UK Speech

UK Speech Conference

Edinburgh

9–10 June 2014

THE UNIVERSITY *of* EDINBURGH

Schedule

Monday

10:30–11:25	Badge pick-up and tea
11:25–11:30	Welcome
11:30–13:00	Tutorial, Heiga Zen
13:00–14:30	Lunch
14:30–15:30	Posters
15:30–16:00	Tea
16:00–17:30	Tutorial, Arnab Ghoshal
17:30–18:30	Drinks reception

Tuesday

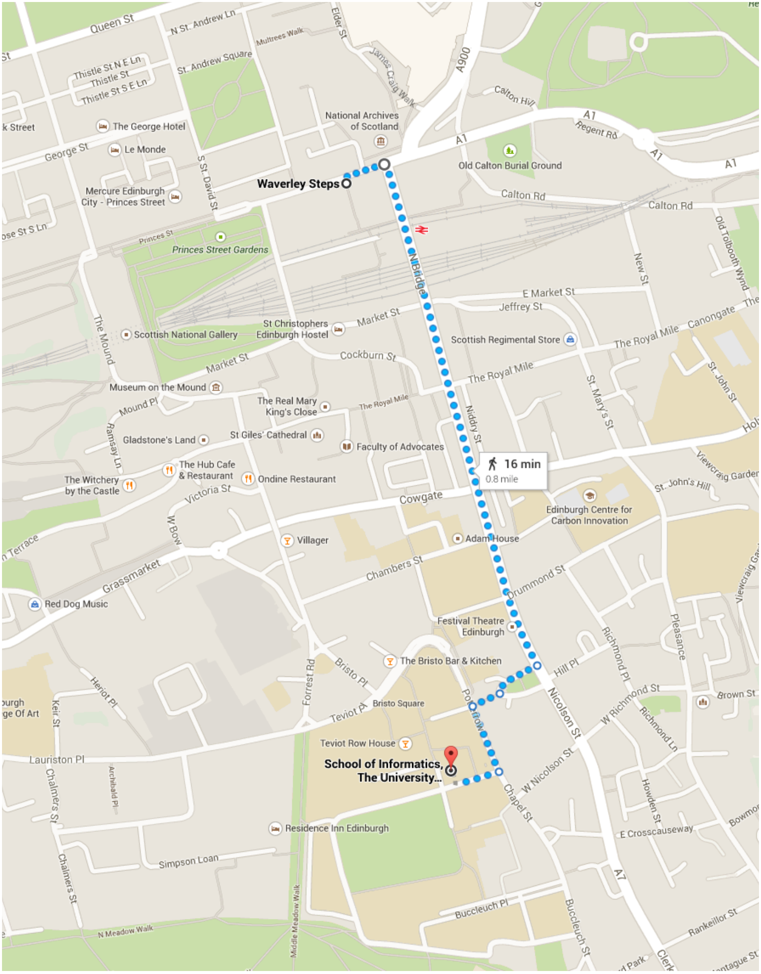
9:30–11:00	Talk, Alessandro Vinciarelli
11:00–11:30	Tea
11:30–12:30	Posters
12:30–13:30	Lunch
13:30–14:30	Posters
14:30–15:00	Tea
15:00–16:00	Panel session - Insights on an Academic Career Path
16:00–16:15	Conclusion/discussion

Map

All events will take place in the School of Informatics, University of Edinburgh:

Informatics Forum / 10 Crichton Street / Edinburgh EH8 9AB

walking directions from the train station:



Site descriptions

The University of Edinburgh
Centre for Speech Technology Research (CSTR)
Institute for Language, Cognition, and Computation (ILCC)
http://www.cstr.ed.ac.uk – http://www.ilcc.inf.ed.ac.uk

The Institute for Language, Cognition, and Computation (ILCC) is one of six research institutes in the School of Informatics, comprising about 150 researchers with expertise in computational linguistics, speech processing, dialogue systems machine learning, multimodal interaction, and cognitive science. Within ILCC, and linking Linguistics, the Centre for Speech Technology Research (CSTR) is an interdisciplinary research centre comprising about 35 researchers (including PhD students, research staff, and teaching staff), plus a number of visiting researchers. CSTR is concerned with research in all areas of speech technology including speech recognition, speech synthesis, speech signal processing, multimodal interaction, and speech perception.

Current research themes at CSTR include bridging the gap between recognition and synthesis, neural networks for acoustic modelling and language modelling in speech recognition and synthesis, the development of factorised acoustic and language models, robustness and adaptivity across domains (e.g. accent, task, and acoustic environment), the development of personalised speech technology systems, and modelling conversational interaction and social cues.

Application areas include voice reconstruction and personalised speech synthesis for assistive technology devices, modelling and tracking for real-time ultrasound-based speech therapy, transcription and subtitling of television and radio, speech translation, and the development of multimodal conversational systems. In addition to industry collaborations, there are also a number of startup and spinout companies associated with CSTR including Cereproc, Quorate, and Speech Graphics.

Apple

Siri is a personal assistant with a voice-controlled natural-language interface that has been an integral part of iOS since 2011. The idea is that Siri will "Understand what you say, and know what you mean". It already works "annoyingly well" [Charlie Brooker] but as you might guess, it doesn't yet do everything you might hope for. Excellent automatic speech recognition is absolutely key to Siri. The Siri team develops and applies large scale systems, spoken language, big data, and artificial intelligence in the service of "the next revolution in human-computer interaction". Apple's growing Siri team is based in Cupertino California, with outposts in Cambridge Massachusetts and Cheltenham Gloucestershire. The Cheltenham team is led by John Bridle and Melvyn Hunt.

CereProc

CereProc, founded in 2005, creates text-to-speech solutions for any type of application. Our core product, CereVoice, is available on any platform, from mobile and embedded devices to desktops and servers. Our voices have character, making them appropriate for a far wider range of applications than traditional text-to-speech systems. Our voices sound engaging when reading long documents and web pages, and add realistic, emotional, voices to animated characters. CereProc has assembled a leading team of speech experts, with a track record of academic and commercial success. We partner with a range of companies and academic institutions to develop exciting new markets for text-to-speech. CereProc works with our language partners to create new versions of CereVoice in any language. www.cereproc.com

Google

Google is full of smart people working on some of the most difficult problems in computer science today. Most people know about the research activities that back our major products, such as search algorithms, systems infrastructure, machine learning, and programming languages. Those are just the tip of the iceberg; Google has a tremendous number of exciting challenges that only arise through the vast amount of data and sheer scale of systems we build. What we discover affects the world both through better Google products and services, and through dissemination of our findings by the broader academic research community. We value each kind of impact, and often the most successful projects achieve both.

Herriot-Watt University

The Interaction Lab at Heriot-Watt University is nearly 5 years old, and is a major research group in Computer Science. It consists of 4 faculty, 8 postdoctoral researchers, and 4 PhD students, and has been a partner in 7 european projects, twice as coordinator. Its mission is to develop intelligent conversational agents which can collaborate effectively and adaptively with humans, by combining a variety of interaction modalities, such as speech, graphics, gesture, and vision. We focus on data-driven machine learning approaches, as well as evaluation of speech and multimodal interfaces with real users. We work with companies such as Yahoo!, BMW, and Orange Labs, to design new conversational speech interfaces. We also do significant work in Human-Robot Interaction. In 2014/2015 we are offering a new masters course in AI with Speech and Multimodal Interaction. www.macs.hw.ac.uk/InteractionLab

http://www.macs.hw.ac.uk/cs/pgcourses/aiws.htm

Imperial College
London

Speech and Audio Processing
Communications and Signal Processing Research Group
Dept. Electrical and Electronic Engineering
Imperial College London

A team of about 10 researchers in the EEE department at Imperial College are working on speech, audio and acoustic signal processing. The technical bases for our work include adaptive signal processing, system identification, speech production analysis and modeling. Our current projects target robot audition and dereverberation. We are aiming to be able to apply dereverberation both for speech recognition and for telecommunications, employing techniques including blind acoustic system identification, system inversion and LPC-based approaches. We have also been working recently on speech processing for law enforcement applications in which the noise levels are severe, aiming to measure and enhance speech intelligibility and quality. Much of our work includes multichannel speech data, and we have are studying spherical microphone arrays for this purpose.

Queen's University
Belfast

Novel Methods for Speech Enhancement Separation and Speaker Recognition
Ji Ming, Darryl Stewart, Danny Crookes
Institute of Electronics, Communications and Information Technology
Queen's University Belfast

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.

Quorate Technology
Quorate Technology is a spin-out from Edinburgh University's Centre for Speech Technology Research (CSTR). The company aims to commercialise the outcomes of the EU-funded AMI/AMIDA research projects through its Automatic Speech Recognition and Analysis suite. The software is targeted towards recognising natural speech involving multiple speakers and it can be adapted to suit a range of different domains. Quorate Technology is based within Edinburgh University's Knowledge Transfer & Commercialisation Suite and the company retains a close working relationship with the School of Informatics in general - and the CSTR in particular.

Trinity College Dublin
The Signal Processing Media Applications Group (Sigmedia) is a research group in the Department of Electronic and Electrical Engineering at Trinity College Dublin in Ireland. Dr. Naomi Harte leads the group. Our research activities are centred on digital signal processing technology. We exploit knowledge from statistics, applied mathematics, computer vision, image and video processing, and speech and language understanding in order to solve very unique problems in a range of domains. The complete group has 3 academics, 4 Research Fellows and 11 PhD students at present. Human Speech Communication is a major theme for the group with active research projects in:

- Audio-visual speech recognition
- Speaker recognition and vocal ageing
- Emotion and affect in speech
- New metrics for speech and audio quality broadcast over the internet
- Forensic analysis of birdsong for species identification

Current projects are funded by Science Foundation Ireland, IRCSET, Enterprise Ireland and Google. Our website at www.sigmedia.tv gives an overview of our research. Please email Naomi Harte at nharte@tcd.ie for further information.

