be estimated and used instantly. Combining these Bayesian motivated approaches with DMTs allows instantaneous unsupervised discriminative adaptation. The theoretical and implementation issues associated with this new approach will be discussed. The proposed method has been evaluated on a large vocabulary conversational telephone speech task. Preliminary results are encouraging with gains over the standard ML and more robust maximum a-posterior linear transform estimation approaches.

B3.

Title: Improving Automatic Phone Alignments for Speech Synthesis

Authors: Catherine Breslin

Address: Toshiba Research Europe Ltd.
St. George House
1st Guildhall Street
Cambridge CB2 3NH

Email: catherine.breslin@crl.toshiba.co.uk

Abstract

Unit selection speech synthesis systems concatenate phones from a large database of speech, so it is important for utterances in the database to be labelled with accurate phone boundaries. Obtaining these boundaries manually is time-consuming and requires expert knowledge. Hence, automatic methods for speech segmentation have been proposed. A common approach is to use HMMs to force-align the speech and thus obtain the phone boundaries.

In this work, the performance of an HMM based alignment system is investigated. In contrast to ASR, the word-level transcription of an utterance is known, and so the task becomes that of finding an accurate phonetic transcription and the corresponding boundaries. First, the selection of appropriate HMM topology is discussed. Then, additional features based on voicing and spectral change are considered for inclusion in the frontend.

Automatic alignment naturally leads to some errors, and it is useful to identify poorly aligned phones. A further post-processing stage can be applied to correct the errors, or they can simply be disregarded from the database. An analysis of errors shows that the HMM aligner is consistently poor for specific classes of boundary (for example, vowel to glide), suggesting that a post-processing stage for specific boundaries might prove useful. Additionally, measures based on duration and likelihood can be used to accurately identify some of the worst errors in the automatic alignment.

Judging the quality of an automatic segmentation remains an open question and, although objective measures can be calculated, it is normally necessary to carry out a listening preference test to assess the impact of the alignment on the voice quality. In this work, a measure based on the overlap between the automatic and manual labelling is used as an objective criterion, and a final listening preference test is carried out.

C1.

Title: Rapid Speaker Adaptation with Limited Data

Authors: Abeer Alwan

Address: Department of Electrical Engineering, UCLA
Speech Processing and Auditory Perception Laboratory
Box 951594, 405 Hilgard Ave.
Los Angeles, CA 90095-1594

Email: alwan@icsl.ucla.edu

Abstract

Maximum Likelihood Linear Regression (MLLR) is widely used in speaker adaptation due to its effectiveness and computational advantages. When the adaptation data are sparse, however, MLLR's performance degrades because of unreliable parameter estimation. We propose several techniques to address this issue.

In one method, spectral mismatch between training and test data is reduced by aligning formant peaks. Regression-tree based phoneme- and state-level spectral peak alignment is proposed for rapid speaker adaptation using linearization of the vocal tract length normalization (VTLN) technique.

Another speaker normalization technique is based on subglottal resonances. Since the subglottal airways do not change for a specific speaker, the subglottal resonances are independent of the sound type and remain constant for a given speaker. This context-free property makes the proposed method suitable for limited data speaker adaptation. This method is computationally more efficient than maximum-likelihood based VTLN.

Compared to MLLR, VTLN, and global peak alignment, improved performance can be obtained for both supervised and unsupervised adaptations for both medium vocabulary (the RM1 database) and connected digits recognition (the TIDIGITS database) tasks. Performance improvements are largest with limited adaptation data which is often the case for ASR applications and these improvements are shown to be statistically significant.

C2.

Title: Robust Lip-Tracking using Rigid Flocks of Selected Linear Predictors

Authors: Eng-Jon Ong and Richard Bowden

Address: Centre for Vision Speech and Signal Processing
University of Surrey
Guildford
Surrey GU2 7XH
United Kingdom

Email: e.ong@surrey.ac.uk

Abstract

This paper proposes a learnt data-driven approach to the accurate, real-time tracking of lip shapes using only intensity information i.e. grey-scale images. This has the advantage that constraints such as a-priori shape models or temporal models for dynamics are not required or used. Tracking the lip shape is simply the independent tracking of a set of points that lie on the lip's contour. This allows us to cope with different lip shapes that were not present in the training data and performs as well as other approaches that have pre-learnt shape models such as the AAM. Tracking is achived via linear predictors, where each linear predictor essentially linearly maps sparse template difference vectors to tracked feature position displacements. Multiple linear predictors are grouped into a rigid flock to obtain increased robustness. To achieve accurate tracking, two approaches are proposed for selecting relevant sets of LPs within each flock. Analysis of the selection results show that the LPs selected for tracking a feature point choose areas that are strongly correlated with that of the tracked target and that these areas are not necessarily the region around the feature point as is commonly assumed in LK based approaches. Experimental results also show that this method is comparable in performance to that of AAMs, despite being much simpler, both in the training and tracking phases, without any a priori shape information and with minimal training examples.

