



www.ukspeech.org.uk

One Day Meeting for Young Speech Researchers

Thursday 12th April 2007

University College London

Programme and Abstracts of Posters and Talks

Organisers:

Prof. Stephen Cox
School of Computing Sciences
University of East Anglia
Norwich NR4 7TJ

sjc@cmp.uea.ac.uk

Dr Mark Huckvale
Department of Phonetics and Linguistics
University College London
Gower Street
London WC1E 6BT

M.Huckvale@ucl.ac.uk

Programme

Coffee available from 10.00

Poster Session 1 10.30–11.55

11.55 Welcome M. Huckvale, University College London

Talks Session 1 Session chair: Mark Huckvale, University College London

12.00 *Speech recognition in university and start-up.*

Tony Robinson, Cantab Research

12.30 *Bayesian Discriminative Adaptation for Speech Recognition*

Chandra Kant Raut, Kai Yu and Mark J. F. Gales, Cambridge University.

Talks Session II 14.15–15.30 Session chair: Stephen Cox, University of East Anglia

14.15 *Human Speech Acquisition and Processing*

Michael Carey, University of Birmingham

15:00 *Allowing for prosodic variation in evaluation against a gold standard
reference corpus*

Claire Brierley and Eric Atwell, University of Leeds

Poster Session II 15.30–17.00

Contents

Talks Session I abstracts.....	1
Talks Session II abstracts.....	3
Poster Session I abstracts.....	5
Poster Session II abstracts.....	16

TALKS SESSION I
12:00–13:00

One Day Meeting for Young Speech Researchers
Thursday 12th April 2007
University College London

T1.

Title: *Speech recognition in university and start-up*

Authors: Tony Robinson

Address: Cantab Research
St John's Innovation Centre
Cowley Road
Cambridge CB4 0WS

Email: tony@tonyRobinson.com

Abstract

This talk will explore the differences in working in the field of speech recognition in a university and start-up company environment. Whilst on the surface the work that is done can appear to be very similar, the people required, working attitude, how the work happens, the timescales the data used and the even the speech recognition task can be very different. It will draw on Tony Robinson's experiences in running research groups in Cambridge University and in taking a speech start-up company through from formation to buy out.

T2.

Title: *Bayesian Discriminative Adaptation for Speech Recognition*

Authors: Chandra Kant Raut, Kai Yu and Mark J. F. Gales

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

Email: ckr21@cam.ac.uk

Abstract

Speech recognition systems are normally adapted for varying acoustic conditions using linear transforms. The standard way to perform unsupervised mode adaptation is to use some initially generated hypothesis to estimate maximum-likelihood linear transforms, which are then used to adapt acoustic models for further recognition. However, this approach does not work well when there is only limited amount of adaptation data.

Bayesian adaptation framework has been investigated for maximum-likelihood based adaptation to overcome this problem. By using prior transform distribution, lower bound approximations, such as MAP and variational Bayes, have been found to significantly outperform the standard ML estimate in limited data case. However, the state-of-art systems widely use discriminative criteria such as MMI/MPE. Therefore, we investigate the formulation of Bayesian framework for discriminative adaptation case. The difficulty of this framework compared to the likelihood based Bayesian framework is that lower bound approximations to likelihood cannot be directly used to optimise the discriminative criteria. Several approximation approaches are investigated in this work to address the problems of Bayesian discriminative adaptation. They include extended Baum-Welch auxiliary function, strong-sense auxiliary function formulated by using Jensen and reverse-Jensen's inequalities, and second-order statistics constrained auxiliary function. The preliminary results of the investigation will be also presented.

T3.

Title: *Human Speech Acquisition and Processing*

Authors: Michael Carey

Address: Dept. of Electronic, Electrical & Computer Engineering,
University of Birmingham
Birmingham B15 2TT.

Email: M.Carey@bham.ac.uk

Abstract

The methods that humans use for processing speech and language is an exciting area of scientific enquiry. Since, almost without exception, humans are also far superior to machines in this task human speech processing is of considerable interest to engineers designing speech processing systems. A key feature of the human system is that it is learnt, we are not born talking. Hence any system proposed as an analogue of human speech processing must take this into account. However there are strong differences of opinion between those like Chomsky who believe some evolutionary pre-adaptation of the human brain is necessary for speech processing and those like Piaget who believe it is solely a consequence of human brain processing power. It's also important to keep in mind that language is acquired through speech and not text.

