UK & IE Speech Meeting, Birmingham, Dec 17-18, 2012
Programme and Abstracts

Monday, December 17, 2012

12:00 -- 12:10
Welcome message; Local information.

12:10 -- 13:10
Talks

13:10 -- 14:00
Catered Lunch

14:00 -- 16:00
Talks

16:00 -- 16:30
Break

16:30 -- 18:00
Posters

Tuesday, December 18, 2012

9:00 -- 10:30
Talks

10:30 -- 11:00
Break

11:00 -- 12:30
Posters

12:30 -- 13:30
Catered Lunch

13:30 -- 14:00
Talks

14:30 -- 15:00
Town hall meeting: future directions

TALKS

1. Audio and Visual Speech Processing at UEA

Much of the work at UEA is concentrated with incorporating visual information into three areas of speech processing:

- speech enhancement
- speech recognition
- speech synthesis

I will describe this work, which includes the technologies of lip-reading, speech enhancement using visual information, and expressive visual synthesis, and also talk about other work in audio-only speech processing.

2. Current Speech Research in Edinburgh

In this talk I'll give an overview of the current research we are doing at CSTR in speech recognition, speech synthesis, speech perception, and related areas. I'll discuss the main challenges that we are working on, and some of our most interesting recent findings. We find that the most compelling scientific challenges in speech technology are often motivated by target applications, and I'll talk about some of the application areas on which we are concentrating. I might also spend a couple of minutes discussing potential future directions.

3. Introduction to the Speech Communication Lab at Trinity College Dublin

The Speech Communication Lab at Trinity College Dublin has expertise in extra-linguistic and nonverbal speech processing, and specialises in the development of tools and corpora for the modelling and control of interaction in spoken discourse. We use audio, video, and biometric sensors to capture signals related to participants' cognitive states, and apply these through a discourse framework model to enable inference about the attitudes and intentions of participants in a spoken dialogue, to enable a form of advanced speech recognition focussed on cognitive states rather than text content, and a form of advanced speech synthesis that is aware of the attentive states of the listener and adjusts its output accordingly.

4. Machine Audition at CVSSP

Machine audition (listening) aims to interpret and understand complex audio scenes in a dynamic environment using computational models and algorithms. It covers a wide range of research topics in audio scene analysis, recognition, separation, tracking and understanding. In this talk, we will give a brief overview of the recent and ongoing activities on machine audition (listening) research in the Centre for Vision, Speech and Signal Processing (CVSSP) at University of Surrey. We will mainly discuss the activities on blind speech separation, audio-visual source separation and tracking, and spatial audio, including some recent results and demonstrations in these areas.

5. Novauris Technologies Ltd

Novauris is a small UK company that develops core automatic speech recognition technology. Our products are software licences and services. Our technology makes use of speech recognition techniques that you may be familiar with, in a way that is useful for a specific but widely useful class of speech recognition tasks, focussed on selection by voice from a large list. Recognition code runs on servers and embedded in cellphones. The talk will include some of the challenges presented by real-world tasks.

6. Novel methods for Speech Enhancement, Separation and Speaker Recognition

Ji Ming, Darryl Stewart, Danny Crookes

The work of the Speech group at QUB has focused on two different and novel research strands in processing speech. The first strand is based on using a corpus-based approach for several problems: speech enhancement in the presence of unpredictable noise, single channel speech separation, and speaker recognition. We use a corpus of clean speech data as our speech model, which enables us to model the speech rather than the noise, and therefore we do not require knowledge of the noise. Enhancement is achieved by finding a sample from the corpus that best matches the underlying speech signal. Key to the success of the method is the use of what we call the longest matching segment (LMS). The technique has also been successfully applied to the problem of speaker recognition.

The second research strand is audio-visual speech processing. We use an analysis of lip movements to supplement the audio information. With a careful choice of image features, lip movements have been shown to increase the accuracy of speech recognition. Lip movements have also been combined with audio-based speaker recognition to give an effective audio-visual speaker recognition system. Lip movements are useful in biometrics because they are unique to the speaker and are hard to emulate. They also give a useful ‘liveness’ test for biometric systems.

7. Overview of the CUED Speech Group

The Speech Group at Cambridge University Engineering Department work in speech and text processing with a primary focus of using statistical modelling and other machine learning techniques to construct advanced speech processing systems. It works in areas including speech recognition, spoken dialogue systems, speech generation, and machine translation. It has 4 members of permanent academic staff, typically 15 Research Associates and Fellows and a further 15-20 PhD students, along with some Masters and undergraduate students. The group is responsible for the widely used HTK HMM toolkit which has more than 100,000 registered users. This talk will give a brief overview of the speech group and discuss the areas of current work, and current funded research projects.

8. Overview of the Sigmedia Group at Trinity College Dublin

Humans love to communicate and as an engineer, my job is to use digital signal processing to enhance, repair or augment that experience. This is whether that communication is human-to-human or human-to machine. This is the underlying theme of my research activities. I have active research in audio-visual speech recognition, ageing speech with applications in biometrics and forensics, speech quality for VoIP, emotion in speech, and speech intelligibility for hearing aids. In this talk I plan to provide an overview of the activities and to introduce my group to the UK Speech Community.

9. Sheffield SPandH 2012

Phil Green, Roger Moore, Thomas Hain, Jon Barker, Guy Brown, Yoshi Gotoh

This is the UK-Speech presentation of work in Computer Science @ Sheffield. It will introduce our current work in:

Computational Hearing: active sound source localisation and separation, auditory basis for noise-robust ASR , hearing for robotics;

Speech Perception: perceptual constancy, turn-taking, speech perception in noise, intelligibility modelling, perceptually motivated speech enhancement ;

Automatic Speech Recognition: environment modelling, domain adaptation, meeting transcription, Arabic ASR, lecture content linking, British conversational telephone speech, speaker & language identification, diarisation, machine translation;

Clinical Applications of Speech Technology: Silent Speech by Permanent Magnetic Articulography, recognition of disordered speech;

Spoken Language Processing by Mind and Machine: the PRESCENCE model, ANTON (ANimatronic TONgue), reactive speech synthesis

Vocal Interactivity: in and between humans, animals and robots.

10. Speech, Audio and Acoustic Signal Processing at Imperial

Patrick A. Naylor and Mike Brookes

An overview of the research activities of the Speech and Audio Signal Processing group at Imperial will be presented. The group comprises around 10 people including 2 academic staff. Our current work related to speech has grown from a background of speech production modelling and now includes topics in noise reduction, dereverberation, microphone array signal processing and some aspects room acoustics. We have developed strong links with the law enforcement community in the UK and with a range of commercial collaborators. The talk will outline some recent project work and highlight the key ongoing themes for future activities and potential scope for interaction with the UK-Speech community.