C3.

Title: Speech-driven Lip Motion Generation with a Trajectory HMM

Authors: Gregor Hofer, Junichi Yamagishi and Hiroshi Shimodaira

Address: The Centre for Speech Technology Research
University of Edinburgh
Informatics Forum
10 Crichton Street
Edinburgh EH8 9AB
United Kingdom

Email: G.Hofer@sms.ed.ac.uk

Abstract

Automatic speech animation remains a challenging problem that can be described as finding the optimal sequence of animation parameter configurations given some speech. In this paper we present a novel technique to automatically synthesise lip motion trajectories from a speech signal. The developed system predicts lip motion units from the speech signal and generates animation trajectories automatically employing a "Trajectory Hidden Markov Model". Using the MLE criterion, its parameter generation algorithm produces the optimal smooth motion trajectories that are used to drive control points on the lips directly. Additionally, experiments were carried out to find a suitable model unit that produces the most accurate results. Finally a perceptual evaluation was conducted, that showed that the developed motion units perform better than phonemes.

D1.

Title: Sound Source Localisation, Acoustic Source Separation and Binaural Audio Processing

Authors: Banu Gunel

Address: Centre for Communication Systems Research
University of Surrey
Guildford
Surrey GU2 7XH
United Kingdom

Email: B.Gunel@surrey.ac.uk

Abstract

Localising robot Sound Source Localisation. Sound source localisation based on the processing of microphone array signals has several applications, one of which is audio interaction for robotics. A real-time source localisation demo of a system developed in I-Lab will be shown. The demo consists of a virtual robot and a compact microphone array. The virtual robot follows a speaker by turning towards him/her as the person moves in the room while speaking.

Acoustic Source Separation. Separating simultaneously active sound sources to obtain individual signals is a well-known problem. Considered as a pre-processing stage for speech and speaker recognition algorithms, acoustic source separation has broad application areas. Although several techniques exist, real-time separation with a compact array is advantageous for most of these applications. This demonstration will show how multiple sound sources are separated in real-time providing interference-free signals using a system developed in I-Lab.

Binaural Audio Processing. This demonstration will show how sounds (e.g. a speech signal), can be reproduced in 3D. The listeners can change the direction of the speaker by clicking on a sphere while listening with headphones. The demonstration uses simplified head related transfer functions (HRTFs) for real-time operation.

D2.

Title: Quality Evaluation of Spatial Transmission and Reproduction by an Artificial Listener

Authors: Martin Dewhirst

Address: Centre for Vision Speech and Signal Processing
University of Surrey
Guildford
Surrey GU2 7XH
United Kingdom

Email: M.Dewhirst@surrey.ac.uk

Abstract

Most current perceptual models for audio quality have concentrated on the audibility of distortions and noises that mainly affect the timbre of reproduced sound. The QESTRAL project, however, is developing a model that is specifically designed to take account of distortions in the spatial domain, such as changes in source location, width and envelopment, and calculates a single measure of the perceived spatial quality. This demonstration uses a surround sound reproduction system to show typical spatial distortions and includes the results of an initial listening test experiment to determine the effect of the distortions on the perceived spatial quality.

D3.

Title: 3D Capture of Dynamic Speech Movements

Authors: James Edge

Address: Centre for Vision Speech and Signal Processing
University of Surrey
Guildford
Surrey GU2 7XH
United Kingdom

Email: J.Edge@surrey.ac.uk

Abstract

We present our work on dense 3D capture of facial movements during speech production. Facial geometry is reconstructed using a stereo capture system, markers on the skin are tracked and these are used to register data over time. Techniques have been developed to parameterize the dynamics of speech which recover the underlying cyclical structure of speech lip movements.

E1.

Title: Automatic Video Language Identification

Authors: Jacob L Newman and Stephen J Cox

Address: University of East Anglia
Norwich NR4 7TJ UK

Email: Jacob.Newman@uea.ac.uk

Abstract

Automatic Language Identification (LID) is the process of classifying utterances of speech with the spoken language they represent. Audio LID is a well established field of research and several successful techniques have been developed to solve the problems it presents. Here, Video LID is characterised by the use of visual features rather than audio features in the language recognition process, namely the face and lips. By contrast to audio LID, video LID is a new and sparse research topic, presently lacking the techniques and datasets necessary to facilitate good research, which this project aims to overcome. It is hoped that this research will benefit speech recognition of video where audio is not available or is inadequate for successful LID. Such an application is in multiuser systems where the appropriate language dependent speech recognizer must be selected to allow the user to interact by issuing speech commands.