The approach described in this talk is to address this problem "bottom-up" starting with the newborn infant's problem of discriminating between speech and noise. We then address a possible method for acquiring the ability to discriminate between significant speech features. We also describe the accommodation of timescale variability using a multi-layered neural network model to recognize phonemes and words.

T4.

Title: *Allowing for prosodic variation in evaluation against a gold standard reference corpus.*

Authors: Claire Brierley and Eric Atwell

Address: School of Computing,
University of Leeds,
Leeds LS2 9JT

Email: claireb@comp.leeds.ac.uk

Abstract

An automatic phrase break prediction system aims to identify prosodic-syntactic boundaries in text which correspond to the way a native speaker might process or chunk that same text as speech. In computational linguistics, Machine Learning from hand-annotated corpus data has become the de-facto standard approach to text annotation problems such as prosodic annotation. This is treated as a classification task in machine learning and output predictions from language models are evaluated against ‘gold standard’ prosodic phrase break annotations in a speech corpus. Despite the application of rigorous metrics such as precision and recall, the evaluation of phrase break models is still problematic because prosody is inherently variable: a given linguist’s set of morphosyntactic analysis and prosodic annotations for a given text may not be fully representative of the range of parsing and phrasing strategies available to, and exhibited by, native speakers.

A fairer approach to evaluation requires POS tagged and prosodically annotated variants of a text to enrich the gold standard and enable more robust ‘noise-tolerant’ measurement of language models. We report on experiments with the AIX-MARSEC spoken English corpus. This has already been richly annotated at several linguistic levels, allowing a range of features to be applied in Machine Learning of phrase break prediction. We have developed a rule-based prosodic phrase break predictor (Brierley and Atwell forthcoming a,b) to enrich the phrase-break mark-up, to expand from a single linguist’s analysis to include a wider range of possible interpretations of the text. This allows for two different predictions to both score well if the prosody is plausible, even if the predicted phrase breaks differ from the corpus linguist’s analysis.

A1.

Title: *Automatic Head Motion Prediction from Speech Data*

Authors: Gregor Hofer and Hiroshi Shimodaira

Address: The Centre for Speech Technology Research,
University of Edinburgh,
2 Buccleuch Place
Edinburgh EH8 9LW

Email: g.hofer@sms.ed.ac.uk

Abstract

In this paper we present a novel approach to generate a sequence of head motion units given some speech. The modelling approach is based on the notion that head motion can be divided into a number of short homogeneous units that can be modelled individually. The system is based on Hidden Markov Models (HMM), which are trained on motion units and act as a sequence generator. They can be evaluated by an accuracy measure. A database of motion capture data was collected and manually annotated for head motion and is used to train the models. It was found that the model is good at distinguishing high activity regions from regions with less activity with accuracies around 75 percent.

Furthermore the model is able to distinguish different head motion patterns based on speech features somewhat reliably, with accuracies reaching almost 70 percent.

A2.

Title: *Acoustic Speech Feature Prediction From MFCC Vectors*

Authors: Jonathan Darch and Ben Milner

Address: School of Computing Sciences,
University of East Anglia,
Norwich, NR4 7TJ

Email: jonathan.darch@uea.ac.uk

Abstract

Acoustic speech features, namely fundamental frequency, formants and voicing class, are important parameters in speech processing and may be used for recognition, synthesis, enhancement and coding.

This work presents an analysis of correlation between acoustic features and mel-frequency cepstral coefficients (MFCCs) made both globally across all speech and within phoneme classes. The analysis leads to the development of two methods of 'predicting' acoustic speech features from MFCCs. The first method uses a Gaussian mixture model (GMM) to model the joint density of acoustic features and MFCCs. The second method exploits phoneme-specific correlations by employing GMMs for each model and state of a set of hidden Markov models (HMMs). An evaluation of prediction accuracy shows that the phoneme-dependent HMM-GMM system is more accurate than the simpler GMM system, which agrees with the correlation analysis.

Acoustic speech features predicted from MFCCs may be used to reconstruct speech within a distributed speech recognition (DSR) environment without the need to transmit fundamental frequency estimates, saving 800bps. There is only a minor degradation in the quality of the reconstructed speech when using predicted, rather than the estimated and transmitted speech.

A3.

Title: *Audio-Visual Speech Fragment Decoding*

Authors: Jon Barker and Xu Shao

Address: Speech and Hearing Research Group
Department of Computer Science,
The University of Sheffield,
Regent Court,
211 Portbello Street,
Sheffield S1 4DP

Email: x.shao@dcs.shef.ac.uk

Abstract

It is well understood that visual speech information can improve the performance of audio-based automatic speech recognition (ASR), particularly in conditions of significant acoustic noise. The conventional view is that visual and acoustic speech features combine to directly aid speech unit classification. What is less often considered is that the visual information may also have a role in *separating* the audio of the target speaker from the acoustic background. Improved audio separation may then lead to better speech recognition.