11. Speech in noise: research into production, perception, measurement and modelling at UCL

Everyday speech communication takes place in a background of noise. This noise affects how we produce speech and how we listen to speech. Noise can be a significant problem when communication is also stressed by the presence of a hearing impairment, a degraded channel, or the occurrence of unfamiliar accents. Over the past few years a number of research activities at UCL have taken as their focus the problems caused by noise in speech communication. The CLEAR project under Mark Huckvale looked at the effects of noise-reduction on the intelligibility of speech and built new signal-based predictors of the effect of speech signal enhancement algorithms. In an MRC funded project, Stuart Rosen aims to better understand how people with normal hearing manage to understand speech in the background of other talkers while people with hearing impairment do not. In a related project, auditory brainstem responses are used to investigate which difficulties in understanding speech in background noises arise from deficits in auditory encoding at the first neural stages of the auditory pathway. An Action on Deafness funded project of Andrew Faulkner looks at the benefits to cochlear implant patients of better exploiting their residual acoustic hearing to deal with noisy environments. In ESRC funded projects of Valerie Hazan, the impact of noisy and degraded channels on the clarity of speech production is recorded using a controlled dialogue task. Finally, Paul Iverson is investigating how the mutual intelligibility of interlocutors in noise is affected by the similarity of their spoken accents.

12. Research in Speech and Language Technology at the School of Electronic, Electrical and Computer Engineering, University of Birmingham

Mike Carey, Peter Jančovič, Emilie Jean-Baptiste, Roozbeh Nabiei, Maryam Najafian, Martin Russell, Saeid Safavi and Masoud Zakeri

Research in speech and language technology in the School of Electronic, Electrical and Computer Engineering (EECE) at Birmingham is part the Human Computer Interaction (HCI) group, which includes relevant research in EECE and the School of Computer Science. EECE has been active in this area since 1998. This presentation will cover seven main areas:
(1) Novel approaches to acoustic modelling for speech recognition: Current approaches to speech recognition use large statistical models, which rely on large speech corpora for training and need substantial computing resources. This makes them difficult to adapt to new domains. The objective of this research is to develop more parsimonious approaches that overcome the need for such large corpora by incorporating more faithful models of speech. In the past, this research was supported by two EPSRC awards, and a new project in this area is about to start.
(2) Paralinguistic speech processing: This refers to research in language, regional accent, speaker, ethnic group, age and gender recognition. We apply a range of acoustic, phonotactic and fused methods to these problems and have achieved good results. This has been a focus for recent PhD research and has been published in Computer Speech and Language and IEEE SPL.
(3) Implications of regional accents for speech recognition: From (2) we know that it is possible to recognise an individual's regional accent, ethnic group, age and gender with accuracies over 90% using just 30s of speech. The objective of this research topic is to use these techniques for rapid speaker adaptation for speech recognition.
(4) Speech processing in noise: The objective of this research is to develop techniques for noise robust extraction of information that characterise speech signal and its incorporation into speech pattern processing and for enhancement of speech signal corrupted by noise. This research was supported by two EPSRC awards.
(5) Audio applications of speech algorithms: This covers recent research in applying the techniques from speech processing to other types of audio data, for instance, bird vocalisations.
(6) Non-audio applications of speech algorithms: The objective of the EU CogWatch project is to develop cognitive rehabilitation technologies for patients who are recovering from a stroke and have difficulty in completing activities of daily living (ADLs). The initial focus is tea making. We will explain how techniques from spoken dialogue systems (SDSs), which combine HMMs for speech recognition with partially observable Markov decision processes (POMDPs) for dialogue processing, are being applied. We will also discuss key differences between this application and SDS.
(7) Speech corpora: The Speech Ark is a spin out company of the University of Birmingham that creates and distributes speech corpora for speech and language technology research. We will describe the corpora that are available and the research that they have supported to-date.

POSTERS

1. A comparative study of adaptive, automatic recognition of disordered speech

Heidi Christensen, Stuart Cunningham, Charles Fox, Phil Green, Thomas Hain

Speech-driven assistive technology can be an attractive alternative to conventional interfaces for people with physical disabilities. However, often the lack of motor-control of the speech articulators results in disordered speech, as condition known as dysarthria. Dysarthric speakers can generally not obtain satisfactory performances with off-the-shelf automatic speech recognition (ASR) products and disordered speech ASR is an increasingly active research area. Sparseness of suitable data is a big challenge. The experiments described here use UAspeech, one of the largest dysarthric databases available, which is still easily an order of magnitude smaller than typical speech databases. This study investigates how far fundamental training and adaptation techniques developed in the LVCSR community can take us. A variety of ASR systems using maximum likelihood and MAP adaptation strategies are established with all speakers obtaining significant improvements compared to the baseline system regardless of the severity of their condition. The best systems show on average 34% relative improvement on known published results. An analysis of the correlation between intelligibility of the speaker and the type of system which would represent an optimal operating point in terms of performance shows that for severely dysarthric speakers, the exact choice of system configuration is more critical than for speakers with less disordered speech.

2. A "Direct" Speech Synthesis Approach for a Medical Speech Aid

Robin Hofe

Speech technology applications typically require the presence of clear, audible speech, which is not always available in noisy environments or when the user is medically impaired. A new emerging trend is therefore to replace or complement the acoustic speech signal, e.g. through lip reading or tracking of tongue movements by various means. These alternative systems are called 'silent speech interfaces' (SSI).

One potential use of SSI is in communication aids for patients who have lost their natural ability to speak, e.g. through laryngectomy. One such system is the Magnetic Voice Communication Aid (MVoCA), that tracks movements of the tongue and lips during speech.

In order to replace the user's natural voice in communicative situations, the synthetic speech signal would ideally be provided in real-time during the user's articulation. This requirement cannot be met by a recognition-and-synthesis approach, where SSI data is used for speech recognition and the recognised utterance is synthesised in a second step. Instead, a "direct synthesis" approach is proposed and investigated, based on a direct mapping between MVoCA signals and control parameters of a parametric speech synthesiser.

3. ABAIR:multi-dialect Irish synthesis

Neasa Ni Chiarain, Christoph Wendler, Harald Berthelsen & Ailbhe Ní Chasaide

The ABAIR project is focused on developing the first text-to-speech systems for the different dialects of the Irish (Gaelic) language. This poster documents some of the challenges encountered in the direction of this goal and some of the key achievements to date. Also highlighted are the applications of the synthesis including its use in: educational language learning games, virtual reality applications, digital talking books and aids for the visually impaired.

4. An analysis of parameter generation considering global variance

Matt Shannon, William Byrne

We present a new analysis of speech parameter generation considering global variance based on Lagrange multipliers. This analysis sheds light on one source of the artifacts that GV generation sometimes introduces, and motivates a new fast approximation to full GV generation.