The Video LID system developed here uses a Vector Quantization (VQ) approach, in which facial features are tracked with an Active Appearance Model (AAM). The collection of vectors produced for each video frame is clustered into a specified number of codewords or VQ symbols, forming a codebook. In training, an AAM frame vector can then be classified as a single VQ symbol from that codebook and bigram language models of those symbols are built for each language. In testing, the vectors produced from the AAMs are expressed in terms of VQ symbols, their bigram probabilities calculated for each language, and the K-nearest bigram probability training vectors classify the test utterance.

E2.

Title: A Generalised Derivative Kernel for Speaker Verification

Authors: Chris Longworth and Mark JF Gales

Address: Fallside Laboratory
Cambridge University Engineering Department,
Trumpington Street
Cambridge, CB2 1PZ.

Email: cl336@eng.cam.ac.uk

Abstract

An important aspect of SVM-based speaker verification systems is the choice of dynamic kernel. For the GLDS kernel, a static kernel is used to map each observation into a higher order feature space. Features are then obtained by taking a simple average over all frames. Derivative kernels, such as the Fisher kernel, use a generative model as a principled way of extracting a fixed set of features from each utterance. However, the model and features are defined using the original observations. Here, a dynamic kernel is described that combines these two approaches. In general, it is not possible to explicitly train a model in the feature space associated with a static kernel. However, by using a suitable metric with approximate component posteriors, this form of dynamic kernel can be computed. This kernel generalises the GLDS and derivative kernel as special cases and is also closely related to parametric kernels such as the GMM-supervector kernel. Preliminary results using this kernel will be presented on the 2002 NIST SRE dataset.

E3.

Title: Direct Posterior-Based Confidence Measurement for Spoken Term Detection

Authors: Dong Wang, Joe Frankel and Simon King

Address: The Centre for Speech Technology Research
University of Edinburgh
Informatics Forum
10 Crichton Street
Edinburgh EH8 9AB
United Kingdom

Email: dwang2@inf.ed.ac.uk

Abstract

Spoken term detection, or STD, is a new task defined by NIST, targeting at a fast retrieval for spoken terms from audio archives. A typical STD system first generates word or sub-word lattices, from which potential searching terms are picked out, and finally decisions are made based on the confidence of each detection.

Our work focuses on the aspect of confidence estimation. Conventional confidence measurements are based on indirect posterior probabilities, i.e., resorting to the Bayesian formula, calculate posteriors from conditional and prior probabilities, divided by the evidence of the whole sentence. Though other factors such as energy, number of frames, frame-wise likelihood can be integrated into a mixed confidence, the basic idea is a posterior based on the Bayesian formula, therefore can be categorised into 'indirect posterior approaches'.

In our work, we try to utilise a multiple layer perceptron (MLP) to generate posterior probabilities, which are immediately used as confidence of the detected terms. In the current implementation, we first train a 3-layer MLP with all phones as targets, and use this structure to generate posterior probabilities of each speech frame regarding to each phone. In lattice searching, these frame-wise posteriors are accumulated to phone-wise posteriors for each arc in the lattice. Finally we achieve the confidence of each detected term by accumulating the posteriors of all the phones in this term, divided by the number of frames.

Compared with indirect posterior approaches, our new method calculates posteriors from MLP directly, which seems more reliable then Bayesian based computing, considering some untrue assumptions. In addition, the frame-wise property in the new approach avoids the complex forward-backward computation for path and lattice evidence. Our experimental results show a 2 points improvement in terms of figure of merit (FOM) with the new approach comparing to the indirect posterior approach.

E4.

Title: HMM-based synthesis of child speech

Authors: Oliver Watts, Junichi Yamagishi, Kay Berkling and Simon King

Address: The Centre for Speech Technology Research
University of Edinburgh
Informatics Forum
10 Crichton Street
Edinburgh EH8 9AB
United Kingdom

Email: O.S.Watts@sms.ed.ac.uk

Abstract

HMM-based speech synthesis offers comparable quality to typical unit selection and concatenation systems. However, it also offers two important capabilities that concatenative methods do not: speaker adaptation, and an integrated data-driven method for dealing with missing units. In the work reported here, we explore the potential of both of these capabilities for synthesising child speech.