University of Birmingham

The Speech & Language Technology group currently consists of two full-time academics, Prof Martin Russell and Dr Peter Jančovič, plus portions of few other academics, plus three postdocs and nine PhD students. We are part of a larger research group called 'Interactive System Engineering (ISE)' and collaborate with other schools in the university, in particular Psychology.

We are active in five main research areas at the moment, funded by the EU and UK funding bodies, UK government and UK and non-UK companies and parts internally by the university:

1. Speech Recognition by Synthesis - development of more compact acoustic speech models, that incorporate more faithful speech knowledge/structure and rely less on estimating large numbers of parameters from data,

2. Children's Speech - speech recognition and paralinguistic processing of children's speech,

3. Regional Accents - implications of regional accents for speech recognition, including collection of the ABI and ABI-2 corpus of accented British English speech,

4. Bird Sound and Music Analysis - recognition of bird species, modelling of bird vocalisations and songs, and analysis of style through ornamentation in music,

5. Non-audio Application of Speech Algorithms - applying methods from speech recognition to development of technology for rehabilitation of stroke patients in CogWatch project.

Speech Research Group, University of Cambridge

The Speech Research Group in Cambridge is part of the Machine Intelligence Laboratory in the Department of Engineering. Its mission is to advance our knowledge of computer-based spoken language processing and develop effective algorithms for implementing applications. Its primary specialism is in large vocabulary speech transcription and related technologies. It also has active research interests in spoken dialogue systems, multimedia document retrieval, statistical machine translation, speech synthesis and machine learning.

University College London

The Department of Speech, Hearing and Phonetic Sciences (SHaPS) at UCL currently employs 9 academic staff and 6 postdoctoral researchers. It is internationally recognised for the excellence of its research into the perception and production of speech, and in applications of speech technology. We combine basic research into the normal mechanisms of speech and hearing, including adaptation to noisy and distorted channels, with applied research into problems caused by hearing impairment, by atypical perceptual and cognitive development, and by second language use. Our work uses a range of methodologies, including behavioural experimentation, computational modelling, acoustic analysis and neuro-imaging. Speech technology expertise covers speech synthesis and recognition, voice conversion and voice measurement techniques, applied to audiovisual speech synthesis in assistive technology for hearing impaired people and in therapy for schizophrenia. Our research laboratory includes air-conditioned listening and recording rooms with state of the art equipment, an anechoic chamber and facilities for EEG, ABR (Auditory Brainstem Response) and TMS (transcranial magnetic stimulation) measurements. Within UCL we have particularly close links to the Ear Institute and the speech group of the Institute of Cognitive Neuroscience, as well as to neighbouring research departments in Linguistics, Language & Communication, and Developmental Science.

University of East Anglia

The Speech Group at UEA currently consists of four faculty members, two Research Associates and ten PhD students. The Group has been active in fundamental research into speech processing algorithms (e.g. speech recognition in noise, speech enhancement, speaker adaptation, confidence measures for speech recognition) and development of applications of speech processing (e.g. call-routing, recognition of speech transmitted using VOIP, dysarthric speech) for many years. More recently, we have been investigating incorporating visual information into several aspects of speech and audio processing. An important current focus is research into automatic lip-reading algorithms, which has been funded by the EPSRC and the Home Office. We are also interested in exploiting visual speech information to improve traditionally audio-only methods of speech enhancement and speaker separation, and in combining audio and visual information to "understand" events such as sports game (EPSRC funding). We have also been active in developing the use of avatars for sign-language, and our research into avatar speech animation is developing avatars that are capable of expressive speech. We have collaborations with Apple and Disney Research as well as with many small companies.

University of Sheffield

The Speech and Hearing Research Group (SpandH) was established in the Department of Computer Science, University of Sheffield, in 1986. Since then, it has gained an international reputation for research in the fields of computational hearing, speech perception, speech technology and its applications. The group is concerned with:

- Computational modelling of auditory and speech perception in humans and machines
- Robustness in speech recognition
- Large vocabulary speech recognition systems and their applications
- Clinical applications of speech technology

An aspect of the group which makes it unique in the United Kingdom is the wide spectrum of research topics covered, from the psychology of hearing through to the engineering of large vocabulary speech recognition systems. It is our belief that studies at different points on this Science to Engineering axis can and should be mutually beneficial.

University of Surrey
Two University of Surrey groups that host speech research are the Centre for Vision, Speech and Signal Processing (CVSSP) and the Institute of Sound Recording (IoSR). In the Department of Electronic Engineering, CVSSP (surrey.ac.uk/cvssp) is a prime centre for audio-visual signal processing & computer vision in Europe with: over 130 researchers, £12M grant portfolio, track-record of pioneering research leading to technology transfer in collaboration with UK industry, world-class audio and video facilities. CVSSP's Machine Audition Group pursues research into sparse audio-visual dictionary learning, source separation and localisation, articulatory modelling for automatic speech recognition, audio-visual emotion classification, speaker tracking and visual speech synthesis, plus robust techniques for spatial audio. Research at the Institute of Sound Recording (iosr.surrey.ac.uk) focuses on psychoacoustic engineering: exploring the connections between acoustic parameters and perceptual attributes, including overall quality and listener preference. This then drives the development of mathematical and computational models of human auditory perception, and of perceptually-motivated audio tools for use with speech, as well as with music and other audio signals. These two groups form part of a Surrey-led consortium recently awarded a five-year EPSRC programme grant to investigate 3D spatial audio for home environment.

Forensic Speech Science research group
University of York
With the practitioner in mind, the Forensic Speech Science research group targets a range of contexts and considerations encountered in legal casework. The group explores how phonetics and acoustics can further inform the use of speech evidence under variable conditions and from numerous perspectives. Work may include building resources and developing current analytical methodologies to approach the variable challenges posed by forensic speech data. This requires a wide combination of subfields including phonetics, acoustics, sociolinguistics, statistics and speech technology. Examples of current and recent projects involve highlighting the effects of physical barriers on a speech signal, lay persons' perception of speech, and the use of population data and likelihood ratios for analysing and presenting expert evidence. The research group closely follows current real-life casework and up-to-date methods as it includes staff and students from the University of York as well as members of J P French Associates, Forensic Speech and Acoustics Laboratory.

VocalIQ
VocalIQ is a spin-out company from the dialogue systems group at Cambridge University. Our goal is to enable people to speak effectively with their devices; smartphones, smart TVs, cars, or robots. We are building a software platform that makes voice interfaces easy to develop and adaptive to the users. It is a machine-learning based system that includes speech recognition, natural language understanding, tracking the user's intentions, and automatically determines the most appropriate response back to the user. Online learning allows the system to optimise these components automatically, which reduces development and maintenance costs, and provides the ability to continue to improve the user experience whilst the system is operational. Our team has successfully participated in several international evaluations of dialogue systems. We are committed to being involved with the research community via joint research grants and internships. VocalIQ has recently received venture investment, and we are actively looking for speech, NLP, machine learning and general software talent and collaboration.

Tutorials

Statistical parametric speech synthesis

Heiga Zen, Google
Heiga Zen received his PhD from the Nagoya Institute of Technology, Nagoya, Japan, in 2006. Before joining Google in 2011, he was an Intern/Co-Op researcher at the IBM T.J. Watson Research Center, Yorktown Heights, NY (2004–2005), and a Research Engineer at Toshiba Research Europe Ltd. Cambridge Research Laboratory, Cambridge, UK (2008–2011). His research interests include statistical speech synthesis and recognition. He was one of the original authors and the first maintainer of the HMM-based speech synthesis system, HTS (http://hts.sp.nitech.ac.jp).

Statistical parametric speech synthesis has grown in popularity over the last years. In this tutorial, its system architecture is outlined, and then basic techniques used in the system, including algorithms for speech parameter generation, are described with simple examples.

The Kaldi Speech (Recognition) Toolkit

Arnab Ghoshal, Apple
Arnab Ghoshal is a Research Scientist at Apple. Prior to Apple, he was a Research Associate at The University of Edinburgh, UK, from 2011 to 2013, and a Marie Curie Fellow at the Saarland University, Saarbrcken, Germany, from 2009 to 2011, during which he made significant contributions to the development of the Kaldi toolkit. He received the B.Tech degree from the Indian Institute of Technology, Kharagpur, India, and the MSE and PhD degrees from the Johns Hopkins University, Baltimore, USA. His primary research interests include acoustic modeling for large-vocabulary automatic speech recognition, multilingual speech recognition, and pronunciation modeling.

This talk will provide an introduction to the Kaldi toolkit. Kaldi was primarily developed as a toolkit for speech recognition research. It is open-source, written in C++ with a modular design, and released under a liberal Apache v2.0 license making it possible for anyone to freely use Kaldi in their work and contribute to it. Kaldi implements state-of-the-art techniques used in speech recognition, including deep neural networks, and provides complete recipes for obtaining state-of-the-art results on several commonly-used speech recognition corpora. Kaldi has been used for other tasks like handwriting recognition, and an extension for parametric speech synthesis is currently under development.

Social Signal Processing: Understanding Social Interactions Through Nonverbal Behavior Analysis

Alessandro Vinciarelli, University of Glasgow

Alessandro Vinciarelli is with the University of Glasgow where he is Senior Lecturer (Associate Professor) at the School of Computing Science and Associate Academic at the Institute of Neuroscience and Psychology. His main research interest is in Social Signal Processing, the domain aimed at modelling analysis and synthesis of nonverbal behaviour in social interactions. In particular, Alessandro has investigated approaches for role recognition in multiparty conversations, automatic personality perception from speech, and conflict analysis and measurement in competitive discussions. Overall, Alessandro has published more than 100 works, including one authored book, five edited volumes, and 26 journal papers. Alessandro has participated in the organization of the IEEE International Conference on Social Computing as a Program Chair in 2011 and as a General Chair in 2012, he has initiated and chaired a large number of international workshops, including the Social Signal Processing Workshop, the International Workshop on Socially Intelligence Surveillance and Monitoring, the International Workshop on Human Behaviour Understanding, the Workshop on Political Speech and the Workshop on Foundations of Social Signals. Furthermore, Alessandro is or has been Principal Investigator of several national and international projects, including a European Network of Excellence, an Indo-Swiss Joint Research Project and an individual project in the framework of the Swiss National Centre of Competence in Research IM2. Last, but not least, Alessandro is co-founder of Klewel, a knowledge management company recognized with several awards.