5. Analysis of Speaker Clustering Strategies for HMM-Based Speech Synthesis

Rasmus Dall, Christophe Veaux, Junichi Yamagishi, Simon King

This work describes a method for speaker clustering, with the application of building average voice models for speaker- adaptive HMM-based speech synthesis that are a good basis for adapting to specific target speakers. Our main hypothesis is that using perceptually similar speakers to build the average voice model will be better than use unselected speakers, even if the amount of data available from perceptually similar speakers is smaller. We measure the perceived similarities among a group of 30 female speakers in a listening test and then apply multiple linear regression to automatically predict these listener judgements of speaker similarity and thus to identify similar speakers automatically. We then compare a variety of average voice models trained on either speakers who were perceptually judged to be similar to the target speaker, or speakers selected by the multiple linear regression, or a large global set of unselected speakers. We find that the average voice model trained on perceptually similar speakers provides better performance than the global model, even though the latter is trained on more data, confirming our main hypothesis. However, the average voice model using speakers selected automatically by the multiple linear regression does not reach the same level of performance.

6. Articulatory features for speech driven head motion synthesis

Atef Ben-Youssef, Hiroshi Shimodaira, David A. Braude

We present speech driven head motion synthesis using hidden Markov models (HMMs) approach. In this approach, multi-stream HMMs are trained jointly on synchronous streams of speech and head motion data, acquired by NaturalPoint OptiTrack motion capture system. We introduce here the articulatory features, that represent an intermediate parametrisation of speech. Articulatory features are predicted from speech using an HMM-based inversion mapping system. Canonical correlation analysis (CCA) shows that the predicted articulatory features are more correlated with head motion than prosodic and/or cepstral speech features. Speech driven head motion synthesis is achieved in two steps. HMMs states decoding is first performed using Viterbi algorithm. Then head motion (i.e. Euler angles trajectories) are inferred from the decoded state sequence using the maximum-likelihood parameter generation algorithm (MLPG).

7. Audio Applications of Speech Algorithms: Automatic Analysis of Bird Vocalisations

Peter Jančovič, Münevver Köküer, Masoud Zakeri, Martin Russell

An automatic analysis of bird vocalisations for the identification of bird species, study of their behaviour, and the way of their communication is important for a better understanding of the environment we are living in and in the context of environmental protection.

The aim of this project is to design and develop a system that can detect and recognise bird species automatically by analysing their songs and calls. The project investigates modifications of techniques which have been demonstrated to be effective for automatic speech pattern processing and develop novel techniques that appropriately account for unique properties of bird acoustic signals.

This poster presents initial outcomes of our research. The first part focuses on an automatic segmentation of acoustic signals recorded in free natural environment, which contain bird vocalisations but also various noises and human speech. The input acoustic signal is converted into its time-frequency representation which provides information about frequency content over time. The frequency-contour features extracted from this representation are then employed for automatic segmentation using Dynamic Time Warping method. The second part focuses on development of an initial bird recognition system based on Gaussian mixture modelling.

8. Automatic Transcription Of Academic Lectures From Diverse Disciplines

Ghada AlHarbi and Thomas Hain

In a multimedia world it is now common to record professional presentations, on video or with audio only. Such recordings include talks and academic lectures, which are becoming a valuable resource for students and professionals alike. However, organising such material from a diverse set of disciplines seems to be not an easy task. One way to address this problem is to build an Automatic Speech Recognition (ASR) system in order to use its output for analysing such materials. In this work ASR results for lectures from diverse sources are presented. The work is based on a new collection of data, obtained by the Liberated Learning Consortium (LLC). The study's primary goals are two-fold: first to show variability across disciplines from an ASR perspective, and how to choose sources for the construction of language models (LMs); second, to provide an analysis of the lecture transcription for automatic determination of structures in lecture discourse. In particular, we investigate whether there are properties common to lectures from different disciplines. This study focuses on textual features. Lectures are multimodal experiences- it is not clear whether textual features alone are sufficient for the recognition of such common elements, or other features, e.g. acoustic features such as the speaking rate, are needed. The results show that such common properties are retained across disciplines even on ASR output with a Word Error Rate (WER) of 30%.

9. C2H: A Computational Model of H&H-based Phonetic Contrast in Synthetic Speech

Mauro Nicolao, Javier Latorre, Roger K. Moore

This paper presents a computational model of human speech production based on the hypothesis that low energy attractors for a human speech production system can be identified, and that interpolation/extrapolation along the key dimension of hypo/hyper-articulation can be motivated by energetic considerations of phonetic contrast. An HMM-based speech synthesiser along with continuous adaptation of its statistical models was used to implement the model. Two adaptation methods were proposed for vowel and consonant models and their effectiveness was tested by showing that such hypo/hyper-articulation control can manipulate successfully the intelligibility of synthetic speech in noise. Objective evaluations with the ANSI Speech Intelligibility Index indicate that intelligibility in various types of noise is effectively controlled. In particular, in the hyper-articulation transforms, the improvement with respect to un-adapted speech is above 25%.

10. CogWatch: an unfamiliar application of some familiar techniques

Chris Baber, Emilie Jean-Baptiste, Roozbeh Nabiei, Manish Parekh, Martin Russell

The objective of the EU 'CogWatch'[1] project is to develop rehabilitation technologies that help patients who are recovering from a stroke to complete a range of activities of daily living (ADL) independently. A third of these patients will experience long term physiological or cognitive disabilities, and a significant proportion can suffer from Apraxia or Action Disorganisation Syndrome (AADS), where symptoms include impairment of cognitive abilities to carry out ADL. The CogWatch system will track a participant's progress as he or she tries to complete an activity, and return an appropriate cue if an error occurs or is judged by the system to be imminent. The focus of the first prototype system is the activity of making a cup of tea, but the project will expand to include other ADLs such as dressing, food preparation and grooming.

An activity, or goal, such as tea making can be broken down into a hierarchy of sub-goals, tasks and sub-tasks. Different sequences of sub-goals can constitute a successful completion of the goal, and there are typically many alternative instantiations of a sub-goal as a sequence of tasks.

In order to monitor the patient's progress, each of the objects involved is fitted with a 'CogWatch instrumented coaster' (CIC). This is an electronic drink mat, containing an accelerometer and three force sensitive resistors (FSRs). The accelerometer indicates the acceleration of an object in its x, y and z planes, while the FSRs show whether it is resting on a surface or raised in the air and when its weight changes. The time varying sequences of data from these sensors are synchronized and passed to an automatic activity recognition (AAR) system, where they are classified as tasks or sub- goals, and these are interpreted by the task model (TM). The job of the TM is to estimate the participant's status with respect to completing the activity at each stage in the interaction, and to intervene appropriately if it judges that an error has occurred or is likely.

This looks very much like a spoken dialogue system (SDS). Indeed, in the first CogWatch prototype we are using hidden Markov models (HMMs) in the AAR to classify the sensor data into sub-goals or tasks, and a partially observable Markov decision process (POMDP) to model the task. However, do these technique lend themselves naturally to this domain, or is it just an instance of Maslow's hammer [2], or perhaps (apologies to Maslow) a variant that should be referred to as Maslow's HMMer - "I suppose it is tempting, if the only tool you have is a HMM, to treat everything as if it were speech recognition"?