Speaker adaptation has become a key technique in many automatic speech recognition (ASR) systems. Methods from the Maximum Likelihood Linear Regression (MLLR) and Maximum A Posteriori (MAP) families are used to transform or adapt the parameters of already-trained HMMs, such that the likelihood of some adaptation data are increased. In speech synthesis, these methods have been used to adapt HMMs to a new speaker using a very limited amount of data - far less data than would be required to build a concatenative system, for example.

Any ”found” speech corpus will inevitably suffer from a greater proportion of missing units than a corpus designed for speech synthesis. When synthesising a sentence that requires a missing unit (e.g., a particular diphone, or a diphone in a certain context), concatenative systems must substitute another unit; typically heuristics are used to select the substitute unit. On the other hand, HMM-based speech synthesis is able to construct a model for any missing unit, by sharing its parameters with existing models. This is achieved using similar data-driven, tree-based state clustering techniques to those used in ASR.

Together with the robustness of HMM-based speech synthesis to imperfect recording conditions, these capabilities are well suited to the task of child speech synthesis. We report an experiment comparing several configurations of an HMM-based speech synthesiser for child speech. We compared speaker-dependent and speaker-adaptive modelling for varying amounts of data in a listening test which evaluated similarity to the target speaker, naturalness and intelligibility. Since there may also be challenges in F0 tracking, spectral estimation and vocoding of child speech, we also evaluated natural and vocoded speech alongside the synthetic speech.

Examples of synthetic speech can be found at: `http://homepages.inf.ed.ac.uk/s0676515/child_speech/`.

E5.

Title: Articulatory resynthesis of EMA data

Authors: Ingmar Steiner and Korin Richmond

Address: The Centre for Speech Technology Research
University of Edinburgh
Informatics Forum
10 Crichton Street
Edinburgh EH8 9AB
United Kingdom

Email: isteiner@inf.ed.ac.uk

Abstract

The 3-D articulatory synthesizer VocalTractLab (http://vocaltractlab.de) allows high-quality speech synthesis with full control over various phonetic parameters, such as intonation, articulatory effort, and voice quality, without degrading the rendered waveform. It uses vocal tract parameters and phone definitions derived from magnetic resonance imaging (MRI) data of a native speaker of German. However, the multi-tier gestural scores which serve as the synthesis input must be skillfully hand-crafted to achieve the high degree of naturalness.

To work around this difficulty, in the work we present here, we use a corpus of speech recorded with electromagnetic articulography (EMA) to automatically derive a corpus of gestural scores suited to resynthesizing the original utterances with VocalTractLab. The data generated by tracking the motion of the articulators of the vocal tract model is correlated with the original EMA trajectories, which yields optimal gestural scores, despite the differences in speaking style and vocal tract anatomy between the speakers used for the parameter configuration of the vocal tract model and EMA data, respectively. The resynthesis gestural scores can then be used to train a statistical model which predicts gestural durations for unseen utterances. This will allow a modular text-to-speech platform to generate gestural scores directly and use VocalTractLab as an articulatory synthesis engine.

Further work will be carried out to adapt VocalTractLab to English by recording MRI and/or ultrasound data for a native English speaker, for whom an EMA corpus is already available. Using the same speaker for both the vocal tract parameter configuration and resynthesis of articulatory training data is expected to enhance the final quality of the synthesis pipeline.

E6.

Title: A Novel Multi stage Approach For Blind Separation of Convolutive Speech Mixtures

Authors: Tariqullah Jan and Wenwu Wang (University of Surrey) and DeLiang Wang (Ohio State University)

Address: Centre for Vision Speech and Signal Processing
University of Surrey
Guildford
Surrey GU2 7XH
United Kingdom

Email: t.jan@surrey.ac.uk

Abstract

Human listeners show remarkable ability to segregate target speech from complex auditory mixtures, such as in a cocktail party environment. However, it remains extremely challenging for machines to replicate even part of such functionalities. One promising technique is to address this problem using blind source separation where the mixing process is described as a linear convolutive model, and independent component analysis (ICA) can then be applied to separate the convolutive mixtures either in the time-domain or in the transform domain. A recent technique, called ideal binary mask (IBM), has shown promising properties in suppressing interference and improving intelligibility of target speech. This simple yet effective approach offers great potential for improving speech separation performance of ICA algorithms.