This paper presents a model called audio-visual speech fragment decoding (AV-SFD), in which the visual signal is used to aid both the separation and the recognition of the speech signal. The model builds on the existing audio-only SFD technique [1,2] which is based on the auditory scene analysis account of perceptual organisation. Primitive processes are used to identify sound fragments, and then model-driven processes search for fragments that can be grouped to make recognisable speech. In AV-SFD, the visual signal is being used in the latter grouping stage allowing the decoder to reliably distinguish between fragments of foreground and background -- even when the acoustic foreground and background are statistically similar.

AV-SFD has been evaluated using an audio-visual version of the simultaneous speaker task employed for the Interspeech'06 Speech Separation Challenge [3]. The video input is shown to resolve foreground/background ambiguities occurring in audio-only SFD, and the system considerably outperforms conventional multistream AV-ASR approaches across a wide range of SNR.

[1] Coy and Barker "An automatic speech recognition system based on the scene analysis account of auditory perception", Speech Communication (in press)

[2] Barker, Coy, Ma and Cooke, "Recent advances in speech fragment decoding techniques", Proc. Interspeech 2006, Pittsburg, PA, 85–88

[3] <http://www.dcs.shef.ac.uk/~martin/SpeechSeparationChallenge.htm>

A4.

Title: *Dynamic Kernels for Speaker Verification*
Authors: Chris Longworth and Mark Gales
Address: Cambridge University Engineering Department
Trumpington Street,
Cambridge CB2 1PZ
Email: cl336@cam.ac.uk

Abstract

Speaker verification is a binary classification task to determine whether a particular speaker uttered a phrase. Support Vector Machine classifiers have proved very effective for speaker verification. However it is necessary to specify an appropriate dynamic kernel to provide a mapping between the variable-length speech sequence and a fixed dimensional feature vector. Many widely-used dynamic kernels can be placed into one of two classes; parameter kernels, where the feature vector consists of parameters extracted from an utterance-dependent model, and derivative kernels, where the features are the derivatives of the utterance log-likelihood with respect to parameters of a generative model.

A5.

Title: *Visually-Derived Wiener Filters for Speech Enhancement*
Authors: Ibrahim Almajai and Ben Milner
Address: School of Computing Sciences,
University of East Anglia,
Norwich, NR4 7TJ
Email: ima@cmp.uea.ac.uk

Abstract

The aim of this work is to use visual speech information to enhance noise-contaminated audio speech. The work begins by examining the correlation between audio and visual speech features and reveals higher correlation to exist within individual phoneme sounds rather than globally across all speech. Utilising this correlation, a visually-derived Wiener filter is proposed in which clean power spectrum estimates are obtained from visual speech features. Two methods of extracting clean power spectrum estimates are made; first from a global estimate using a single Gaussian mixture model (GMM), and second from phoneme-specific estimates using a hidden Markov model (HMM)-GMM structure. Measurement of estimation accuracy reveals that the phoneme-specific (HMM-GMM) system leads to lower estimation errors than the global (GMM) system. Finally, the effectiveness of visually-derived Wiener filtering is examined.

A6.

Title: *Joint multigrams for dictionary generation for ASR*

Authors: Fiona Couper Kenney and Simon King

Address: The Centre for Speech Technology Research,
University of Edinburgh,
2 Buccleuch Place,
Edinburgh EH8 9LW

Email: f.kenney@ed.ac.uk

Abstract

We present a fully automatic method for learning dictionaries directly from word- and sub-word-unit transcriptions, and evaluate its use for automatic speech recognition. Joint multigrams are used to align word and sub-word sequences, and these are then used to generate dictionaries. Recognition experiments show that a dictionary produced automatically in this way outperforms a simple baseline dictionary with an absolute word error rate decrease of 3.44% (36% relative) and also achieve lower word error rates than a range of probabilistic dictionaries learnt from the same word- and sub-word-unit transcriptions in a supervised manner.

A7.