Social Signal Processing is the domain aimed at modelling, analysis and synthesis of nonverbal behaviour in social interactions. The core idea of the field is that nonverbal cues, the wide spectrum of nonverbal behaviours accompanying human-human and human-machine interactions (facial expressions, vocalisations, gestures, postures, etc.), are the physical, machine detectable evidence of social and psychological phenomena non otherwise accessible to observation. Analysing conversations in terms of nonverbal behavioural cues, whether this means turn-organization, prosody or voice quality, allows one to automatically detect and understand phenomena like conflict, roles, personality, quality of rapport, etc. In other words, analysing speech in terms of social signals allows one to build socially intelligent machines that sense the social landscape in the same way as people do. This talk provides an overview of the main principles of Social Signal Processing and some examples of their application.

Panel session

Insights on an Academic Career Path

A panel session with Roger Moore (University of Sheffield), Patrick Naylor (Imperial College London) and Simon King (University of Edinburgh)

This informal session will be chaired by Naomi Harte. The panel will be asked to give their views on issues relevant to careers in academia for all, from the early stage researcher to established academic. Topics touched upon will include a diverse range of issues such as: favourite advice to PhD students, pitfalls for the early stage researcher, converting conference publications to journal papers, finding time to write, best/worst things about being an academic, and the h-index or other metrics. It is hoped that this will be a lively and informal session. Audience participation mandatory!

Posters

Poster session 1: Monday 14:30–15:30

POSTER BOARD 1

Acoustic Data-driven Pronunciation Lexicon for Large Vocabulary Speech Recognition

Liang Lu, The University of Edinburgh
Arnab Ghoshal, The University of Edinburgh
Steve Renals, The University of Edinburgh

Speech recognition systems normally use handcrafted pronunciation lexicons designed by linguistic experts. Building and maintaining such a lexicon is expensive and time consuming. This paper concerns automatically learning a pronunciation lexicon for speech recognition. We assume the availability of a small seed lexicon and then learn the pronunciations of new words directly from speech that is transcribed at word-level. We present two implementations for refining the putative pronunciations of new words based on acoustic evidence. The first one is an expectation maximization (EM) algorithm based on weighted finite state transducers (WFSTs) and the other is its Viterbi approximation. We carried out experiments on the Switchboard corpus of conversational telephone speech. The expert lexicon has a size of more than 30,000 words, from which we randomly selected 5,000 words to form the seed lexicon. By using the proposed lexicon learning method, we have significantly improved the accuracy compared with a lexicon learned using a grapheme-to-phoneme transformation, and have obtained a word error rate that approaches that achieved using a fully handcrafted lexicon.

POSTER BOARD 2

Language Independent and Unsupervised Acoustic Models for Speech Recognition and Keyword Spotting

Kate Knill, Cambridge University
Mark Gales, Cambridge University
Anton Ragni, Cambridge University
Shakti Rath, Cambridge University

Developing high-performance speech processing systems for low-resource languages is very challenging. One approach to address the lack of resources is to make use of data from multiple languages. A popular direction in recent years is to train a multi-language bottleneck DNN. Language dependent and/or multi-language (all training languages) Tandem acoustic models are then trained. This work considers a particular scenario where the target language is unseen in multi-language training and has limited language model training data, a limited lexicon, and acoustic training data without

transcriptions. A zero acoustic resources case is first described where a multi-language AM is directly applied to an unseen language. Secondly, in an unsupervised training approach a multi-language AM is used to obtain hypotheses for the target language acoustic data transcriptions which are then used in training a language dependent AM. 3 languages from the IARPA Babel project are used for assessment: Vietnamese, Haitian Creole and Bengali. Performance of the zero acoustic resources system is found to be poor, with keyword spotting at best 60% of language dependent performance. Unsupervised language dependent training yields performance gains. For one language (Haitian Creole) the Babel target is achieved on the in-vocabulary data.

POSTER BOARD 3

Noise-robust detection of peak-clipping in decoded speech

James Eaton, Department of Electrical and Electronic Engineering, Imperial College, London, UK
Patrick A. Naylor, Department of Electrical and Electronic Engineering, Imperial College, London, UK

Clipping is a commonplace problem in voice telecommunications and detection of clipping is useful in a range of speech processing applications. We analyse and evaluate the performance of three previously presented algorithms for clipping detection in decoded speech in high levels of ambient noise. We identify a baseline method which is well known for clipping detection, determine experimentally the optimized operation parameter for the baseline approach, and use this in our experiments. Our results indicate that the new algorithms outperform the baseline except at extreme levels of clipping and negative signal-to-noise ratios.

POSTER BOARD 4

A Fixed Dimension and Perceptually based Dynamic Sinusoidal Model of Speech

Qiong Hu,University of Edinburgh
Yannis Stylianou, Toshiba Research Europe Ltd
Korin Richmond, University of Edinburgh
Ranniery Maia, Toshiba Research Europe Ltd
Junichi Yamagishi, University of Edinburgh
Javier Latorre, Toshiba Research Europe Ltd

This paper presents a fixed- and low-dimensional, perceptually based dynamic sinusoidal model of speech referred to as PDM (Perceptual Dynamic Model). To decrease and fix the number of sinusoidal components typically used in the standard sinusoidal model, we propose to use only one dynamic sinusoidal component per critical band. For each band, the sinusoid with the maximum spectral amplitude is selected and associated with the centre frequency of that critical band. The model is expanded at low frequencies by incorporating sinusoids at the boundaries of the corresponding bands while at the higher frequencies a modulated noise component is used. A listening test is conducted to compare speech reconstructed with PDM and state-of-the-art models of speech, where all models are constrained to use an equal number of parameters. The results show that PDM is clearly preferred in terms of quality over the other systems.

POSTER BOARD 5

Using Neural Network Front-ends on Far Field Multiple Microphones Based Speech Recognition

Yulan Liu, University of Sheffield, Sheffield, UK
Pengyuan Zhang, Key Laboratory of Speech Acoustics and Content Understanding, IACAS, Beijing, China
Thomas Hain, University of Sheffield, Sheffield, UK

This paper presents an investigation of far field speech recognition using beamforming and channel concatenation in the context of Deep Neural Network (DNN) based feature extraction. While speech enhancement with beamforming is attractive, the algorithms are typically signal-based with no information about the special properties of speech. A simple alternative to beamforming is concatenating multiple channel features. Results presented in this paper indicate that channel concatenation gives similar or better results. On average the DNN front-end yields a 25% relative reduction in Word Error Rate (WER). Further experiments aim at including relevant information in training adapted DNN features. Augmenting the standard DNN input with the bottleneck feature from a Speaker Aware Deep Neural Network (SADNN) shows a general advantage over the standard DNN based recognition system, and yields additional improvements for far field speech recognition.

POSTER BOARD 6

Data augmentation for low resource languages

Anton Ragni, University of Cambridge
Kate Knill, University of Cambridge
Shakti Rath, University of Cambridge
Mark Gales, University of Cambridge

Recently there has been interest in the approaches for training speech recognition systems for languages with limited resources. Under the IARPA Babel program such resources have been provided for a range of languages to support this research area. This paper examines a particular form of approach, data augmentation, that can be applied to these situations. Data augmentation schemes aim to increase the quantity of data available to train the system, for example semi-supervised training, multi-lingual processing, acoustic data perturbation and speech synthesis. To date the majority of work has considered individual data augmentation schemes, with few consistent performance contrasts or examination of whether the schemes are complementary. In this work two data augmentation schemes, semi-supervised training and vocal tract length perturbation, are examined and combined on the Babel limited language pack configuration. Here only about 10 hours of transcribed acoustic data are available. Two languages are examined, Assamese and Zulu, which were found to be the most challenging of the Babel languages released for the 2014 Evaluation. For both languages consistent speech recognition performance gains can be obtained using these augmentation schemes. Furthermore the impact of these performance gains on a down-stream keyword spotting task are also described.

POSTER BOARD 7

Statistical Parametric Speech Synthesis based on Recurrent Neural Networks

Heiga Zen, Google UK
Hasim Sak, Google NYC
Alex Graves, Google DeepMind
Andrew Senior, Google NYC

Neural network-based acoustic modeling has been successfully applied to statistical parametric speech synthesis. This poster presentation reports Google's recent research works for statistical parametric speech synthesis using various types of recurrent neural networks.

POSTER BOARD 8

Charisma in Political Speech

Ailbhe Cullen, Trinity College Dublin
Naomi Harte, Trinity College Dublin

The rise of streaming has enabled political debates and speeches to reach much wider audiences. The challenge for the viewer is to sort through this information, in order to find something appealing, enjoyable, or informative. In this paper, we explore the nature of charisma in political speech, with a view to the automatic detection of charismatic recordings. We present a novel database which has been collated from a variety of on-line sources, containing a wide range of recording and noise conditions. Compared to previous paralinguistic databases, this is more representative of conditions which must be tolerated by real-world systems. A subset of this database has been annotated for four attributes: charisma; likeability; enthusiasm; and inspiration. Preliminary results of regression using these labels are presented, and in light of these results, future plans to annotate a larger portion of the database are discussed.