Here we propose a novel algorithm for the separation of convolutive speech mixtures using binaural recordings, based on the combination of ICA and IBM, together with a post-filtering process in the cepstral domain. Essentially, the proposed algorithm consists of three steps. First, a constrained convolutive ICA algorithm is applied to binaural recordings. Listening tests show that the separated target speech contains a considerable amount of interference from other sources. In order to reduce the interference, in the second step, we use IBM to process the outputs from the previous step. Listening tests from this step suggest that estimated IBM can considerably improve the separation performance by suppressing the interference to a much lower level. However, a typical problem occur in this step is the musical noise. The third and also last step in our algorithm is therefore to remove musical noise using cepstral smoothing. The proposed algorithm is observed to have improved performance for convolutive recordings with reverberation time up to 100 ms.

E7.

Title: Towards deriving compact and meaningful articulatory representations: an analysis of feature extraction techniques

Authors: Veena D Singampalli and Philip JB Jackson

Address: CVSSP, University of Surrey, Guildford
Surrey GU2 7XH, United Kingdom

Email: V.Singampalli@surrey.ac.uk

Abstract

We present an analysis of linear feature extraction techniques to derive a compact and meaningful representation of the articulatory data. We used 14-channel EMA (ElectroMagnetic Articulograph) data from two speakers from the MOCHA database[1]. As representations, we considered the registered articulator fleshpoint coordinates, transformed PCA (Principal Component Analysis) and LDA (Linear Discriminant Analysis) features. Various PCA schemes were considered, grouping coordinates according to correlations amongst the articulators. For each phone, critical dimensions were identified using the algorithm in [2]: critical articulators with registered coordinates, and critical modes with PCA and LDA. The phone distributions in each representation were modelled as univariate Gaussians and the average number of critical dimensions was controlled using a threshold on the 1-D Kullback Leibler divergence (identification divergence). The 14-D KL divergence (evaluation divergence) was employed to measure goodness of fit of the models to estimated phone distributions. Phone recognition experiments were performed using coordinate, PCA and LDA features, for comparison.

We found that, of all representations, the LDA space yielded the best fit between the model and phone pdfs. The full PCA representation (including all articulatory coordinates) gave the next best fit, closely followed by two other PCA representations that allowed for correlations across the tongue. At the threshold where average number of critical dimensions matched those obtained from IPA, the goodness of fit improved by 34% (22%/46% for male/female data) when LDA was used over the best PCA representation, and by 72% (77%/66%) over articulatory coordinates. For PCA and LDA, the compactness of the representation was investigated by discarding the least significant modes. No significant change in the recognition performance was found as the dimensionality was reduced from 14 to 8 (95% confidence t-test), although accuracy deteriorated as further modes were discarded. Evaluation divergence also reflected this pattern. Experiments on LDA features increased recognition accuracy by 2% on average over the best PC representation. An articulatory interpretation of the PCA and LDA modes is discussed. Future work focuses on articulatory trajectory generation in feature spaces guided by the findings of this study.

1. A.A. Wrench. A new resource for production modelling in speech technology. In Proc. Inst. of Acoust., Stratford-upon-Avon, UK, 2001.
2. Veena D Singampalli and Philip JB Jackson. Statistical identification of critical, dependent and redundant articulators. In Proc. Interspeech, Antwerp, Belgium, pages 70-73, 2007.

E8.

Title: Parallel model combination and digit recognition with soccer audio

Authors: Jack Longton and Philip JB Jackson

Address: Centre for Vision Speech and Signal Processing
University of Surrey
Guildford
Surrey GU2 7XH
United Kingdom

Email: J.Longton@surrey.ac.uk

Abstract

Audio from broadcast soccer can be used for identifying highlights from the game. We can assume that the basic construction of the auditory scene consists of two additive parallel audio streams, one relating to commentator speech and the other relating to audio captured from the ground level microphones. Audio cues derived from these sources provide valuable information about game events, as can the detection of key words used by the commentators, which are useful for identifying highlights. We investigate word recognition in a connected digit experiment providing additive noise that is present in broadcast soccer audio. A limited set of background soccer noises, extracted from the FIFA World Cup 2006 recordings, were used to create an extension to the Aurora-2 database. The extended data set was tested with various HMM and parallel model combination (PMC) configurations, and compared to the standard baseline, with clean and multi-condition training methods. It was found that incorporating SNR and noise type information into the PMC process was beneficial to recognition performance with a reduction in word error rate from 17.5% to 16.3% over the next best scheme when using the SNR information.Future work will look at non stationary soccer noise types and multiple statenoise models.

E9.