This poster will describe the CogWatch system. It will argue that methods from speech and language technology are appropriate because of the nature of the problems, that these new applications can benefit from the experience and investment of the speech and language research community, and that, conversely, challenges in these new areas might give new insights into difficult speech and language processing problems.

[1] http://www.cogwatch.eu

[2] Abraham H. Maslow (1966), The psychology of science

11. Compensating for Ageing and Quality variation in Speaker Verification

Naomi Harte

Performing speaker verification in the simultaneous presence of ageing progression and changing speech sample quality is an important, open problem. The issues of ageing and quality variation go hand in hand; the effect of ageing increases with time, while variations in quality are also more likely to be encountered as time passes. In this work we demonstrate the effect of ageing on speaker verification performance, and show the relationship between quality variation and verification score via a range of established quality measures. We employ a stacked classifier framework to combine the output of the baseline verification system with ageing information and quality measures. This new approach to long-term speaker verification allows for a multi-dimensional decision boundary that significantly improves upon the baseline performance. The proposed framework is evaluated on the Trinity College Dublin Speaker Ageing Database.

12. Confusion Modelling using Weighted Finite-State Transducers for Automated Lip-reading

Dominic Howell

The accuracy of current state-of-the-art automated lip-reading systems is significantly lower than that obtained by acoustic speech recognisers. These poor results are likely due to the lack of information from the visual signal. Confusions provide challenging problems for visual speech recognisers. Previous studies in dysarthric speech recognition have used a network of weighted finite-state transducers to improve accuracy. A dysarthric speaker is unable to produce a full set of speech sounds and hence has a limited phonemic repertoire: this is a similar situation to lip-reading, where some sounds are impossible to detect. Here, we explore the use of weighted finite-state transducers to improve lip-reading recognition by modelling visual confusions in speech.

13. Contrasting the Effects of Different Frequency Bands on Speaker and Accent Identification

Saeid Safavi, Abualsoud Hanani, Martin Russell, Peter Jančovič, Michael J Carey

This poster presents an experimental study investigating the effect of frequency sub-bands on regional accent identification (AID) and speaker identification (SID) performance on the ABI-1 corpus. The AID and SID systems are based on Gaussian mixture modeling. The SID experiments show up to 100% accuracy when using the full 11.025 kHz bandwidth. The best AID performance of 60.34% is obtained when using band-pass filtered (0.23–3.4 kHz) speech. The experiments using isolated narrow sub-bands show that the regions (0–0.77 kHz) and (3.40–11.02 kHz) are the most useful for SID, while those in the region (0.34–3.44 kHz) are best for AID. AID experiments are also performed with intersession variability compensation, which provides the biggest performance gain in the (2.23–5.25 kHz) region.

14. Cross-Lingual Knowledge Transfer In DNN-Based LVCSR

Pawel Swietojanski, Arnab Ghoshal, Steve Renals

We investigate the use of cross-lingual acoustic data to initialise deep neural network (DNN) acoustic models by means of unsupervised restricted Boltzmann machine (RBM) pretraining. DNNs for German are pretrained using one or all of German, Portuguese, Spanish and Swedish. The DNNs are used in a tandem configuration, where the network outputs are used as features for a hidden Markov model (HMM) whose emission densities are modeled by Gaussian mixture models (GMMs), as well as in a hybrid configuration, where the network outputs are used as the HMM state likelihoods. The experiments show that unsupervised pretraining is more crucial for the hybrid setups, particularly with limited amounts of transcribed training data. More importantly, unsupervised pretraining is shown to be language-independent. Additionally, we show that finetuning the hidden layers of the DNNs using data from multiple languages improves the recognition accuracy compared to a monolingual DNN-HMM hybrid system.

15. Dirichlet Process Mixture of Experts models in Speech Recognition

Jingzhou Yang, Mark Gales, Rogier van Dalen

Bayesian non-parametric models recently have been widely employed in machine learning. Compared with their parametric counterparts, the problem of choosing model complexity can be side-stepped; as Bayesian models, the problem of over-fitting can be mitigated. Meanwhile, discriminative models, which model the class posterior distribution directly, are potentially more suitable for the classification tasks compared with generative models. Previous work on discriminative model on generative kernels have shown great improvement in speech recognition. However, for input data with complicated distributions, it is hard to model the whole data set by using a single classifier. A mixture-of-experts model could solve this problem. We introduce the mixture-of-experts model with gating network based on Dirichlet process in speech recognition, which takes the advantage of Bayesian non-parametric models to handle the problem of choosing the number of experts, and discriminative models for the classification tasks.

16. Efficient Decoding with Generative Score-Spaces Using the Expectation Semiring

Rogier C. van Dalen, Anton Ragni, and Mark J. F. Gales

Speech recognisers are usually based on hidden Markov models (HMMs). These are finite-state models, with observation vectors conditionally independent given the state sequence. This assumption yields efficient algorithms, but it limits the power of the model. An alternative type of model that allows a wide range of features and dependence structures is a log-linear model. For example, it can use longer-span, variable-length features. For decoding continuous speech, however, the optimal combination of segmentation of the utterance into words and word sequence must be found. Features must therefore be extracted for each possible segment of audio. For many types of features, this becomes slow. This poster will discuss how long-span features can be derived from the likelihoods of word HMMs. Standard adaptation techniques can then be used. Derivatives of the log-likelihoods, which break the Markov assumption, are appended. The poster will show how to decode with this specific class of model in quadratic time.

17. Enhancing Speech by Reconstruction from Robust Acoustic Features

Philip Harding, Ben Milner

A method of speech enhancement is developed that reconstructs clean speech from a set of acoustic features using a sinusoidal model of speech. This is a significant departure from traditional filtering-based methods of speech enhancement. A major challenge with this approach is to estimate accurately the acoustic features (voicing, fundamental frequency, spectral envelope) from noisy speech. This is achieved using maximum a-posteriori estimation methods that operate on the noisy speech. Objective results are presented to optimise the proposed system and a set of subjective tests compare the approach with traditional enhancement methods.

18. Exploiting Long-Range Temporal Dynamics Of Speech For Noise-Robust Speaker Recognition

Ayeh Jafari, Ramji Srinivasan, Danny Crookes, Ji Ming

Temporal dynamics is an important feature of speech that distinguishes speech from noise, as well as distinguishing between different speakers. In this paper, we present an approach to maximally extract this feature of speech to improve the robustness against background noise, for text-independent speaker recognition. The new approach identifies and compares the longest matching speech segments between the training and test speech to increase noise immunity. Experiments have been conducted on the NIST 2002 SRE database in the presence of various types of noise including fast-varying song and music. The new approach has shown significantly improved performance over conventional noise-robust techniques.