POSTER BOARD 9

Trajectory Analysis of Speech using Continuous-State Hidden Markov Models

Philip Weber, University of Birmingham
Steve M. Houghton, University of Birmingham
Colin J. Champion, University of Birmingham
Martin J. Russell, University of Birmingham

Peter Jančovič, University of Birmingham

Many current speech models used in recognition involve thousands of parameters, whereas the mechanisms of speech production are conceptually very simple. We present and evaluate a new continuous state probabilistic model (CS-HMM) for recovering dwell-transition and phoneme sequences from dynamic speech production features. We show that with very few parameters, these features can be tracked, and phoneme sequences recovered, with promising accuracy.

POSTER BOARD 10

A New Phase-based Feature Representation for Robust Speech Recognition

Erfan Loweimi, Speech and Hearing Research Group (SpandH), Department of Computer Science, University of Sheffield
Seyed Mohammad Ahadi, Speech Processing Research Laboratory (SPRL), Electrical Engineering Department, Amirkabir University of Technology, Tehran, Iran
Thomas Drugmann, Circuit Theory and Signal Processing Lab (TCTS), Mons, Belgium

The aim of this paper is to introduce a novel phase-based feature representation for robust speech recognition. This method consists of four main parts: autoregressive (AR) model extraction, group delay function (GDF) computation, compression, and scale information augmentation. Coupling GDF with AR model results in a high-resolution estimate of the power spectrum with low frequency leakage. The compression step includes two stages similar to MFCC without taking logarithm from the output energies. The fourth part augments the phase-based feature vector with scale information which is based on the Hilbert transform relations and complements the phase spectrum information. In the presence of additive and convolutional noises, the proposed method has led to 15% and 12% reductions in the averaged error rates, respectively (SNR ranging from 0 to 20 dB), compared to the standard MFCCs.

POSTER BOARD 11

Speech Enhancement by Speech Reconstruction Using Hidden Markov Models

Akihiro Kato, University of East Anglia
Ben Milner, University of East Anglia

This work presents an approach to speech enhancement that operates using a speech production model to reconstruct a clean speech signal from a set of speech parameters that are estimated from the noisy speech. The motivation is to remove the distortion and residual and musical noises that are associated with conventional filtering-based methods of speech enhancement. The STRAIGHT vocoder forms the model for speech reconstruction and requires a time-frequency surface and fundamental frequency information. Hidden Markov model synthesis is used to create an estimate of the time-frequency surface and this is combined with the noisy surface using a perceptually motivated signal-to-noise ratio weighting. Experimental results compare the

proposed reconstruction-based method to conventional filtering-based approaches of speech enhancement.

POSTER BOARD 12

Paraphrastic Neural Network Language Models

Xunying Liu, University of Cambridge, United Kingdom
Mark Gales, University of Cambridge, United Kingdom
Phil Woodland, University of Cambridge, United Kingdom

Expressive richness in natural languages presents a significant challenge for statistical language models (LM). As multiple word sequences can represent the same underlying meaning, only modelling the observed surface word sequence can lead to poor context coverage. To handle this issue, paraphrastic LMs were previously proposed to improve the generalization of back-off n-gram LMs. Paraphrastic neural network LMs (NNLM) are investigated in this paper. Using a paraphrastic multi-level feedforward NNLM modelling both word and phrase sequences, significant error rate reductions of 1.3% absolute (8% relative) and 0.9% absolute (5.5% relative) were obtained over the baseline n-gram and NNLM systems respectively on a state-of-the-art conversational telephone speech recognition system trained on 2000 hours of audio and 545 million words of texts.

POSTER BOARD 13

Unsupervised Model Selection for Recognition of Regional Accented Speech

Maryam Najafian, University of Birmingham
Martin Russell, University of Birmingham

This paper is concerned with automatic speech recognition (ASR) for accented speech. Given a small amount of speech from a new speaker, is it better to apply speaker adaptation to the baseline, or to use accent identification (AID) to identify the speaker's accent and select an accent-dependent acoustic model? Three accent-based model selection methods are investigated: using the "true" accent model, and unsupervised model selection using i-Vector and phonotactic-based AID. All three methods outperform the unadapted baseline. Most significantly, AID-based model selection using 43s of speech performs better than unsupervised speaker adaptation, even if the latter uses five times more adaptation data. Combining unsupervised AIDbased model selection and speaker adaptation gives an average relative reduction in ASR error rate of up to 47%.

POSTER BOARD 14

The Effect of Encoding and Equipment on Perceived Audio Quality

Andrew Hines, Trinity College Dublin

Naomi Harte, Trinity College Dublin

Subjective listener tests provide the ground truth data necessary to develop objective models for speech and audio quality. For streaming audio, channel bandwidth usage is conserved using lossy compression schemes and a perceived link between bit rate and quality is commonly reported. This work investigated this link along with the additional factor of presentation hardware. MUSHRA tests were used to assess a number of audio codecs and bit rates typically used by streaming services. Three presentation modes were used, namely consumer and studio quality headphones and loudspeakers. Listeners with consumer quality headphones could not differentiate between codecs with bit rates greater than 48 kb/s. For studio quality headphones and loudspeakers 128 kb/s and higher was differentiated over other codecs. The results provide insights into quality of experience that will guide future development of objective audio quality metrics.

POSTER BOARD 15

Combining Tandem and Hybrid Systems for Improved Speech Recognition and Keyword Spotting on Low Resource Languages

Shakti Rath, Cambridge University Engineering Department
Kate Knill, Cambridge University Engineering Department
Anton Ragni, Cambridge University Engineering Department
Mark Gales, Cambridge University Engineering Department

In recent years there has been significant interest in Automatic Speech Recognition (ASR) and Key Word Spotting (KWS) systems for low resource languages. One of the driving forces for this research direction is the IARPA Babel project. This paper examines the performance gains that can be obtained by combining two forms of deep neural network ASR systems, Tandem and Hybrid, for both ASR and KWS using data released under the Babel project. Baseline systems are described for the five option period 1 languages: Assamese; Bengali; Haitian Creole; Lao; and Zulu. All the ASR systems share common attributes, for example deep neural network configurations, and decision trees based on rich phonetic questions and state-position root nodes. The baseline ASR and KWS performance of Hybrid and Tandem systems are compared for both the “full”, approximately 80 hours of training data, and limited, approximately 10 hours of training data, language packs. By combining the two systems together consistent performance gains can be obtained for KWS in all configurations.

POSTER BOARD 16

Avatar Therapy: an audio-visual dialogue system for treating auditory hallucinations

Mark Huckvale, Department of Speech, Hearing and Phonetics, UCL
Geoff Williams, Department of Speech, Hearing and Phonetics, UCL

Julian Leff, Department of Mental Health Sciences, UCL

This paper presents a radical new therapy for persecutory auditory hallucinations ("voices") which are most commonly found in serious mental illnesses such as schizophrenia. In around 30% of patients these symptoms are not alleviated by anti-psychotic medication. This work tackles the problem posed by the inaccessibility of the patients' experience of voices to the clinician. Patients are invited to create an external representation of their dominant voice hallucination in the form of a talking head, or avatar. We use 3D animation technology to give a persona to the voice, and custom real-time voice morphing software to modify the therapist's voice to simulate the internal voice. The therapist then conducts a dialogue between the avatar and the patient, with a view to gradually bringing the avatar, and ultimately the hallucinatory voice, under the patient's control. Results of a pilot study indicate that the approach has potential for dramatic improvements in patient control of the voices after a series of only six short sessions. The focus of this poster is on the audio-visual speech technology which delivers the central aspects of the therapy.

POSTER BOARD 17

Loose Coupling of Speech Recognition and Machine Translation Systems

Raymond W. M. Ng, Department of Computer Science, The University of Sheffield, United Kingdom
Thomas Hain, Department of Computer Science, The University of Sheffield, United Kingdom
Trevor Cohn, Department of Computer Science, The University of Sheffield

Spoken language translation (SLT) is an important problem, that requires a combination of automatic speech recognition (ASR) and machine translation (MT). In previous work we have investigated the case where the acoustic signal is available along with its text translation in another language. We have shown that recognition results in the source language can be improved by coupling the ASR with MT outputs. In this paper we focus on the structure of our loose coupling approach, its efficiency and performance, and extend the approach to full end-to-end SLT. We compare utterancebased coupling with talk-based coupling on the TED lectures dataset, and show that using general knowledge present in translated talks only has a small effect on performance of 1.4% WER absolute. A second set of experiments considered loose coupling approaches for domain adaptation of the MT system. Experimental results indicate that in-domain translation models tuned with the coupled system output have comparable performance to tuning on the reference. Together these findings imply a reduction the data requirements, allowing training of SLT systems on bilingual speech and text corpora without the need for transcripts or strictly parallel translations.

POSTER BOARD 18

Roomprints for forensic audio applications

Alastair H. Moore, Imperial College London
Mike Brookes, Imperial College London
Patrick A. Naylor, Imperial College London

A roomprint is a quantifiable description of an acoustic environment which can be measured under controlled conditions and estimated from a monophonic recording made in that space. We here identify the properties required of a roomprint in forensic audio applications and review the observable characteristics of a room that, when extracted from recordings, could form the basis of a roomprint. Frequency-dependent reverberation time is investigated as a promising characteristic and used in a room identification experiment giving correct identification in 96% of trials.

POSTER BOARD 19

Auditory adaptation to static spectra

Cleo Pike, University of Surrey
Russell Mason, University of Surrey
Tim Brookes, University of Surrey

Auditory adaptation is thought to reduce the perceptual impact of static spectral energy and increase sensitivity to spectral change. Research suggests that this adaptation helps listeners to extract stable speech cues across different talkers, despite inter-talker spectral variation caused by differing vocal tract acoustics. This adaptation may also be involved in compensation for transmission channels more generally (e.g. distortions caused by the room or loudspeaker through which a sound has passed).