Title: Voice Impersonation and Speaker Recognition

Authors: Zargham Haider, Stephane Villette and Ahmet Kondoz

Address: Centre for Communication Systems Research
University of Surrey
Guildford
Surrey GU2 7XH
United Kingdom

Email: Z.Haider@surrey.ac.uk

Abstract

Voice-based recognition systems offer a number of advantages over other biometric recognition systems. Speaker Identification (SID) Systems offer good performance in the case of clean speech, and most of the current ongoing research aims at improving their reliability under various operating conditions. Error rates down to 1% can be achieved with the existing SID systems, making them suitable for real-life applications. However, in most of the existing research material on SID systems, it is assumed that the speakers make no attempts at disguising their voices. With the advancements in speech synthesis and voice transformation techniques, it is highly likely that SID will be attacked by an impostor using these synthesis and transformation techniques to mimic another speaker. The SID systems have been shown to falter under such deliberate attacks. The aim of this work is to analyze how various SID techniques can resist such deliberate attacks. To accomplish this task, an SID system which has previously been developed and tested, employing Gaussian Mixture Models (GMM) and using Mel-frequency Cepstrum Coefficients (MFCC) for representing the spectral properties of a speaker, has been used. An impersonation system is also being developed using the Line Spectral Frequencies (LSF) as the feature vectors, representing the spectral properties of the impostor and the target i.e. the speaker being mimicked. The correspondences between the feature vector spaces of the two speakers are represented as a mapping function, which is used to map the spectral properties of impostor onto that of the target. The impersonation system is being used to intrude the SID and the results of the intrusion being analyzed. This knowledge will be used to develop SID systems which will offer performance similar to that of the existing systems, and are either resilient to such attacks, or are able to detect whether an impersonation is being attempted.

E10.

Title: Index assignment based channel coding for speech/audio communications

Authors: Huseyin Oztoprak, Stephane Villette and Ahmet Kondoz

Address: Centre for Communication Systems Research
University of Surrey
Guildford
Surrey GU2 7XH
United Kingdom

Email: h.oztoprak@surrey.ac.uk

Abstract

Despite the advances in multimedia coding which enabled the realization of real time multimedia communications streaming in an error prone mobile network, the end quality of a real time multimedia transmission system depends to a large extent on the resillience of the coder to the bit errors. Channel coding, i.e. adding a number of redundant bits to the source bits, is widely used to increase the robustness.

A scheme called named Index Assignment based Channel Coding (IACC) has been developed for resilience against the bit errors commonly seen in wireless channels. Different parameters of coded speech may have varying sensitivities against the channel errors. This has been efficiently exploited by Unequal Error Protection (UEP) schemes which provide different levels of protection for the different parameters of the coded media. The proposed scheme takes this approach further and adjusts the amount of protection according to the sensitivity of the different values of the source parameters. This scheme has been evaluated using the AMR-WB+ audio codec under random error conditions. Results show that the performance of the proposed scheme concatenated with convolutional coding can be superior to that of conventional convolutional coding, especially at high BERs.

E11.

Title: Synthesising personalised voices for individuals with progressive speech loss

Authors: Sarah Creer, Phil Green and Stuart Cunningham

Address: ICOSS, 219 Portobello Street
Sheffield, S1 4DP

Email: S.Creer@dcs.shef.ac.uk

Abstract

Can Speech Technology help individuals with speech disorders to interact more easily, reducing the impact of their disability? Many individuals with speech disorders use voice output communication aids (VOCAs), which 'speak' synthesised or pre-stored phrase digitised voice output. The aim of this work is to investigate techniques to capture voices that will be lost or are deteriorating due to progressive speech disorders for use with a personalised speech synthesiser for a communication aid.

The voice is an identifier of the person to whom it belongs. It provides clues about the gender, age, size, ethnicity and geographical identity of the person along with identifying them as a particular individual to family members, friends and, once interaction has begun, to new communication partners. Maintaining that vocal identity will help the ease of interaction and preservation of social relationships. VOCAs currently only allow voice personalisation of gender, language and to a limited extent, age.

The work presented will show the progress being made in this area using Hidden Markov Model (HMM) synthesis (HTS-AVSS toolkit), taking small amounts of dysarthric speech data and adapting average voice HMMs towards those of the target speaker. In order to not recreate the disordered speech patterns of the target speaker, duration and aperiodicity information from the average voice model can be imposed onto the adapted HMMs, which retain the adapted spectral and F0 features of the target speaker. This method has the potential to capture the essence of the voice of an individual whose speech is deteriorating and provide an intelligible and personalised speech synthesiser for such a user.

E12.