19. Expressive Visual Speech Synthesis

Felix Shaw

Much work has gone into creating realistic computer generated humanoid avatars. The video game and movie industries spend millions on animation and CGI created scenes and characters every year. Most of these techniques rely on artists modelling key frames, and having 3D software packages interpolate the intermediate frames. More recently, Model based approaches have been proposed, which can be used to synthesise entire video segments of talking animated faces. It is thought that if a way can be found of projecting emotional expression into such video sequences, then it will make it possible to automatically synthesise complex, realistic looking video without the need for manual creation of keyframes or the large corpora of video required for concatenative synthesis. This poster presents preliminary work to this end.

20. GloRi: the Glottal Research Instrument

John Dalton, John Kane, Christer Gobl

This poster presents the new voice analysis system, GlorRi (or Glottal Research Instrument), developed in the Phonetics and Speech laboratory, in Trinity College Dublin. We adopt a parallel approach to voice analysis. On the one hand, for certain purposes one may require precise fine-grained characterisation of the glottal excitation source, for other purposes, one may require more coarse-grained measurements for robust differentiation of different aspects of tone-of-voice/voice quality. The GloRi system facilitates this parallel approach for carrying out precise estimation and modelling of the glottal source, while also providing a batch analysis component for automatic analysis of large volumes of data. The system incorporates ongoing algorithm development within the lab, to help further automate and improve the robustness of the various analysis workflows.

21. Improved Detection Of Anomaly Ball-Hit Events In A Tennis Game Using Multimodal Information

Qiang Huang, Stephen J. Cox

In a tennis game, some anomaly match events, such as ``fault serve'' and ``ball out'' reported by the line judges, play important roles in video analysis as they explicitly indicate the match progress. However, some conventional methods, using only audio information, often fail to accurately detect these match events because of the acoustic mismatch between the training and the test data and interfering noise caused by spectators' applauses and players' yells. In order to improve the detection of these anomaly events, in this paper, we present a framework using both audio and visual information. In addition to the detection of the sound of line judges shouts, we also take advantage of visual information to identify whether the related event occurs in play-shot scene and whether the ball bounces off in an invalid region by locating the court lines and ball's position. To evaluate the effectiveness and robustness of our approach, we test it on three different tennis matches. The generated results show the use of our approach significantly outperforms several baselines by 35% on the three test matches.

22. Improving computer lip-reading with machine learning

Yogi Bear

Shapelets and Motifs are currently used machine learning techniques for Time Series Classification (TSC). It is hypothesised that these techniques may be applicable for computer lip-reading from visual data in order to improve the robustness of visual speech recognition.

23. inEvent: Accessing Dynamic Networked Multimedia Events

Fergus McInnes, Catherine Lai, Steve Renals, Jean Carletta

Events such as conferences, lectures and videoconferenced meetings are increasingly generating large quantities of online multimedia material, comprising video recordings plus speakers' presentations, documents, online comment and discussion streams, and metadata such as topic keywords and participant lists. inEvent is a three-year European project (2011-2014) developing techniques for automatically analysing, annotating, indexing and linking this material so as to make it more readily accessible to users. The poster will present an overview of the project, and highlight the research at Edinburgh which is focused on speech recognition with topic adaptation and on social and semantic analysis of meetings.

24. Iterative Classification Of Regional British Accents In I-Vector Space
Andrea DeMarco, Stephen J. Cox

Joint-Factor Analysis (JFA) and I-vectors have been shown to be effective for speaker verification and language identification. Channel factor adaptation has also been used for language and accent identification. In this paper, we show how these techniques can be used successfully in the task of accent classification, and we achieve good accuracy on a 14 accent problem using a novel iterative classification framework based on an iterative linear/quadratic classifier. These results compare favourably with recent results obtained using other non-fused acoustic techniques.

25. Language Identification by using IPA(International Phonetic Alphabet)
Zhuoyi Dai

The language identification has been studied for decades. However, the IPA is not widely used for speech recognition. My concentration focus on whether the IPA can improve the language identification performance.

26. Low-level and high-level models of perceptual compensation for reverberation
Amy V. Beeston and Guy J. Brown

Mechanisms of perceptual constancy allow us to compensate for typical variation in our everyday environment, seen for instance in the recent finding that speech perception improves with prior listening in a reverberant room (Brandewie & Zahorik, 2010). However, compensation for reverberation can be disrupted in a predictable fashion by applying incongruous reverberation to a test word and its preceding context (Watkins, 2005). In a speech identification task involving matched and mis-matched context- and test-reverberation, human listener responses are qualitatively matched by results from low-level and high-level computational models of perceptual compensation for reverberation. The low-level model uses an efferent feedback loop to monitor and control the dynamic range of the simulated auditory nerve response that results after peripheral processing in the afferent pathway. This helps to recover dips in the temporal envelope which would otherwise be filled with reflected energy. In an alternative approach, the high-level model views compensation for reverberation as an acoustic model selection process where analysis of the preceding context speech informs the selection of an appropriate (or mis-matched) acoustic model for the test word.

27. Maximum a Posteriori Adaptation of Subspace Gaussian Mixture Models for Cross-lingual Speech Recognition
Liang Lu, Arnab Ghoshal, Steve Renals

This paper concerns cross-lingual acoustic modeling in the case when there are limited target language resources. We build on an approach in which a subspace Gaussian mixture model (SGMM) is adapted to the target language by reusing the globally shared parameters estimated from out-of-language training data. In current cross-lingual systems, these parameters are fixed when training the target system, which can give rise to a mismatch between the source and target systems. We investigate a maximum a posteriori (MAP) adaptation approach to alleviate the potential mismatch. In particular, we focus on the adaptation of phonetic subspace parameters using a matrix variate Gaussian prior distribution. Experiments on the GlobalPhone corpus using the MAP adaptation approach results in word error rate reductions, compared with the cross-lingual baseline systems and systems updated using maximum likelihood, for training conditions with 1 hour and 5 hours of target language data.

28. Model-based approaches to robust speech recognition in reverberant environments

Y.-Q. Wang, M. J. F. Gales

Model-based approaches to handling additive background noise and channel distortion, such as vector Taylor series (VTS) compensation has been intensively studied and extended. However, less work has been done on applying these approaches for reverberant environment robustness. In this work, VTS is extended to compensate acoustic models for the effect of both reverberant and additive noise, yielding a new compensation scheme, reverberant VTS Joint (RVTSJ), in which noise model parameters can be obtained via an EM algorithm. Furthermore, a neural canonical model can be adaptively trained by using RVTSJ to factor out unwanted additive and reverberant noise variations in multi-conditional training data. Experimental results demonstrated that RVTSJ adaptation significantly improved model robustness in reverberant environments while reverberant adaptive training (RAT) yields a neural canonical model which is more amenable to adaptation in unseen noise conditions.

Y.-Q. Wang and M. J. F. Gales, "Model-based approaches to adaptive training in reverberant environments", in Proc. Interspeech-2012.