The magnitude of this adaptation and its ecological importance has not been established. The physiological and psychological mechanisms behind adaptation are also not well understood. The current research confirmed that adaptation to transmission channel spectrum occurs when listening to speech produced though two types of transmission channel: loudspeakers and rooms. The loudspeaker is analogous to the vocal tract of a talker, imparting resonances onto a sound source which reaches the listener both directly and via reflections. The room-affected speech however, reaches the listener only via reflections - there is no direct path. Larger adaptation to the spectrum of the room was found, compared to adaptation to the spectrum of the loudspeaker. It appears that when listening to speech, mechanisms of adaptation to room reflections, and adaptation to loudspeaker/vocal tract spectrum, may be different.

Poster session 2: Tuesday 11:30–12:30

POSTER BOARD 1

Interpreting voice communications in search and rescue: data collection in simulated environment

Saeid Mokaram, University of Sheffield
Roger K. Moore, University of Sheffield

Radio voice communications is the key element of the C3I infrastructure in any Search and Rescue operation. Clearly accessing to the huge amount of valuable information flowing on these channels will enhance the situation awareness (SA) and decision-making in crisis response system. The main objective of this research is to investigate the solutions for interpreting these voice communications in order to improve and update primary estimations about the lay of the land. Providing suitable speech data set with proper annotations is a preliminary issue in this research. This poster reports the data collection in a simulation system which is designed based on an abstract model of search and rescue two party remote communications. In this model, one participant explores a simulated indoor environment and reports his/her observations and actions back to the other participant which only has access to a rough building map. While the volunteers' voices and environment noise were recorded in separate channels, EX's location, actions and list of the objects in his/her field of view were also recorded simultaneously for annotation. At the early stage of this research, pilot recordings were performed and the main recording phase is in progress.

POSTER BOARD 2

Investigating Automatic and Human Filled Pause Insertion for Speech Synthesis

Rasmus Dall, The Centre for Speech Technology Research, University of Edinburgh
Marcus Tomalin, Cambridge University Engineering Department, University of Cambridge
Mirjam Wester, The Centre for Speech Technology Research, University of Edinburgh
William Byrne, Cambridge University Engineering Department, University of Cambridge
Simon King, The Centre for Speech Technology Research, University of Edinburgh

Filled pauses are pervasive in conversational speech and have been shown to serve several psychological and structural purposes. Despite this, they are seldom modelled overtly by state- of-the-art speech synthesis systems. This paper seeks to motivate the incorporation of filled pauses into speech synthesis systems by exploring their use in conversational speech, and by comparing the performance of several automatic systems inserting filled pauses into fluent text. Two initial experiments are described which seek to determine whether people's predicted insertion points are consistent with actual practice and/or with each other. The experiments also investigate whether there

are ‘right’ and ‘wrong’ places to insert filled pauses. The results show good consistency between people’s predictions of usage and their actual practice, as well as a perceptual preference for the ‘right’ placement. The third experiment contrasts the performance of several automatic systems that insert filled pauses into fluent sentences. The best performance (determined by F-score) was achieved through the by-word interpolation of probabilities predicted by Recurrent Neural Network and 4gram Language Models. The results offer insights into the use and perception of filled pauses by humans, and how automatic systems can be used to predict their locations.

POSTER BOARD 3

Adaptation of Deep Neural Network Acoustic Models Using Factorised I-Vectors

Penny Karanasou, University of Cambridge
Yongqiang Wang, University of Cambridge
Mark J.F. Gales, University of Cambridge
Philip C. Woodland, University of Cambridge

The use of deep neural networks (DNNs) in a hybrid configuration is becoming increasingly popular and successful for speech recognition. One issue with these systems is how to efficiently adapt them to reflect an individual speaker or noise condition. Recently speaker i-vectors have been successfully used as an additional input feature for unsupervised speaker adaptation. In this work the use of i-vectors for adaptation is extended to incorporate acoustic factorisation. In particular, separate i-vectors are computed to represent speaker and acoustic environment. By ensuring “orthogonality” between the individual factor representations it is possible to represent a wide range of speaker and environment pairs by simply combining ivectors from a particular speaker and a particular environment. In this work the i-vectors are viewed as the weights of a cluster adaptive training (CAT) system, where the underlying models are GMMs rather than HMMs. This allows the factorisation approaches developed for CAT to be directly applied. Initial experiments were conducted on a noise distorted version of the WSJ corpus. Compared to standard speaker-based i-vector adaptation, factorised i-vectors showed performance gains.

POSTER BOARD 4

Front-end Filters for Bird Call Feature Extraction

Colm O’Reilly, Trinity College Dublin
Nicola Marples, Trinity College Dublin
Naomi Harte, Trinity College Dublin

Distinguishing the calls and songs of different bird populations is important to Ornithologists. Together with morphological and genetic information, these vocalisations can yield an increased understanding of population diversity. This paper investigates the optimal front-end filterbank used for extracting cepstrum features to classify bird populations. The mel-scale is compared to a linear scale and species specific filterbanks,

optimised by inspecting the spectrum of bird species vocalisations. Experiments are conducted on island populations of Olive-backed Sunbirds and Black-naped Orioles from Indonesia. Results show an improvement in classification rates when using an optimised front-end for each species.

POSTER BOARD 5

Conversational skill development strategies for cochlear implant users

Amy V Beeston, Department of Computer Science, University of Sheffield;
Guy J Brown, Department of Computer Science, University of Sheffield;
Emina Kurtic, Department of Computer Science, University of Sheffield;
Bill Wells, Department of Human Communication Sciences, University of Sheffield
Erica Bradley, Sheffield Teaching Hospitals NHS Foundation Trust
Harriet Crook, Sheffield Teaching Hospitals NHS Foundation Trust

Until recently, many cochlear implant (CI) users would need optimum conditions to hold a satisfactory conversation e.g., a quiet environment, one-to-one setting, and communication awareness to avoid both parties talking at once. Recent technological improvements in CI devices mean that it is now more realistic for users to attempt to engage in natural conversations in which overlapping talk is a common occurrence. However, currently there are no established training materials that hearing professionals can use to help CI users deal with the problem of simultaneous talk. Acoustic analysis of typical turn-taking behaviour has suggested various strategies that normal-hearing listeners employ to manage their conversational exchanges. Some acoustic cues relevant to the management of turn-taking are transmitted through the cochlear implant (e.g. the intensity contour), however, other aspects of the signal that are crucial to a normal-hearing listener's perception and action (e.g., interpreting a rising or falling pitch pattern) still remain inaccessible to listeners using a CI. Drawing material from a pre-recorded audio-visual corpus of natural conversational, our project has begun to devise training materials to promote key conversational competencies in CI-users. We suggest graded tasks to enable CI-users to repeatedly practise (i) crucial listening skills (identifying the main speaker, recognising the semantic content of the speech signal, and understanding the social action underlying the conversational exchange) and (ii) speaking skills fundamental to multi-party conversation (using competitive and non-competitive overlaps appropriately).

POSTER BOARD 6

Unsupervised Learning of Lexical Categories from Speech Using Fixed-Dimensional Acoustic Embeddings

Herman Kamper, University of Edinburgh
Aren Jansen, Johns Hopkins University

Sharon Goldwater, University of Edinburgh

Our long-term aim is to learn lexical and syntactic structure from raw speech without supervision. This requires both an unsupervised acoustic model to relate segments of the speech signal to unidentified word categories and a language model over those categories. In this work we explore a novel lexical acoustic model in which clustering is performed on recently proposed fixed-dimensional embeddings of word segments. We evaluate several clustering algorithms and find that the best methods allow for large variation in cluster sizes, as is inherently the case for natural language. The best probabilistic approach is an infinite Gaussian mixture model (IGMM) which chooses its own number of components. Performance is comparable to that of the non-probabilistic Chinese Whispers and average-linkage hierarchical clustering algorithms, with the latter performing slightly better. We conclude that IGMM clustering on fixed-dimensional embeddings holds promise for unsupervised acoustic modelling.

POSTER BOARD 7

Automatic Speech Recognition Using Neural Networks

Linxue Bai, University of Birmingham
Martin Russell, University of Birmingham
Peter Jančovič, University of Birmingham

Current automatic speech recognition systems rely on very large complex statistical models. To develop new more compact models of speech which require less training corpora and are more transferable, we consider replacing the statistical models with the Neural Networks in the Hidden Markov Models. Meanwhile, we are doing research on new representations of speech using neural networks, especially features with dynamics. This project started at the end of September, 2013.

POSTER BOARD 8

SpeechCity: Conversational Interfaces for Urban Environments

Verena Rieser, Heriot-Watt University
Srini Janarthanam, Heriot-Watt University
Andy Taylor, Heriot-Watt University
Yanchao Yu, Heriot-Watt University
Oliver Lemon, Heriot-Watt University

We demonstrate a conversational interface that assists pedestrian users in navigating and searching urban environments. Locality-specific information is acquired from open data sources, and can be accessed via intelligent interaction. We therefore combine a variety of technologies, including Spoken Dialogue Systems and Geographical Information Systems (GIS) to operate over a large spatial database. In this demo, we present a system for tourist information within the city of Edinburgh. We harvest points of interest from Wikipedia and social networks, such as Foursquare, and we calculate walking directions from Open Street Map (OSM). In contrast to existing mobile applications, our Android agent is able to simultaneously engage in multiple tasks, e.g. navigation and tourist information, by using a multi-threaded dialogue manager. For demonstrating the full functionality of the system, we simulate a (user-specified) walking route, where the system "pushes" relevant information to the user. Through the use of open data, the agent is easily portable and extendable to new locations and domains. Future possible versions of the systems include an Edinburgh Festival app, a tourist guide for San Francisco and the Bay Area, and a conference system for the SemDial'14 workshop (to be held at Heriot-Watt University in September).