Title: *Building Multiple Complementary Systems using Directed Decision Trees*
Authors: Catherine Breslin and Mark Gales
Address: Cambridge University Engineering Department
Trumpington Street,
Cambridge CB2 1PZ
Email: cb404@cam.ac.uk

Abstract

Large vocabulary speech recognition often uses a multipass framework with a combination of multiple systems to obtain the final hypothesis. For the combination to give gains, the systems being combined must be complementary, i.e. they must make different errors. Often, complementary systems are chosen simply by training a variety of systems, performing all combinations, and selecting the best. This approach becomes time consuming as more potential systems are considered, and it is not guaranteed that any of the systems will be complementary. Hence, recent work has looked at explicitly building systems that are complementary to each other.

This work considers building complementary systems using directed decision trees. Decision trees are used to cluster states for parameter tying, and have the property that clustered states share the same output distribution.

Hence higher level information, such as language model and context, is needed to distinguish clustered states. Directed decision trees aim to build complementary systems by biasing the tree generation towards separating confusable states and so making different errors. This is done by using a second set of weighted statistics in the tree generation, where the training data is weighted to reflect confusions. The weighting is obtained by generating confusion networks for each training utterance, aligning these with the reference, and calculating a weight for each reference word based on its posterior probability in the confusion network.

Multiple complementary systems can be built within an iterative framework, like that used in boosting.

Experiments have been performed on a large vocabulary Broadcast News Arabic task. Two complementary systems have been built using directed decision trees, starting from an MPE trained baseline system. Decoding is performed within a multipass framework using CMLLR speaker adaptation. Combination of the three systems yield gains over the individual system performances alone.

A8.

Title: *Building personalised voices for speech synthesis for individuals with progressive speech loss*
Authors: Sarah Creer, Stuart Cunningham and Phil Green
Address: Speech and Hearing Research Group
Department of Computer Science,
The University of Sheffield,
Regent Court,
211 Portbello Street,
Sheffield S1 4DP
Email: s.creer@dcs.shef.ac.uk

Abstract

Applications of speech technology can be used to try and 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 use 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 that 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 maintenance of social relationships. VOCAs currently only allow voice personalisation of gender, language and to a limited extent, age. The requirements set out for the task of building a voice for a speaker whose own voice is about to or has already begun to deteriorate are: the amount of recording has to be minimal, the synthesiser has to have a small memory footprint and the output has to be intelligible, natural sounding and as similar to the individual's voice as possible.

This work will aim to investigate synthesis techniques to provide a personalised voice for those speakers with progressive disorders which will satisfy these requirements, particularly focussing on HMM synthesis using adaptation techniques (HTS toolkit) and concatenative synthesis with minimal data and introducing donor segments into the dataset (Festvox). Experiments have begun on building unimpaired voices using these toolkits using a phonetically balanced dataset (Arctic) and a plan of the next stages of this work will be presented.

A9.

Title: *Modelling speech dynamics using trajectory HMMs*
Authors: Hongwei Hu and Martin Russell
Address: Dept. of Electronic, Electrical & Computer Engineering,
University of Birmingham,
Birmingham, B15 2TT.
Email: hwh400@bham.ac.uk

Abstract

A Multiple-level Segmental HMM (M-SHMM) is a novel acoustic model in which the relationship between the symbolic and acoustic representation (e.g. MFCC) of a speech signal is regulated by an intermediate articulatory-based layer. The motivation for incorporating an intermediate layer is to better modeling speech dynamics and take good advantage of underlying sources of variability in speech. In previous work, the intermediate layer is based on formant parameters (Russell & Jackson, 2005). Speech dynamics are modeled as linear, piecewise constant trajectories and mapped to the acoustic layer by linear or non-linear articulatory-to-acoustic mappings.

A linear representation of speech dynamics is not sufficient to capture the variability of speech production. The objective of this research is to investigate better way to model speech dynamics in the intermediate layer. A trajectory HMM (Tokuda et al, 2002) is investigated and employed in this research. The motivation of trajectory HMMs is to overcome the piecewise constant and independent assumptions of standard HMMs. A trajectory HMM is derived by imposing explicit relationship between the static and dynamic features (e.g. delta and delta-delta coefficients) of a standard HMM. As a result, a smoothed, continuous trajectory is synthesized in the static feature space, which is most consistent with the static and dynamic features.

To investigate the synergy between speech recognition and synthesis, a basic speech synthesizer is built based on monophone HMMs. The models are trained on Parallel Formant Synthesizer (PFS) control parameters using TIMIT corpus. In the synthesis phase, two kind of speech parameters are generated, i.e piecewise constant control parameters based on HMM themselves and smoothed control parameters by applying trajectory HMM. Subjective evaluation are conducted to test the performance of trajectory HMMs on speech synthesis.