Y.-Q. Wang and M. J. F. Gales, "Improving reverberant VTS for hands-free robust speech recognition", in Proc. ASRU-2011.

M. J. F. Gales and Y.-Q. Wang, "Model-based approaches to handling additive noise in reverberant environments", in Workshop on Hands-free Speech Communication and Microphone Arrays, May 2011

29. Movement Based Parametrisation for Speech Driven Head Motion Synthesis

David Braude, Hiroshi Shimodaira

In this research we propose an alternative system for parametrising head motion used for speech driven animation models. Traditionally the methods used are based around a frame-wise generation approach for instance position or velocity. In this research we have based the parametrisation on movement features for instance when the head changes direction. This change addresses the 'jerkiness' issue that frame-wise methods can have.

30. New algorithms in voice processing

John Kane, Christer Gobl

Many recent speech technology developments have sought to exploit measurements which can characterise and differentiate different aspects of voice quality. However, there is clear room for improvement in terms of the effectiveness of such measurements and in terms of their robustness to non-ideal recording conditions. This poster give details and experimental results for two newly developed algorithms which attempt to address this issue. One method applies wavelet decomposition of the speech signal for deriving a measurement used for discriminating breathy to tense voice, while the other algorithm involves two newly developed acoustic features used as inputs to a decision tree classifier for detecting creaky voice.

31. Objective Evaluation of Naturalness Perception in Singing Voice Synthesis

Ryan Stables, Munevver Kokuer, Cham Athwal

In this study, a metric to objectively evaluate the naturalness of synthesized singing voices is developed. We use a modified bag-of-frames model to identify the low level audio descriptors that are salient at selected segmental levels, against a subjectively labelled dataset of synthesized voices. First, a large corpus of audio descriptors commonly used in the analysis of speech and music signals is extracted from the dataset, followed by a linear-correlation based variable ranking procedure. A classification and cross-validation stage is then performed using a support vector machine, iteratively trained with selected feature subsets. Each subset is based on the input representation, segmental level and data source of the feature vector, evaluated against a series of target vectors representing high-level attributes of naturalness perception. The results show that the decision boundary of the classifier can be represented as a function of MFCC and Peak-spectra based feature vectors, taken from a combination of short-term and suprasegmental frames of the sound source, with an observed accuracy of 83.13%, compared to 74.46% when trained with the global feature-set. Finally, we infer relationships between the feature salience and classification accuracies observed in the study, to a series of semantic descriptors, derived from a qualitative survey of naturalness perception.

32. Paraphrastic Language Models

Xunying Liu, Mark Gales & Phil Woodland

In natural languages multiple word sequences can represent the same underlying meaning. Only modelling the observed surface word sequence can result in poor context coverage, for example, when using *n*-gram language models (LM). To handle this issue, this paper presents a novel form of language model, the paraphrastic LM. A phrase level paraphrase model that is statistically learned from standard text data is used to generate paraphrase variants. LM probabilities are then estimated by maximizing their marginal probability. Significant error rate reductions of 0.5%-0.6% absolute (3%-5% relative) were obtained over the baseline *n*-gram LMs on two state-of-the-art recognition tasks for English conversational telephone speech and Chinese broadcast speech using a paraphrastic multi-level LM modelling both word and phrase sequences. When it is further combined with neural network LMs, significant error rate reduction of 0.9% absolute (9% relative) and 0.5% absolute (5% relative) were obtained over the baseline *n*-gram and neural network LMs respectively.

33. Performance-Based Measurement of Speech Quality

The extent to which speech communication systems deliver speech signals which are 'fit for purpose' can be assessed through a range of evaluation techniques including intelligibility and quality measures. But when signal quality is high, performance-based scales of intelligibility reach ceiling values and evaluation is performed instead using opinion-based rating scales. In this work we investigate whether performance-based testing can be applied even when intelligibility is high by measuring the cognitive effort listeners employ to understand speech. If effective, such testing could be applied to the evaluation of speech signal enhancement algorithms, to the performance of hearing aids and to the comparison of speech synthesis systems.

We outline three experimental approaches to performance-based measurement of speech signal quality that we have explored in our lab: (i) reaction time for digit recognition, (ii) error detection in audio proof-reading, and (iii) reaction time difference between congruent and incongruent trials in a response competition paradigm.

Through these approaches, we have shown that noise can have a measurable effect on task performance even when speech signal intelligibility is high. However, evidence so far is that so-called 'noise reduction' techniques - which are often said to improve 'quality' - do not improve these measures. Future work is designed to increase the statistical power of the tests and to understand further the perceptual and cognitive mechanisms by which signal quality affects listener performance.

Huckvale, M., Leak, J., "Effect of Noise Reduction on Reaction Time to Speech in Noise", Interspeech 2009, Brighton, U.K.

Huckvale, M., Hilkhuysen, G. "Performance-based Measurement of Speech Quality with an Audio Proofreading Task", J. Audio Engineering Society, June 2012

34. Progress and Prospects for Speech Technology: Results from Three Sexennial Surveys

Roger K. Moore

In 1997, and again in 2003, the author was invited to conduct a survey at the IEEE workshop on 'Automatic Speech Recognition and Understanding' (ASRU) in which attendees were offered a set of statements about putative future events relating to progress in various aspects of speech technology R&D. The task of the respondents was to assign a date to each possible event. The 1997 and 2003 results were published at INTERSPEECH 2005 in Lisbon. Six years later, the author was invited by the organisers of ASRU'2009 to repeat the survey for a third time, and this paper presents the combined results from all three 1997, 2003 and 2009 surveys. The overall conclusion is that, over the twelve year period progress is perceived as slow, and the future appears to be generally no nearer than it has been in the past. However, on a positive note, the 2009 survey confirmed that the market for speech technology applications on mobile devices would be highly attractive over the next ten or so years.

35. Rapid Compact Nonlinear Adaptation for Large-Vocabulary Speech Recognition

Zoi Roupakia, Mark Gales

Recently, kernel eigenvoice adaptation was revisited using kernel representations of distributions for rapid nonlinear speaker adaptation. These representations ensure the validity of the adapted distribution functions and enable expectation-maximisation training to be applied. In addition, compact representations of the eigenvoices have been used to scale up its application to large-vocabulary speech recognition. The resulting speaker and state-dependent model is an expanded GMM. Though gains have been shown in terms of word error rate for rapid speaker adaptation, this approach slows down the decoding process due to an increase in the number of likelihood evaluations. To overcome this issue, a cost function has been introduced to reduce the model complexity and resulting recognition cost, whilst yielding equally powerful adapted models. This criterion is based on a matched- pair bound approximation of the Kullback-Leibler divergence between the original speaker dependent GMM and a compact target GMM along with a clustering scheme. This work compares different upper-bound approximations for the Kullback-Leibler distance between the original and the compact target adapted model. Experimental results are presented for two different large vocabulary domains: conversational telephone speech and noise corrupted speech.