POSTER BOARD 9

Modelling hearing impaired-listeners' perception of speaker intelligibility in noise

Lindon Falconer, University of Sheffield
Jon Barker, University of Sheffield
Andre Coy, University of the West Indies

The ability of hearing aids to increase speech intelligibility in multi-source environments is still relatively limited. One of the main problems for developing new algorithms is the time and expense of testing algorithms on human subjects. Further, variability between listeners and the time it takes listeners to acclimatise to new algorithms makes it hard to design robust experiments. A possible solution would be to replace hearing-impaired listeners with a computational model, i.e. a model able to predict a specific listener's judgement of speech intelligibility in given noise conditions. This study is testing the feasibility of this approach. The work uses an auditory-based model of hearing impairment that is able to mimic measured hearing thresholds and loudness recruitment of a specific listener. This is paired with a microscopic intelligibility model that employs statistical speech models and knowledge of the noise background. Can such a system predict the intelligibility judgements of a hearing-impaired listener? If so, is it possible to use this model as a tool for rapid hearing aid signal processing development and evaluation? We will present preliminary results and plans for future results.

POSTER BOARD 10

Efficient Lattice Rescoring Using Recurrent Neural Network Language Models

Xunying Liu, University of Cambridge, United Kingdom
Yongqiang Wang, Cambridge University, United Kingdom
Xie Chen, University of Cambridge, United Kingdom
Mark Gales, University of Cambridge, United Kingdom
Phil Woodland, University of Cambridge, United Kingdom

Recurrent neural network language models (RNNLM) have become an increasingly popular choice for state-of-the-art speech recognition systems due to their inherently strong generalization performance. As these models use a vector representation of complete history contexts, RNNLMs are normally used to rescore N-best lists. Motivated by their intrinsic characteristics, two novel lattice rescoring methods for RNNLMs are investigated in this paper. The first uses an *n*-gram style clustering of history contexts. The second approach directly exploits the distance measure between hidden history vectors. Both methods produced 1-best performance comparable with a 10k-best rescoring baseline RNNLM system on a large vocabulary conversational telephone speech recognition task. Significant lattice size compression of over 70% and consistent improvements after confusion network (CN) decoding were also obtained over the N-best rescoring approach.

POSTER BOARD 11

Speech recognition and related technologies in the inEvent portal

Fergus McInnes, University of Edinburgh
Jean Carletta, University of Edinburgh
Catherine Lai, University of Edinburgh
Steve Renals, University of Edinburgh

The inEvent project (2011-2014) aims to make online multimedia material, such as video recordings of lectures and meetings, more useful and accessible by automatically analysing, annotating, indexing and linking the content. The project has developed a portal for users to browse, search and navigate within and between recordings, and user evaluations are in progress. This poster presents ways in which speech recognition and related technologies (such as speaker diarisation and sentiment analysis) contribute to the process, and the interfaces being developed to present their outputs to the user. The interface to speech recognition output makes use of word confidence scores to present a filtered transcript to the user, and a word cloud can be presented in place of the transcript (in case of low overall confidence) or as a compact summary of the content.

POSTER BOARD 12

Identification of Age-Group from Children's Speech by Computers and Humans

Saeid Safavi, School of Electronic, Electrical & Computer Engineering, University of Birmingham, UK
Martin Russell, School of Electronic, Electrical & Computer Engineering, University of Birmingham, UK
Peter Jančovič, School of Electronic, Electrical & Computer Engineering, University of Birmingham, UK

This paper presents results on age identification (Age-ID) for children's speech, using the OGI Kids corpus and GMM-UBM, GMM-SVM and i-vector systems. Regions of the spectrum containing important age information for children are identified by conducting Age-ID experiments over 21 frequency sub-bands. Results show that the frequencies above 5.5 kHz are least useful for Age-ID. The effect of using gender-independent and gender-dependent age-group modelling is explored. The GMM-UBM and i-vector systems considerably outperform the GMM-SVM system. The best Age-ID performance of 85.77% is obtained by the i-vector system applied to band-limited speech to 5.5 kHz. Experiments on human Age-ID were also conducted and the results show that the humans do not achieve the performance of the machine.

POSTER BOARD 13

Speaker Specific Layer Training for Speaker Adaptation in ASR

Rama Doddipatla, University of Sheffield
Madina Hasan, University of Sheffield
Thomas Hain, University of Sheffield

Speaker adaptation of deep neural networks (DNN) is difficult, and most commonly performed by changes to the input of the DNNs. Here we propose to learn speaker dependent discriminative feature transformations to obtain speaker normalised bottleneck (BN) features. This is achieved by interpreting the final two hidden layers as a speaker specific matrix and update the weights with speaker specific data to learn speaker-dependent discriminative feature transformations. Such simple implementation lends itself to rapid adaptation and flexibility to be used in Speaker Adaptive Training frameworks. The performance of this approach is evaluated on a meeting recognition task, using the official NIST RT'07 and RT'09 evaluation sets. CMLLR adaptation only yields 3.4% and 2.5% relative word error rate (WER) improvement on the RT'07 and RT'09 respectively, where the baselines include speaker based CMVN. The combined CMLLR and BN layer speaker adaptation yields a relative WER gain of 4.5% and 4.2% respectively. SAT style BN layer adaptation is attempted and combined with conventional CMLLR SAT, to show that it provides a relative gain of 1.43% and 2.02% on the RT'07 and RT'09 data sets over CMLLR SAT. While the overall gain from BN layer adaptation is small, the results are found to be statistically significant on both the test sets.

POSTER BOARD 14

An Initial Investigation of Long-Term Adaptation for Meeting Transcription

X. Chen, Cambridge University Engineering Department
M.J.F. Gales, Cambridge University Engineering Department
K. Knill, Cambridge University Engineering Department
C. Breslin, Cambridge University Engineering Department
L. Chen, Toshiba Research Europe Ltd, Cambridge
K. K. Chin, Toshiba Research Europe Ltd, Cambridge
V. Wan, Toshiba Research Europe Ltd, Cambridge

Meeting transcription is a very useful and challenging task. The majority of speech recognition research to date has focused on transcribing individual meetings, or a small set of meetings. In many practical deployments, multiple related meetings will take place over a long period of time. This paper describes an initial investigation of how this

long-term data can be used to improve meeting transcription. A corpus of technical meetings was recorded over a two year period. A microphone array located in the center of the meeting room was used for the data collection. This yielded a total of 179 hours of meeting data. An advanced baseline system based on deep neural network acoustic models, in both Tandem and Hybrid configurations, and neural network-based language models is described. The impact of supervised and unsupervised adaptation of the acoustic models is then evaluated, as well as the impact of improved language models.

POSTER BOARD 15

Dealing with Transcription Errors: Towards Active Learning in Audio Books

Chenhao Wu, University of Sheffield
Thomas Hain, University of Sheffield

The objective of this project is personalise speech recognisers over long periods of time, despite only having access to errorful data. Large amounts of data can help to adapt an ASR system very precisely to a speaker, but this is often only practical to do that with output from the system itself. Errors in the adaptation data labels degrade the performance of adaptation, resulting in poorer results overall. This project investigated the use of information about the errors to steer adaptation, for example to re-transcribe errorful sections. One of the options for dealing with errors is to perform data selection. Active learning is a state-of-the-art sample selection strategy based on the labels' confidence scores. For the experiments we used data from audio book recordings in the public domain. These are especially relevant as they contain a large amount of data from individual speakers.

POSTER BOARD 16

Development and evaluation of an improved Reverberation Decay Tail metric as a measure of perceived late reverberation

Hamza Javed, Dept. of Electrical and Electronic Engineering, Imperial College London, UK
Patrick Naylor, Dept. of Electrical and Electronic Engineering, Imperial College London, UK

In this paper the development and evaluation of an improved Reverberation Decay Tail (RDT) metric is described. The signal-based metric predicts the perceived impact of reverberation on captured speech, by identifying and characterising energy decay curves in the signal Bark spectra. The measure, based on earlier research, is extended to operate on wideband speech and incorporates an improved perceptual model and decay curve detection scheme. Experimental testing of the metric on simulated and recorded reverberant speech shows positive correlation with objective measures such as C50 and subjective listening test scores. Potential applications of the measure include use as a developmental tool for dereverberation research.

POSTER BOARD 17

Semi-Supervised DNN Training in Meeting Recognition

Pengyuan Zhang, Key Laboratory of Speech Acoustics and Content Understanding, IACAS, China
Yulan Liu, University of Sheffield, UK
Thomas Hain, University of Sheffield, UK

Due to domain specificity, there are low resource scenarios where annotated training data can be especially expensive to obtain. Existing research based on advanced DNN front-end utilized semi-supervised training to improve the recognition performance of a seed system which is trained with limited amount of annotated data. In this work, semi-supervised training of two typical low resource scenarios was explored. The performance of semi-supervised training with confidence score based hypothesis transcription selection is verified and extended with analysis on hypothesis label accuracy. By comparing hypothesis labels of different resolution, the semi-supervised training is further improved with an optimal balance between label resolution and accuracy achieved at monophone level.

POSTER BOARD 18

Signal Processing for Embodied Audition for RobotS (EARS)

Christine Evers, Imperial College London
Alastair H. Moore, Imperial College London
Patrick A. Naylor, Imperial College London

The success of natural intuitive human-robot interaction (HRI) depends heavily on effective speech interaction and dialogue systems. However, current limitations in robot audition do not allow for natural acoustic human-robot communication in real-world environments due to the severe degradation of the desired acoustic signals by noise, interference and reverberation when captured by the robot's microphones. To overcome these limitations, the project Embodied Audition for RobotS (EARS), funded by the European Union's Seventh Framework Programme, aims to provide intelligent 'ears' with close-to-human auditory capabilities and use it for HRI in complex real-world environments. Novel microphone arrays and powerful signal processing algorithms will be developed to localise and track multiple sound sources of interest and to extract and recognise the desired signals. After fusion with robot vision, embodied robot cognition will then derive HRI actions and knowledge on the entire scenario, and feed this back to the acoustic interface for further auditory scene analysis. This poster provides an overview of the EARS project goals, focusing in particular on the development of a signal processing system for speaker localisation, identification, and tracking as well as signal enhancement for speech recognition purposes.