A10.

Title: *Natural language generation for service oriented chatbot systems.*

Authors: Marie Claire Jenkins and Stephen Cox

Address: School of Computing Sciences,
University of East Anglia,
Norwich, NR4 7TJ

Email: mcjenkins@gmail.com

Abstract

Service oriented chatbot systems are used to assist customers find information on large complex websites. The service oriented chatbot acts as a virtual customer service representative, giving answers to natural language queries and offering more targeted information during the course of its conversation with the user. Interaction between the user and the system happens via a chat window, similar to an instant messaging interface.

HCI experiments show that these systems are well accepted by users; however the level of expectation is quite high. The conversation needs to be as human-like as possible whilst also providing all of the relevant information. In order to produce the correct language and the correct behaviour it has been necessary to observe humans. This has led us to study in greater depth the interaction between humans in an online chat situation and humans and machines under the same circumstances.

To this date chatbot systems have relied heavily on “canned answers” stored in a database, and keyword matching in user queries. Large corporate websites change often and this method would be far too laborious to work adequately. We propose a method based on “fluid construction grammar” where the system is able to create its own sentences and extract the features necessary for information retrieval and sentence understanding from the user input.

“Construction Grammar” is metonymic: there is a pairing of form and content, and it evolves through user interaction. “Fluid construction grammar” makes use of semantic and syntactic poles, is bi-directional, selects meanings and maps them into the “real world”. It is called “fluid” because it takes into consideration the fact that users change and update their grammars often. This method means that a user input can be broken down syntactically in order to gain meaning from the grammatical components, whilst also being able to map the semantic relationships. The information gained this way enables the system to define the topic and sub-topics involved as well as giving information on the construction of the sentence to be volunteered to the user as a response.

A11.

Title: *Acoustic model development using HLDA for robust embedded in-car speech recognition.*

Authors: Antoni Abella, James Nealand and Kate Knill

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

Email: antoni.abella@crl.toshiba.co.uk

Abstract

Automatic Speech Recognition (ASR) is an important part of embedded systems for in-car navigation and personal media delivery applications. In-car embedded ASR solutions need to be robust over a pre-defined set of operating conditions. Typically a fixed microphone type, microphone position, and a range of operating environments are specified by the customer. Furthermore, embedded systems need to comply with strict memory footprint and computational cost specifications, making the development of robust acoustic models in a timely manner an important challenge.

Heteroscedastic Linear Discriminant Analysis (HLDA), a linear feature space transformation maximising the between to within class covariance ratio in a maximum likelihood scheme, has been widely used in LVCSR systems leading to a significant performance boost. The HLDA transform can be calculated off-line, improving accuracy for a minimal increase in computation cost. In this paper the sensitivity of HLDA to specific channel and environmental conditions is investigated. Initial evaluations show a consistent boost in performance over multiple tasks and languages. The observed gains do however vary significantly across different channel and environmental conditions.

The research reported in this paper explores the source of the variation in gains due to HLDA, with the objective of biasing performance improvements towards the target channel and operating environment. Delivering reliable gains for the target conditions through HLDA will significantly improve acoustic model quality, while also reducing the tuning effort required to produce robust acoustic models.

Initial investigations show that the gains from global HLDA can not be significantly biased by estimating the transform using data matched to the target conditions. One current further avenue of investigation is multiple-HLDA; estimating a set of transforms for a set of predefined regression classes. Current results and planned future work in this on-going project are reported

B1.

Title: *The Role of 'Delta' Features in Speaker Verification*
Authors: Ying Liu, Martin J. Russell and Michael J. Carey
Address: Dept. of Electronic, Electrical & Computer Engineering,
University of Birmingham,
Birmingham, B15 2TT.
Email: y.liu@bham.ac.uk

Abstract

We present an analysis of the role of 'delta' features in GMM-based Text- Independent Speaker Verification (TI-SV). We compare the models and verification performance obtained with 'delta' MFCC parameters alone with the more conventional 'static plus delta' parameters. We show that the values of the 'delta' parameters in 'delta only' and 'static plus delta' models are very different, and that the latter only outperform the former after RASTA filtering. This raises the question of whether the utility of the delta parameters arises from modelling non-stationary speech dynamics or their relative immunity to time-varying noise. Unfortunately, the scores obtained with the 'delta only' and 'static plus deltas' systems are highly correlated, and a fused system gives little improvement over the 'statics plus deltas' system.

B2.