36. Recovering training data from rough transcriptions in a damaged found corpus

Charles W Fox, Thomas Hain

Large corpora of transcribed speech are rare and expensive to acquire, yet are of great use for developing and improving ASR systems. Of particular research interest are corpora of natural speech, such as far-field recordings of multiple speakers in noisy environments. The ESDS database contains many thousands of hours of such recordings, but made for non-ASR purposes. We describe one example corpus from this database, called Family Life, discuss the challenges it presents for data recovery prior to ASR and algorithms for dealing with some of them, and give baseline recognition results. In particular, the transcriptions have no timing annotations, and many of the audio files are mislabelled. Family Life is one of many corpora in ESDS and if data cleansing is possible on it, then a potentially large collection of natural speech corpora could become available from ESDS.

37. Separation and Enhancement of Reverberant Speech Mixtures using Binaural cues, Statistical properties and Precedence effect

Atiyeh Alinaghi, Wenwu Wang & Philip Jackson

Underdetermined reverberant speech separation is a challenging problem in source separation that has received considerable attention in both computational auditory scene analysis (CASA) and blind source separation (BSS). Recent studies suggest that, in general, the performance of frequency domain BSS methods suffer from the permutation problem across frequencies which degrades in high reverberation, meanwhile, CASA methods perform less effectively for closely spaced sources.

This paper presents a method to address these limitations, based on the combination of binaural and BSS cues for the automatic classification of time-frequency (T-F) units of the speech mixture spectrogram. By modeling the interaural phase difference, the interaural level difference and frequency-bin mixing vectors, we integrate the coherent information for each source within a probabilistic framework. The Expectation- Maximization (EM) algorithm is then used iteratively to refine the soft assignment of T-F regions to sources and re-estimate their model parameters. The coherence between the left and right recordings is also calculated to model the precedence effect which is then incorporated to the algorithm to reduce the effect of reverberation.

Binaural room impulse responses for 5 different rooms with various acoustic properties have been used to generate the source images and the mixtures. The proposed method compares favorably with state-of-the- art baseline algorithms by Mandel et al. and Sawada et al., in terms of signal-to-distortion ratio (SDR) of the separated source signals.

38. Single-Channel Speaker Separation Using Visual Speech Features

Faheem Khan, Ben Milner

This work proposes a method of single-channel speaker separation that uses visual speech information to extract a target speaker,Äôs voice from a mixture of speakers. The method requires a single audio input and visual inputs from each speaker in the mixture. The visual information from speakers is used to create a visually-derived Wiener filter. The Wiener filter gains are then non-linearly adjusted by a perceptual gain transform to improve the quality and intelligibility of the target speech. Experimental results are presented that measure the quality and intelligibility of the extracted target speaker and a comparison made of different perceptual gain transforms. These show that significant gains are achieved by the application of the perceptual gain function.

39. Single-Channel Speech Separation, Speech Enhancement and Bandwidth Expansion Based on Composition of Longest Segments - CLOSE

Ji Ming, Danny Crookes

This presentation introduces our recent work on a new method for single-channel based speech separation, speech enhancement and speech bandwidth expansion. We call our new method CLOSE, which is an abbreviation of Composition of Longest Segments.

Given a single-channel mixture of two speech utterances, the separation of the underlying constituent speech utterances is achieved by seeking the longest mixed speech segments that can be accurately matched by composite training segments. The longer the mixed segments match, the more specific the constituent training segments. Therefore separation based on the longest matching segments reduces the error of separation. We further extend the method to speech enhancement by finding the longest matching speech segments from a clean speech corpus. Longer speech segments can be identified more accurately in the presence of noise than shorter speech segments. Therefore enhancement based on the longest matching speech segments improves the ability of dealing with nonstationary noise which can be difficult to predict. Finally, we extend the new method to bandwidth expansion of speech subject to bandwidth reduction, resulting from the codec effect in wireless (e.g., Bluetooth) applications. We assume no prior knowledge about the codec characteristics. All experiments are conducted using large-vocabulary speech data with realistic noise and codec effects. The results are evaluated using both objective and subjective tests, including the challenge of large-vocabulary speech recognition.

40. Speech Graphics audio-driven facial animation demo

Michael Berger, Gregor Hofer

Speech Graphics are emerging as the leading provider of multilingual lip sync technology for highly realistic next-generation video games, avatars and simulations. An award-winning spinoff from the University of Edinburgh's School of Informatics and the Centre for Speech Technology Research, our technology combines state-of-the-art speech recognition techniques with modelling of articulation dynamics, facial modelling and computer graphics.

41. The BUDS POMDP Spoken Dialogue System

Matthew Henderson, Martin Szummer, Catherine Breslin, Milica Gasic, Dongho Kim, Blaise Thomson, Pirros Tsiakoulis, Steve Young

Bayesian update of dialogue state (BUDS) is a state-of-the art system for human-computer conversation in dialogues. Here, it is employed to build a speech-driven intelligent assistant. The system manages the conversation to help the user achieve their goal as quickly as possible. The main challenge is to converse in a way that overcomes mistakes made by the speech recognizer, or ambiguous utterances by the user. The system can ask for confirmations, pose choices, and ask for additional information, all in order to gain certainty while maximizing dialogue utility.

The system contains a long machine learning pipeline. It preserves a large number of speech recognition hypotheses by representing them as a confusion network (a compact form of an HMM lattice), and applies a semantic decoder directly to this network. The dialogue state is tracked via a Dynamic Belief Network. The system chooses actions according to a policy that has been learned using a POMDP. The ability of the system to maintain uncertainty significantly improves dialogue utility compared to rule-based dialogue systems.

42. The role of accent in fast adaptation for automatic speech recognition

Maryam Najafian, Martin Russell, Mike Carey

Is the notion of "regional accent" useful for adapting a speech recognition system to a new user? From previous research we know that it is possible to identify an individual's accent with an accuracy of 97% using just 30 seconds of his or her speech [1]. This research asks how this can be exploited for rapid speaker adaptation. Since the techniques that are applied to accent recognition in [1] involve representing an utterance as a 'supervector' in a high-dimensional vector space, this approach is related to previous work on "eigenvoices" [2].

This poster has three sections. In the first part, we will describe techniques to visualise the space where accent recognition is performed and show the extent to which the structure which emerges is consistent with subjective notions of accent. Then we will present speech recognition results on accented speech and interpret them in terms of the structure of the "accent recognition" space. Next, we will present base line results on accent adaptation using conventional adaptation techniques. Finally we will suggest future work on how accent recognition techniques can be used for rapid speaker adaptation.

[1] Hanani, A., M.J. Russell, and M.J. Carey, "Human and computer recognition of regional accents and ethnic groups from British English speech", *Computer Speech & Language (27) 2013, 59-74.*

[2] Kuhn, R., Nguyen, P., Junqua, J-L., Goldwasser, L. Niedzielski, N., Fincke, S., Field, K. And Contolini, M., "Eigenvoices for speaker adaptation", Proc. ICSLP. 1998.