POSTER BOARD 19

Resolution Limits on Visual Speech Recognition

Helen L. Bear, University of East Anglia, Norwich
Richard Harvey, University of East Anglia, Norwich
Yuxuan Lan, University of East Anglia, Norwich
Barry Theobald, University of East Anglia, Norwich

Visual-only speech recognition is dependent upon a number of factors that can be difficult to control, such as: lighting; identity; motion; emotion and expression. But some factors, such as video resolution are controllable, so it is surprising that there is not yet a systematic study of the effect of resolution on lip-reading. Here we use the Rosetta Raven data (a new data set), to train and test recognizers so we can measure the affect of video resolution on recognition accuracy.

Poster session 3: Tuesday 13:30–14:30

POSTER BOARD 1

Investigating the Effects of Knowledge Transfer in Multi-Domain Speech Recognition Systems

Mortaza Doulaty, Speech and Hearing Group, University of Sheffield
Thomas Hain, Speech and Hearing Group, University of Sheffield

This poster investigates the effects of knowledge transfer in multi-domain and cross-domain speech recognition systems. The common belief is that adding more data always helps. In this study, data from six different ASR domains is used, exhibiting negative transfer effects. An unsupervised method for identifying the parts of data causing negative transfer is proposed and its effectiveness in different cross-domain and multi-domain scenarios is studied. It is shown that a data selection technique based on the proposed method improves the performance of the recognition system. This study further shows that certain accepted domains in speech recognition do not appear to be as well defined in terms of results.

POSTER BOARD 2

Speech Technologies for Children

Eva Fringi, University of Birmingham
Martin Russel, University of Birmingham

Automatic speech recognition (ASR) is a very promising technological advancement which can be utilized in various applications to assist children's learning and entertainment. However, despite the fact that ASR systems can reach great levels of accuracy on adult speech, when it comes to children speech they perform significantly worse. The majority of research on ASR for children has been conducted using systems trained on adults' speech and focusing on the acoustic differences between adult and children speech, aiming to produce new methods which would modify adults' ASR systems to yield the same results as if they were trained on children speech. As a consequence, several techniques have been introduced to normalize the acoustic variability, which is prominent in children speech, namely pitch normalization (PN), vocal tract length normalization (VTLN) and speech rate normalization (SRN). At the same time studies on children speech trained recognisers indicate that the use of the appropriate training data improves the systems' performance, but not to such an extent as to elicit results comparable to those of systems trained and tested on adult speech. The hypothesis of the current study holds that apart from the discrepancies in acoustic components, it is also the constant phonological development that children speech is undergoing, which affects the performance of ASR on children. Studies on language acquisition have shown that a full set of phonemes is not acquired until the age of seven and even after that verbal instability remains in some cases until adolescence. The aim of the present research is to investigate the possible correlation between poor ASR performance and stages of speech development in children, in order to improve the former. The project is a collaboration with Disney Research Lab in Pittsburgh, USA.

POSTER BOARD 3

Measuring the perceptual effects of modelling assumptions in speech synthesis using stimuli constructed from repeated natural speech

Gustav Eje Henter, University of Edinburgh
Thomas Merritt, University of Edinburgh
Matt Shannon, University of Cambridge
Catherine Mayo, University of Edinburgh
Simon King, University of Edinburgh

Acoustic models used for statistical parametric speech synthesis typically incorporate many modelling assumptions. It is an open question to what extent these assumptions limit the naturalness of synthesised speech. To investigate this question, we recorded a speech corpus where each prompt was read aloud multiple times. By combining speech parameter trajectories extracted from different repetitions, we were able to quantify the perceptual effects of certain commonly used modelling assumptions. Subjective listening tests show that taking the source and filter parameters to be conditionally independent, or using diagonal covariance matrices, significantly limits the naturalness that can be achieved. Our experimental results also demonstrate the shortcomings of mean-based parameter generation.

POSTER BOARD 4

Diarisation of multi-channel TV studio recordings

Rosanna Milner, University of Sheffield
Yanxiong Li, South China University of Technology
Thomas Hain, University of Sheffield

Diarisation addresses the question "who speaks when" in audio recordings. This is typically performed by automatic segmentation producing speaker-pure fragments, followed by clustering. In this study diarisation is carried out on studio recordings of a TV series, where both individual and mixed channels are available. Although microphones are assigned to speakers, these are placed within reach of all speakers and hence a large amount of crosstalk is present. Firstly, state of the art diarisation systems are evaluated and used as baselines. Next, deep neural networks are used for speech/nonspeech detection with the aim of determining which speech segment belongs to which channel. This is contrasted with alignment of approximate transcripts, and subsequent energy based methods of speaker diarisation.

POSTER BOARD 5

Dialogue Context Sensitive Speech Synthesis using Context Adaptive Training

with Factorized Decision Trees

Pirros Tsiakoulis, University of Cambridge
Catherine Breslin, University of Cambridge
Milica Gasic, University of Cambridge
Matthew Henderson, University of Cambridge
Dongho Kim, University of Cambridge
Steve Young, University of Cambridge

Our recent work has shown significant improvements to the appropriateness for spoken dialogue system of HMM-based synthetic voices by introducing dialogue context into the decision tree state clustering stage. Continuing in this direction, we investigate the performance of dialogue context-sensitive voices in different domains. The Context Adaptive Training with Factorized Decision trees (FD-CAT) approach was used to train a dialogue context-sensitive synthetic voice which was then compared to a baseline system using the standard decision tree approach. Preference-based listening tests were conducted for two different domains. The first domain concerned restaurant information and had significant coverage in the training data, while the second dealing with appointment bookings had minimal coverage in the training data. No significant preference was found for any of the voices when tested in the restaurant domain whereas in the appointment booking domain, listeners showed a statistically significant preference for the adaptively trained voice.

POSTER BOARD 6

Introducing the TCD-TIMIT database as a resource for AVSR research

Eoin Gillen, TCD

Automatic audio-visual speech recognition currently lags behind its audio-only counterpart in terms of research progress. One of the reasons commonly cited by researchers is the scarcity of suitable research corpora. This issue motivated the creation of TCD-TIMIT, a new corpus designed for continuous audio-visual speech recognition research. TCD-TIMIT consists of high-quality audio and video footage of 62 speakers reading a total of 6913 sentences. Each sentence was recorded from two camera angles; straight and 30 degrees to the speaker's right. Three of the speakers are professional lipspeakers. In addition to the database itself, results from visual speech recognition tests done on the database using DCT, PCA and Optical Flow features are available. It is hoped that TCD-TIMIT will now be used to further the state of audio-visual speech recognition research.

POSTER BOARD 7

Infinite Structured Support Vector Machines for Speech Recognition

Jingzhou Yang, University of Cambridge
Rogier van Dalen, University of Cambridge
Shi-Xiong Zhang, University of Cambridge

Mark Gales, University of Cambridge
Discriminative models, like support vector machines (SVMs), have been successfully applied to speech recognition and improved performance. A Bayesian non-parametric version of the SVM, the infinite SVM, improves on the SVM by allowing more flexible decision boundaries. However, like SVMs, infinite SVMs model each class separately, which restricts them to classifying one word at a time. A generalisation of the SVM is the structured SVM, whose classes can be sequences of words that share parameters. This paper studies a combination of Bayesian non-parametrics and structured models. One specific instance called infinite structured SVM is discussed in detail, which brings the advantages of the infinite SVM to continuous speech recognition.

POSTER BOARD 8

Speaker individuality in head motion

Kathrin Haag, University of Edinburgh
Hiroshi Shimodaira, University of Edinburgh
Personality is an important factor in audio-visual human-computer interaction. While state-of-the-art synthetic voices for virtual assistants and avatars have become reasonably natural and intelligible, their body movement is often randomized and generated from speaker independent key poses. Our goal is to create a speech driven talking head with a personality, who learns individual head motion from speech, in order to make the user experience more natural and realistic. The question is in how far head motion differs between speakers and if speakers can be distinguished exclusively on the basis of their head movement. Does head movement between speakers differ considerably enough to justify an embedding of this individuality into talking heads? A preliminary data analysis suggested that this is indeed the case. In order to determine the degree of speaker individuality, we applied GMM based speaker recognition based only on head motion and achieved accuracy scores of up to 71.44%. We went further and recognized speaker dependent head motion trajectories with the help of aligned cluster analysis proposed by Zhou et al. (2008). Training a HMM based speaker recognition system on these clusters gave us promising results, and our findings lead us to conclude that speaker individuality should be integrated into speech driven talking heads.

POSTER BOARD 9

Adaptive Speech Recognition and Dialogue Management for Users with Speech Disorders

Iñigo Casanueva, University of Sheffield
Heidi Christensen, University of Sheffield
Thomas Hain, University of Sheffield
Phil Green, University of Sheffield
Spoken control interfaces are very attractive to people with severe physical disabilities who often also have a type of speech disorder known as dysarthria. This condition

is known to decrease the accuracy of automatic speech recognisers (ASRs) especially for users with moderate to severe dysathria. In this paper we investigate how applying probabilistic dialogue management (DM) techniques can improve interaction performance of an environmental control system for such users. The effect of having access to different amounts of adaptation data, as well as using different vocabulary size for speakers of different intelligibilities is investigated. We explore the effect of adapting the DM models as the ASR performance increases, such as is the case in systems where more adaptation data is collected through system use. Improvements compared to a non-probabilistic DM baseline are seen both in terms of dialogue length and success rate, 9% and 25% mean relative improvement respectively. Looking at just the more severe dysarthric speakers these numbers rise 25% and 75% mean relative improvement. These improvements are higher when the ASR data adaptation amount is small. Further results show that a DM trained on data from multiple speakers outperform a DM trained on data from a single speaker.