Title: *Language Identification: Insights from the Classification of Hand Annotated Phone Transcripts*

Authors: Timothy Kempton and Roger K. Moore

Address: Speech and Hearing Research Group
Department of Computer Science,
The University of Sheffield,
Regent Court,
211 Portbello Street,

Email: t.kempton@dcs.shef.ac.uk

Abstract

Language Identification (LID) of speech can be split into two processes; phone recognition and language modeling. This two stage approach underlies the most successful LID systems. As phone recognizers become more accurate it is useful to simulate a very accurate phone recognizer to determine the effect on the overall LID accuracy. This can be done by using phone transcripts. In this paper LID is performed on phone transcripts from six different languages in the OGI multi-language telephone speech corpus. By simulating a phone recognizer that classifies phones into ten broad classes, a simple n-gram model gives low LID equal error rates (EER) of <1% and 7% on 30 seconds and 3 seconds respectively. This indicates that for even broad phone classes, improving the phone recognition accuracy can significantly increase the overall accuracy of the LID system.

B3.

Title: *An assessment of the effect of physical differences in the articulation of consonants and vowels on audiovisual temporal perception.*

Authors: Argiro Vatakis and Charles Spence

Address: Crossmodal Research Laboratory,
Department of Experimental Psychology,
University of Oxford,
9 South Parks Road,
Oxford OX1 3UD.

Email: argiro.vatakis@psy.ox.ac.uk

Abstract

We investigated the extent to which physical differences associated with the articulation of consonants and vowels affect the temporal aspects of audiovisual speech perception. A series of video clips of three different speakers (British English) uttering various different consonants and vowels was presented at a range of stimulus onset asynchronies (SOAs) using the method of constant stimuli. Participants made unspeeded temporal order judgments (TOJs) regarding which speech stream (auditory or visual) appeared to have been presented first. The sensitivity of the participants' temporal discrimination responses (measured in terms of the just noticeable difference; JND) and the point of subjective simultaneity (PSS) were analyzed in terms of variations in several different articulatory features including the place and manner of articulation and voicing for consonants, and the height and backness of the tongue and roundedness of the lips for vowels. The visual-speech stream had to precede the auditory-speech stream for most of the consonants tested in order for synchrony to be perceived. By contrast, for vowels, auditory leads were required. More importantly, these results also demonstrate that people are more sensitive to the temporal order of highly-visible speech stimuli (e.g., for bilabials and rounded vowels) and that the visual stimuli had to lead by a smaller interval in order for the PSS to be reached. This was not the case for the less-visible speech tokens (e.g., for velars and unrounded vowels), where higher JNDs and larger visual leads were observed. These findings suggest that in audiovisual speech perception, the visual-speech signal provides substantial cues to the auditory signal that modulate the relative processing times (i.e., less processing times for highly-visible stimuli) required for the perception of the speech signal.

B4.

Title: *Modelling of Formulaic Language*
Authors: Christopher Watkins and Stephen Cox
Address: School of Computing Sciences
University of East Anglia
Norwich
Email: cjw@cmp.uea.ac.uk

Abstract

There is evidence to suggest that humans use formulaic language to increase fluency in conversations. This is language that is retrieved whole from memory i.e. requires no novel processing or construction and has a predefined meaning attached to it, for example, the phrase "by and large".

Given this observation, a new method for clustering phrases is described. Training data is first segmented into variable length phrases using multigrams in a Maximum-Likelihood (ML) Viterbi approach. These phrases are then searched for in their corresponding parse trees and given a hybrid constituent label (not all phrases align with one constituent - they may cover one or more) with left and right context (the labels of surrounding phrases in the sentence). Phrases are then grouped together into classes based on identical constituent and left context labels. These initial classes are then merged to some pre-defined number of classes using the vector based measure, cosine similarity.

For evaluation, a hybrid language model is constructed which combines ngrams and stochastic finite state automata (SFSAs). Each phrase class is modelled using the SFSAs, while the ngram models external probabilities between SFSAs. This method results in a large decrease on the training set perplexity (~ 45%), while only providing a modest decrease on the development set (~ 6%). We suggest these results are due to the small variations that occur in the phrases between the test and dev set - meaning that the decoder chooses single words over phrases, which results in an increase in the perplexity due to a smaller number of singletons modelled in the training data. We suggest a generalisation method for the SFSAs which will allow for unseen phrases to be modelled in the SFSAs by inserting open slots for alternative words.

B5.