Title: Cross-language Voice Conversion: Analysis from a Phonetic Perspective

Authors: Kayoko Yanagisawa and Mark Huckvale

Address: Department of Speech, Hearing and Phonetic Sciences
UCL,Chandler House
2 Wakefield Street,
London WC1N 1PF

Email: k.yanagisawa@ucl.ac.uk

Abstract

Voice Conversion (VC) is a technique which aims to convert the identity of the speaker of a given signal without changing its linguistic content. It uses a mapping function between the source speaker and the target speaker trained from paired and aligned utterances. Cross-language VC has been explored as a way of achieving Spoken Language Conversion, whereby a speaker is made to appear to speak a text in a language other than the one it was originally uttered in. The problem with cross-language VC is how to generate parallel training utterances between two speakers who speak different languages. Typically, equivalence is assumed between speech materials produced by the two speakers, either in acoustic or phonetic terms. However, different languages have different phoneme inventories and they also exploit different allophonic variations. We hypothesize that parallel utterances based on such ”phonetic equivalence” would lead to inaccurate mapping and hence conversion.

This study investigates the issue of phonetic equivalence, by comparing the conversion performance of mono-lingual and cross-lingual VC systems in terms of intelligibility. Two systems were constructed to perform VC in Japanese from speaker S1 (Japanese TTS) to S2 (English-Japanese bilingual) so that the output should sound like S2 speaking Japanese. The mono-lingual system was trained using only Japanese sentences, while the cross-lingual system was trained using only English sentences. The English sentences were constructed to have the phonotactics of Japanese but English phonetic forms by making use of a phoneme equivalence table.

We show that VC impairs intelligibility in both conditions, but the effect is significantly worse in the cross-lingual system. A comparison of the phonetic confusions made by the two systems highlights a number of problems which may be attributed to the phonetic equivalence assumption underlying the cross-lingual training.

E13.

Title: The Development of the "TalkMaths" Speech Interface at Kingston University

Authors: Angela Wigmore, Gordon Hunter, Eckhard Pflgel and James Denholm-Price

Address: Penrhyn Road Campus
Faculty of CISM, Kingston University
Penrhyn Road Kingston-upon Thames, Surrey, KT1 2EE UK

Email: k0330947@kingston.ac.uk

Abstract

Computer speech recognition technology has now developed to a level such that the control of computer applications and devices by voice commands is now possible using technology that is becoming commonplace. This has the potential to enable people to use devices hands-free, which could be beneficial for a wide variety of reasons: convenience, safety (e.g. whist driving or performing a surgical operation) or necessity. There are numerous potential benefits for people with a wide range of disabilities. Furthermore, to make Information and Communications Technology (ICT) more accessible, recent government legislation in many countries has been introduced and "good practice" guidelines have been produced by bodies within the industry, such as the World-Wide Web Consortium's (W3C) Web Accessibility Initiative (WAI). However, in order for the benefits of this technology to be genuinely realised, it will need to be both reliable and easy to use.

Even for able-bodied people, typing and editing mathematical text can be very tedious, time-consuming and error prone - yet it is an essential part of the production of many scientific and technical documents. In this paper, we describe the details of the "TalkMaths" system, which is currently under development at Kingston University and will enable University staff and/or students to perform various computer-based tasks hands-free, including dictating mathematical expressions in the LaTeX and MathML languages for mathematical typesetting. This prototype system greatly simplifies the user's task by not requiring the user to have extensive knowledge of the syntax of either typesetting language. The prototype also has the potential to be used as an educational tool where students can practice their pronunciation of mathematical formulae and equations, as a means of learning the "correct" or "valid" ways of reading such expressions aloud and checking the output produced on a screen or printed page.

There will also be a short demonstration of the current system.