POSTER BOARD 10

Neural net word representations for phrase-break prediction without a part of speech tagger

Oliver Watts, University of Edinburgh
Siva Reddy Gangireddy, University of Edinburgh
Junichi Yamagishi, University of Edinburgh
Simon King, University of Edinburgh
Steve Renals, University of Edinburgh
Adriana Stan, Technical University of Cluj-Napoca
Mircea Giurgiu, Technical University of Cluj-Napoca

The use of shared projection neural nets of the sort used in language modelling is proposed as a way of sharing parameters between multiple text-to-speech system components. We experiment with pretraining the weights of such a shared projection on an auxiliary language modelling task and then apply the resulting word representations to the task of phrase-break prediction. Doing so allows us to build phrase-break predictors that rival conventional systems without any reliance on conventional knowledge-based resources such as part of speech taggers.

POSTER BOARD 11

CogWatch: Technologies for Stroke Patient Rehabilitation - an unfamiliar application of familiar techniques

Roozbeh Nabiei, University of Birmingham
Emilie Jean-Baptiste, University of Birmingham
Martin Russel, University of Birmingham

CogWatch is an EU project developing technologies to help stroke patients complete a range of activities of daily living (ADL) independently. A third of these patients have long term physiological or cognitive disabilities, and many 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 patient's progress through an ADL, and return a cue if an error occurs or is imminent. The initial ADL is tea making, but others will be addressed. Much of the inspiration for our approach to CogWatch comes from Spoken Dialogue Systems.

POSTER BOARD 12

Multi-pass approach for sentence end detection in lecture speech

Madina Hasan, The University of Sheffield
Rama Doddipatla, The University of Sheffield
Thomas Hain, The University of Sheffield

Making speech recognition output readable is an important task. The first step here is automatic sentence end detection (SED). We introduce novel F0 derivative-based features and sentence end distance features for SED that yield significant improvements in slot error rate (SER) in a multi-pass framework. Three different SED approaches are compared on a spoken lecture task: hidden event language models, boosting, and Conditional Random Fields (CRFs). Experiments on reference transcripts show that CRF-based models give best results. Addition of pause duration features yields an improvement of 11.1% in SER. The addition of the F0-derivative features yield a further reduction of 3.0% absolute, and an additional 0.5% reduction is gained by backward distance features. In the absence of audio, the use of backward features alone give 2.2% absolute reduction in SER.

POSTER BOARD 13

Standalone Training of Context-Dependent Deep Neural Network Acoustic Models

Chao Zhang, Cambridge University Engineering Dept, Cambridge, U.K.
Philip Charles Woodland, Cambridge University Engineering Dept, Cambridge, U.K.

Recently, context-dependent (CD) deep neural network (DNN) hidden Markov models (HMMs) have been widely used as acoustic models for speech recognition. However, the standard method to build such models requires target training labels from a system using HMMs with Gaussian mixture model output distributions (GMM-HMMs). In this paper, we introduce a method for training state-of-the-art CD-DNN-HMMs without relying on such a pre-existing system. We achieve this in two steps: build a context-independent (CI) DNN iteratively with word transcriptions, and then cluster the equivalent output distributions of the untied CD-DNN HMM states using the decision tree based state tying approach. Experiments have been performed on the Wall Street Journal corpus and the resulting system gave comparable word error rates (WER) to CD-DNNs built based on GMM-HMM alignments and state-clustering.

POSTER BOARD 14

Towards a Spoken Dialogue API

Martin Szummer, VocalIQ
Blaise Thomson, VocalIQ

There exist a few established standards for specifying spoken dialog systems: VoiceXML and SCXML. These specifications were developed before the arrival of machine-learning based approaches to natural language understanding, dialogue state tracking and decision making. In the light such data-driven techniques, we discuss an API that enables dialog systems to be specified in a declarative rather than a procedural way. Instead of writing specific grammars and rules for how to understand and carry on the dialogue in given situations, we leave these to be learned automatically. We demonstrate a simple database-driven system built using the API.

POSTER BOARD 15

Efficient GPU-based Training of Recurrent Neural Network Language Models Using Spliced Sentence Bunch

X. Chen, Cambridge University, Engineering Department
Y. Wang, Cambridge University, Engineering Department
X. Liu, Cambridge University, Engineering Department
M.J.F. Gales, Cambridge University, Engineering Department
P. C. Woodland, Cambridge University, Engineering Department

Recurrent neural network language models (RNNLMs) are becoming increasingly popular for a range of applications including speech recognition. However, an important issue that limits the quantity of data, and hence their possible application areas, is the computational cost in training. A standard approach to handle this problem is to use class-based outputs, allowing systems to be trained on CPUs. This paper describes an alternative approach that allows RNNLMs to be efficiently trained on GPUs. This enables larger quantities of data to be used, and networks with an unclustered, full output layer to be trained. To improve efficiency on GPUs, multiple sentences are "spliced" together for each mini-batch or "bunch" in training. On a large vocabulary conversational telephone speech recognition task, the training time was reduced by a factor of 27 over the standard CPU-based RNNLM toolkit. The use of an unclustered, full output layer also improves perplexity and recognition performance over class-based RNNLMs.

POSTER BOARD 16

Objective Voice Quality Assessment using Digital Signal Processing and Machine Learning

Farideh Jalalinajafabadi, School of computer science, University of Manchester
Chaitanya Gadepalli, University department of Otolaryngology, Head and Neck Surgery, Central Manchester Foundation trust

Frances Ascott, University department of Otolaryngology, Head and Neck Surgery, Central Manchester Foundation trust
Jarrod Homer, University department of Otolaryngology, Head and Neck Surgery, Central Manchester Foundation trust
Mikel Luján, School of computer science, University of Manchester
Barry Cheetham, School of computer science, University of Manchester

Voice disorders may be caused by voice-strain due to speaking or singing, vocal cord damage, infection, side effects of inhaled steroids as used to treat asthma or more serious disease including laryngeal cancer and neurological disease. The resulting loss of voice quality can be measured subjectively or objectively. For clinical and research use the Japanese Society of Logopeadics and Phoniatrics and the European Research Group recommended a standard referred to as 'GRBAS' which is an acronym for a five dimensional scale of measurements of voice properties. The properties are 'grade', 'roughness', 'breathiness', 'asthenia' and 'strain'. Each property is quantified by one dimension of the scale, and it is standard to use a range between 0 and 3; 0 for normal, 1 for mild impairment, 2 for moderate impairment and 3 for severe impairment. The GRBAS scale has the advantage of being widely understood and recommended by many professional bodies, but its subjectivity and reliance on highly trained personnel are significant limitations. The aim of this research is to design and evaluate objective measurement of voice quality conforming to the GRBAS standard. Overall, a recorded voice signal will be fed into a digital system consisting digital signal processing and mapping techniques based on machine learning. Different voice features such as voice power, low-to high spectral energy, tremor and Harmonic-to noise ratio will be extracted from the voice and used as features by the mapping techniques.

POSTER BOARD 17

Intelligibility of fast synthesized speech

Cassia Valentini-Botinhao, University of Edinburgh
Markus Toman, Telecommunications Research Center Vienna (FTW), Austria
Michael Pucher, Telecommunications Research Center Vienna (FTW), Austria
Dietmar Schabus, Telecommunications Research Center Vienna (FTW), Austria
Junichi Yamagishi, University of Edinburgh

We analyse the effect of speech corpus and compression method on the intelligibility of synthesized speech at fast rates. We recorded English and German language voice talents at a normal and a fast speaking rate and trained an HSMM-based synthesis system based on the normal and the fast data of each speaker. We compared three compression methods: scaling the variance of the state duration model, interpolating the duration models of the fast and the normal voices, and applying a linear compression method to generated speech. Word recognition results for the English voices show that generating speech at normal speaking rate and then applying linear compression resulted in the most intelligible speech at all tested rates. For the German voices, interpolation was found to be better at moderate speaking rates but the linear method was again more successful at very high rates.

POSTER BOARD 18

Multiple-Average-Voice-based Speech Synthesis

Pierre Lanchantin, Cambridge University Engineering Department
Mark Gales, Cambridge University Engineering Department
Simon King, CSTR, Edinburgh
Junichi Yamagishi, CSTR, Edinburgh

This paper describes a novel approach for the speaker adaptation of statistical parametric speech synthesis systems based on the interpolation of a set of average voice models (AVM). Recent results have shown that the quality/naturalness of adapted voices depends on the distance from the average voice model used for speaker adaptation. This suggests the use of several AVMs trained on carefully chosen speaker clusters from which a more suitable AVM can be selected/interpolated during the adaptation. In the proposed approach a set of AVMs, a multiple-AVM, is trained on distinct clusters of speakers which are iteratively re-assigned during the estimation process initialised according to metadata. During adaptation, each AVM from the multiple-AVM is first adapted towards the target speaker. The adapted means from the AVMs are then interpolated to yield the final speaker adapted mean for synthesis. It is shown, performing speaker adaptation on a corpus of British speakers with various regional accents, that the quality/naturalness of synthetic speech of adapted voices is significantly higher than when considering a single factor-independent AVM selected according to the target speaker characteristics.

POSTER BOARD 19

The UEDIN English ASR System for the IWSLT 2013 Evaluation

Peter Bell, University of Edinburgh
Fergus McInnes, University of Edinburgh
Siva Reddy Gangireddy, University of Edinburgh
Mark Sinclair, University of Edinburgh
Alexandra Birch, University of Edinburgh
Steve Renals, University of Edinburgh

This paper describes the University of Edinburgh (UEDIN) English ASR system for the IWSLT 2013 Evaluation. Notable features of the system include deep neural network acoustic models in both tandem and hybrid configuration, cross-domain adaptation with multi-level adaptive networks, and the use of a recurrent neural network language model. Improvements to our system since the 2012 evaluation – which include the use of a significantly improved n-gram language model – result in a 19% relative WER reduction on the tst2012 set.

Notes

Organizing committee	Finance & Website	Local arrangements
Cassia Valentini Botinhao Naomi Harte Peter Jančovič Rogier van Dalen	Mark Huckvale	Cassia Valentini Botinhao Rasmus Dall