Title: *The Listening Room—a Speech-Based Interactive Art Installation*

Authors: Alexa Wright, Alun Evans and Mike Lincoln

Address: The Centre for Speech Technology Research,
University of Edinburgh,
2 Buccleuch Place
Edinburgh EH8 9LW

Email: mlicoll@inf.ed.ac.uk

Abstract

In this poster we describe 'The Listening Room' - an interactive art installation that incorporates a number of speech technologies. In The Listening Room up to three small sculptures are displayed on exhibition plinths. People entering the space are automatically tracked using webcams positioned overhead. An individual standing close to one of the sculptures triggers a disembodied voice, which will then try to engage that person in conversation. Using keywords to interpret what is said in reply, the disembodied voice will pursue a more or less meaningful dialogue that can be heard only at particular location in the space. 'The voice' will be able to conduct conversations at up to three different locations.

In the contemporary field of affective computing, the desire to include machines as part of the human world is very current and is extended to attempts to give the machine social and emotional intelligence. Following in the tradition of chatbots such as Eliza and Jabberwacky, The Listening Room is an artwork that explores the idea and experience of a seamless and unencumbered convergence of the 'real' and the 'virtual'. In this work, a softly spoken, but just perceptibly synthesized, female voice renders the interface almost intangible. The work playfully questions whether it is possible for humans to have meaningful social exchange with machines. For each user, the illusion of meaningful social exchange is mediated by the extent to which he or she is led to project personality or emotional content into the synthesized voice. The exhibition incorporates a number of interacting technologies to achieve its aim.

B6.

Title: *Long Audio Alignment Using Grapheme-Based Speech Recognition*

Authors: Dong Wang

Address: The Centre for Speech Technology Research,
University of Edinburgh,
2 Buccleuch Place
Edinburgh EH8 9LW

Email: dwang2@inf.ed.ac.uk

Abstract

Long audio indexing is the first step for retrieving some information from the audio/text pair. Traditionally, a phone based speech recognition system is used to generate a recognition transcription and then a matching algorithm is applied to find possible positions the audio among which can then be force aligned. In this work, we use a grapheme based decoder to index the long audio. Free of complex lexicon construction, easy to handle special words in the text reference of the audio, grapheme system, with proper online adaptation and a local language model trained from the reference text, can achieve enough accuracy for the alignment. The simplicity of the system helps quick implementation and rapid porting.

B7.

Title: *Application of accent morphing to improve the intelligibility of foreign-accented speech*
Authors: Kayoko Yanagisawa
Address: Department of Phonetics and Linguistics,
University College London
Wolfson House,
4 Stephenson Way,
London NW1 2HE
Email: k.yanagisawa@ucl.ac.uk

Abstract

Accent morphing is a technique in which the accent of a speaker is modified whilst the speaker identity is preserved. It would have application in speech-to-speech translation systems as well as in foreign language learning. In a speech-to-speech translation context, the accent of the output of a text-to-speech (TTS) system – built using the source language audio data – may be morphed towards the accent of a native speaker of the target language, making it more native-like and thus more intelligible.

This study implements and evaluates an accent-morphing system and shows how accent morphing can improve the intelligibility of English-accented Japanese sentences to Japanese listeners (from 57% to 84% words correct) by manipulating both spectral and prosodic features. It also investigates the relative contributions of segmental quality, pitch and timing to any change in intelligibility, and whether there are any interactions between these factors.

Recordings of Japanese sentences were generated by an English TTS system and a Japanese TTS system. Segmental and suprasegmental information were taken from the English TTS speech, and morphing towards the Japanese TTS was performed on the low-frequency spectral envelope in voiced regions, together with pitch and rhythm. Manipulations were carried out in such a way that any aspect unrelated to accent remained as unmodified as possible. Analysis of the different processing conditions revealed that morphing segmental quality, pitch or rhythm individually does not have a large impact on intelligibility, but the combination of segmental and suprasegmental changes has a super-additive effect. This interaction between the segmental and suprasegmental aspects suggests that segmental changes need to be matched to the correct prosodic context to have a big impact on intelligibility.

B8.

Title: *Modelling Confusion Matrices to Improve Speech Recognition Accuracy, with an Application to Dysarthric Speech*

Authors: Omar Caballero Morales and Stephen Cox

Address: School of Computing Sciences,
University of East Anglia,
Norwich, NR4 7TJ

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. In this paper, we introduce a technique that can increase recognition accuracy in speakers with low intelligibility by incorporating information from an estimate of the speaker's phoneme confusion matrix. The technique performs much better than standard speaker adaptation when the number of sentences available from a speaker for confusion matrix estimation or adaptation is low, and has similar performance for larger numbers of sentences.