List of participants

Abeer Alwan, [alwan@icsl.ucla.edu], University of California, LA
Anna Barney, [ab3@soton.ac.uk], University of Southampton
Peter Bell, [peter.bell@ed.ac.uk], University of Edinburgh
Michael Berger, [m.a.berger@sms.ed.ac.uk], University of Edinburgh
Richard Bowden, [r.bowden@surrey.ac.uk], University of Surrey
Andrew Breen, [abreen@nuance.com], Nuance Communications
Catherine Breslin, [catherine.breslin@crl.toshiba.co.uk], Toshiba Research Europe
Luis Buera, [lbuera@unizar.es], Toshiba Research Europe
Omar Caballero, [a027295@uea.ac.uk], University of East Anglia
Siripinyo Chantamunee , [s.chantamunee@dcs.shef.ac.uk], University of Sheffield
Stephen Cox, [s.j.cox@uea.ac.uk], University of East Anglia
Sarah Creer, [s.creer@dcs.shef.ac.uk], University of Sheffield
Ian Cushing, [i.cushing@ucl.ac.uk], University College London
Martin Dewhirst, [m.dewhirst@surrey.ac.uk], University of Surrey
Rakkrit Duangsoithong, [r.duangsoithong@surrey.ac.uk], University of Surrey
James Edge, [j,edge@surrey.ac.uk], University of Surrey
Emma Greenland, [emmag@sandybrown.com], Institute of Acoustics
Jon Gudnason, [jon.gudnason@imperial.ac.uk], Imperial College
Banu Gunel, [b.gunel@surrey.ac.uk], University of Surrey
Zargham Haider, [z.haider@surrey.ac.uk], University of Surrey
Sana-ul Haq, [s.haq@surrey.ac.uk], University of Surrey
Gregor Hofer, [g.hofer@sms.ed.ac.uk], University of Edinburgh
Hongwei Hu, [hwh400@bham.ac.uk], University of Birmingham
Mark Huckvale, [m.huckvale@ucl.ac.uk], University College London
Gordon Hunter, [G.Hunter@kingston.ac.uk], Kingston University
Tariqullah Jan, [t.jan@surrey.ac.uk], University of Surrey
Adam Kalbassi, [m.kalbassi@ntlworld.com], University of Surrey
Sertan Kaymak, [s.kaymak@surrey.ac.uk], University of Surrey
Tim Kempton, [acp06tk@shef.ac.uk], University of Sheffield
Ilias Kolonias, [ik0001@surrey.ac.uk], University of Surrey
Joe Kornycky, [j.kornycky@surrey.ac.uk], University of Surrey
Yuxuan Lan, [yl@cmp.uea.ac.uk], University of East Anglia
Caroline Leathem, [caroline.leathem@cayton-consulting.com], Cayton Consulting
Ying Liu, [y.liu@bham.ac.uk], University of Birmingham
Jack Longton, [J.Longton@surrey.ac.uk], University of Surrey
Chris Longworth, [cl336@cam.ac.uk], Cambridge University
Piers Messum, [p.messum@gmail.com],
Ben Milner, [b.milner@uea.ac.uk], University of East Anglia
Roger Moore, [r.k.moore@dcs.shef.ac.uk], University of Sheffield
Nataliya Nadtoka, [N.Nadtoka@surrey.ac.uk], University of Surrey
Jacob L Newman, [Jacob.Newman@uea.ac.uk], University of East Anglia
Eng-Jon Ong, [e.ong@surrey.ac.uk], University of Surrey
Huseyin Oztoprak, [h.oztoprak@surrey.ac.uk], University of Surrey
Medha Pandit, [m.pandit@eim.surrey.ac.uk], University of Surrey
Norman Poh, [n.poh@surrey.ac.uk], University of Surrey
Chandra Raut, [ckr21@eng.cam.ac.uk], Cambridge University
Korin Richmond, [korin@cstr.ed.ac.uk], University of Edinburgh

Paul Rocca, [paul.rocca@audiosoft.co.uk], AudioSoft Ltd
Zoi Roupakia, [zr216@cam.ac.uk], Cambridge University
Antoine Serrurier, [A.Serrurier@soton.ac.uk], University of Southampton
Syed Raza Shahid, [syed.raza@spinvox.com], SpinVox: Speech Recognition Engineer
Veena D Singampalli, [v.singampalli@surrey.ac.uk], University of Surrey
Ingmar Steiner, [ingmar.steiner@ed.ac.uk], Centre for Speech Technology Research
Mark Thomas, [mark.r.thomas@imperial.ac.uk], Imperial College London
Emre Unver, [e.unver@surrey.ac.uk], University of Surrey
Rogier van Dalen, [rcv25@cam.ac.uk], Cambridge University
Stephane Villette , [s.villette@surrey.ac.uk], University of Surrey
Ravichander Vipperla , [r.c.vipperla@sms.ed.ac.uk], University of Edinburgh
Dong Wang, [dwang2@inf.ed.ac.uk], University of Surrey
WenWu Wang, [w.wang@syrrey.ac.uk], University of Surrey
Oliver Watts, [O.S.Watts@sms.ed.ac.uk], University of Edinburgh
Angela Wigmore, [k0330947@kingston.ac.uk], Kingston University
Siew Yeung Wong, [s.wong@surrey.ac.uk], University of Surrey
Junichi Yamagishi, [jyamagis@inf.ed.ac.uk], University of Edinburgh
Kayoko Yanagisawa, [k.yanagisawa@ucl.ac.uk], University College London
Kai Yu, [ky219@cam.ac.uk], Cambridge University
Dang Cong Zheng, [d.c.zheng@wlv.ac.uk], University of Wolverhampton