43. Transcription of multi-genre media archives using out-of-domain data

Peter Bell, Mark Gales, Pierre Lanchantin, Xunying Liu, Yanhua Long, Steve Renals, Pawel Swietojanski, Phil Woodland

We describe our work on developing a speech recognition system for multi-genre media archives. The high diversity of the data makes this a challenging recognition task, which may benefit from systems trained on a combination of in-domain and out-of-domain data. Working with tandem HMMs, we present Multi-level Adaptive Networks (MLAN), a novel technique for incorporating information from out-of-domain posterior features using deep neural networks. We show that it provides a substantial reduction in WER over other systems, with relative WER reductions of 15% over a PLP baseline, 9% over in-domain tandem features and 8% over the best out-of-domain tandem features.

44. Transformation of the Glottal Parameters of an HMM-based Speech Synthesiser for Controlling Voice Quality

João P. Cabral, Julie Carson-Berndsen

One of the great challenges for state-of-the-art text-to-speech (TTS) synthesis systems is to allow the control over voice characteristics while preserving the high-quality of the synthetic speech. Such flexibility is necessary for emerging applications of TTS systems which expect a wide variety of speaker's voices or expressive speech, such as computer games, computer dialogue systems, audiobooks, and computer interfaces for people who have speaking impairments.

HMM-based speech synthesis offers great flexibility for controlling voice characteristics by directly modifying the parameters generated by the trained voice models (HMMs) to produce speech or by using adaptation algorithms to transform the HMMs from a relatively small amount of speech data associated with the target voice type.

However, the speech parameters used to train the HMMs generally do not represent aspects related to the glottal source component of the human speech production system, which are important for voice quality. This limitation can be explained by the increased complexity of a speech model incorporating the glottal source compared with a model which uses a simpler excitation, such as the impulse train, and the difficulty to accurately estimate the glottal parameters from the speech signal at the analysis stage.

Recently, an acoustic glottal source model was successfully incorporated into a HMM-based speech synthesiser [1]. In the present work, the advantages of using the glottal parameters by this system for synthesising speech with a creaky voice are investigated.

[1] Cabral, J. P., Renals, S., Richmond, K. and Yamagishi, J., "HMM-based speech synthesiser using the LF-model of the glottal source", Proc. of the ICASSP, Prague, May, 2011.

45. Unsupervised Domain Adaptation

Raymond Ng, Thomas Hain, Matt Gibson

Language model and wordlist adaptation are studied for unsupervised domain adaptation. The goal of the task is to adapt an ASR system such that it exhibits improved performance upon a test set which is sourced from a previously unseen domain. Unsupervised adaptation of the background language model with relevant text, which comes as powerpoint slides of some lecture data, is compared with full adaptation where in-domain training and development text are given. For wordlist adaptatin, pronunciation rules are derived for 174 OOV words which account for 40% of all OOV instances in the training set. Incorporation of these words in the language model training gives a moderate WER reduction. Also, some analysis of abbreviated words is done.

46. Unsupervised Learning for Text-to-Speech Synthesis

Oliver Watts, Junichi Yamagishi, Simon King

We present a general method for incorporating the distributional analysis of textual and linguistic objects into text-to-speech (TTS) conversion systems. Conventional TTS conversion uses intermediate layers of representation to bridge the gap between text and speech. Collecting the annotated data needed to produce these intermediate layers is a far from trivial task, possibly prohibitively so for languages in which no such resources are in existence. Distributional analysis, in contrast, proceeds in an unsupervised manner, and so enables the creation of systems using textual data that are not annotated. The method therefore aids the building of systems for languages in which conventional linguistic resources are scarce, but is not restricted to these languages. The distributional analysis presented here places the textual objects analysed in a continuous-valued space, rather than specifying a hard categorisation of those objects. This space is then partitioned during the training of acoustic models for synthesis, so that the models generalise over objects' surface forms in a way that is acoustically relevant.

47. Using HMM-based Speech Synthesis to Reconstruct the Voice of Individuals with Degenerative Speech Disorders

Christophe Veaux, Junichi Yamagishi, Simon King

When individuals lose the ability to produce their own speech, due to degenerative diseases such as motor neuron disease (MND) or Parkinson's, they lose not only a functional means of communication but also a display of their individual and group identity. In order to build personalized synthetic voices, attempts have been made to capture the voice before it is lost, using a process known as voice banking. But, for some patients, the speech deterioration frequently coincides or quickly follows diagnosis. Using HMM-based speech synthesis, it is now possible to build personalized synthetic voices with minimal data recordings and even disordered speech. In this approach, the patient's recordings are used to adapt an average voice model pre-trained on many speakers. The structure of the voice model allows some reconstruction of the voice by substituting some components from the average voice in order to compensate for the disorders found in the patient's speech. We have now started a collaborative project to move from research prototype to large-scale clinical trial.

48. View Independent Computer Lip-Reading

Yuxuan Lan, Barry-John Theobald, Richard Harvey

Computer lip-reading systems are usually designed to work using a full-frontal view of the face. However, many human experts tend to prefer to lip-read using an angled view. In this paper we consider issues related to the best viewing angle for an automated lip-reading system. In particular, we seek answers to the following questions: 1) Do computers lip-read better using a frontal or a non-frontal view of the face? 2) What is the best viewing angle for a computer lip-reading system? 3) How can a computer lip-reading system be made to work independently of viewing angle? We investigate these issues using a purpose built audio-visual dataset that contains simultaneous recordings of a speaker reciting continuous speech at five angles.

We find that the system performs best on a non-frontal view, perhaps because lip gestures, such as lip-protrusion and lip-rounding, are more pronounced when viewing from an angle. We also describe a simple linear mapping that allows us to map any view of the face to the view that we find to be optimal. Hence we present a view-independent lip-reading system.

49. VISQOL: The Virtual Speech Quality Objective Listener

Andrew Hines

The Virtual Speech Quality Objective Listener (ViSQOL) model is a signal based full reference metric that uses a spectro-temporal measure of similarity between a reference and a test speech signal to provide an objective measure for predicting subjective quality. This poster describes the algorithm and compares the results for VISQOL and PESQ for common problems in VoIP: clock drift, associated time warping and jitter. The results indicate that ViSQOL is less prone to underestimation of speech quality in both scenarios than the ITU standard.

50. Voice prosody: a holistic approach

Irena Yanushevskaya, Christer Gobl, Ailbhe Ní Chasaide

The poster illustrates a fine grained analysis (a micro study) of voice source dynamics in linguistic (focal accentuation and deaccentuation) and paralinguistic prosody. The findings suggest that focal accentuation involves dynamic shifts in phonatory quality both at the utterance level (the relative phonatory strength of sentence constituents) and at the syllable level (enhancement of CV contrast in focal syllable). Similar kinds of shifts are involved in paralinguistic signalling, along with a more 'global' shifts from the neutral baseline.