B9.

Title: *Multilingual phone classification using Artificial Neural Networks*

Authors: Partha Lal

Address: The Centre for Speech Technology Research,
University of Edinburgh,
2 Buccleuch Place
Edinburgh EH8 9LW

Email: p.lal@sms.ed.ac.uk

Abstract

Large amounts of speech data are needed to train a speech recognition model but there are hundreds of languages for which sufficient data doesn't exist. This work is motivated by the possibility of using speech data in one language to improve the performance of a recogniser for another language.

Experiments were performed using the Croatian, German, Spanish and Swedish sections of the GlobalPhone corpus, consisting of newspaper text read by variety of speakers. Multilayer Perceptrons (MLPs) were trained using that data to perform phone classification, with phone labels derived from a forced alignment of a prior Hidden Markov Model (HMM). Data in each language was used separately, as well as being pooled across languages. Pooling the training data was made possible by using a shared phoneset.

The trained MLP has one output unit per phone, where the activation at that unit relates to the probability of that particular phone. Those activations can then be used, indirectly, as additional features in a conventional HMM, resulting in what is known as a tandem model. The MLP used there could be trained on a larger corpus in a different language, providing another opportunity for sharing data between languages.

B10.

Title: *A statistical technique for identifying articulatory roles in speech production*

Authors: Veena D.Singampalli and Philip J.B. Jackson

Address: School of Electronics and Physical Sciences,
University of Surrey,
Guildford,
Surrey GU2 7XH.

Email: V.Singampalli@surrey.ac.uk

Abstract

In describing speech production in the articulatory domain, many researchers have used distinctive features associated with phones through which coarticulatory effects are specified by phonological rules. Others have introduced coarticulation resistance and degree of articulatory constraint as a means of prioritising articulatory movements involved in speech gestures. On similar lines, the roles played by an articulator during an utterance can be termed critical, dependent or redundant at any time. We propose a statistical approach for identifying such roles from annotated articulatory data. The critical articulators for each phone are identified based on the strength of Kullback-Leibler divergence between phone-specific distributions and the global distributions. The effect of the critical articulators' configuration on dependent articulators is estimated using significant articulatory correlations. Two versions of the algorithm are considered, 1D and 2D, which either treat x and y coordinates of the articulatory data independently or incorporate covariation within the motion of an articulator in the mid-sagittal plane. The 2D case uses canonical correlation analysis. The performance of the critical articulator identification approach is evaluated by comparing the results obtained with our algorithm to the distinctive phonetic features. Such a model which captures the phonetic invariance in a statistical way from articulatory measurements has the potential to quantify cross-language differences in the realisation of the phonetic inventory, and provide a compact representation of critical articulatory events in speech production.

B11.

Title: *Language Models for Multiparty Meetings*
Authors: Songfang Huang and Steve Renals
Address: The Centre for Speech Technology Research,
University of Edinburgh,
2 Buccleuch Place
Edinburgh EH8 9LW
Email: s.f.huang@ed.ac.uk

Abstract

Statistical language model (LM) is an essential component of speech and language processing for human-computer interaction. We consider the research question of language modeling in the context of multiparty meetings. Although multiparty meetings are group- and multimodality-based, traditional language models for ASR in meetings have only considered the lexical information, without taking other available multimodal cues into account. We propose a multimodal language model, which attempts to exploit multimodal cues in meetings, such as prosodic, semantic, and visual information, to augment language models for ASR.

In the preliminary experiments, we exploited prosodic features for language modeling. Using an automatic syllable detection algorithm, the syllable-based prosodic features are extracted to form the prosodic representation for each word. Two modeling approaches are then investigated. One is based on a factored language model [Bilmes et al., 2003], which directly uses the prosodic representation and treats it as a 'word'. Instead of direct association, the second approach provides a richer probabilistic structure within a hierarchical Bayesian framework [AGelman, 1995] by introducing an intermediate latent variable to represent similar prosodic patterns shared by groups of words. Four-fold cross-validation experiments on the ICSI Meeting Corpus show that exploiting prosody for language modeling can significantly reduce the perplexity, and also have marginal reductions in word error rate.

Multimodal cues could be tightly incorporated into language models via a two-stage model [Goldwater et al., 2006] within the hierarchical Bayesian framework. We are currently working on Bayesian language models for meetings based on hierarchical Pitman-Yor processes [Teh, 2006], and also looking forward to present some preliminary results on this work during the meeting