UK Speech Conference

Cambridge

17–18 September 2013

Schedule

Tuesday

10:30–11:30 LT1 Badge pick-up and tea for arriving delegates

11:30–11:35 LT1 Welcome

11:35–12:20 LT1 What's Happening in Speech Enhancement and Acoustic Signal Processing?

12:30–13:30
ST CATHARINE'S COLLEGE Lunch

13:45–15:15 LT1
Tutorial: (Deep) Neural Networks for Speech Recognition — Steve Renals

15:15–15:45 LR3 Tea

15:45–17:00 LR3 Posters

18:00–19:30 HILTON DOUBLETREE
Drinks reception

Wednesday

9:00–10:30 LT1
Tutorial: An Introduction to Spoken Dialog Systems — Blaise Thomson

10:30–11:00 LR3 Tea

11:00–12:15 LR3 Posters

12:30–13:30
ST CATHARINE'S COLLEGE Lunch

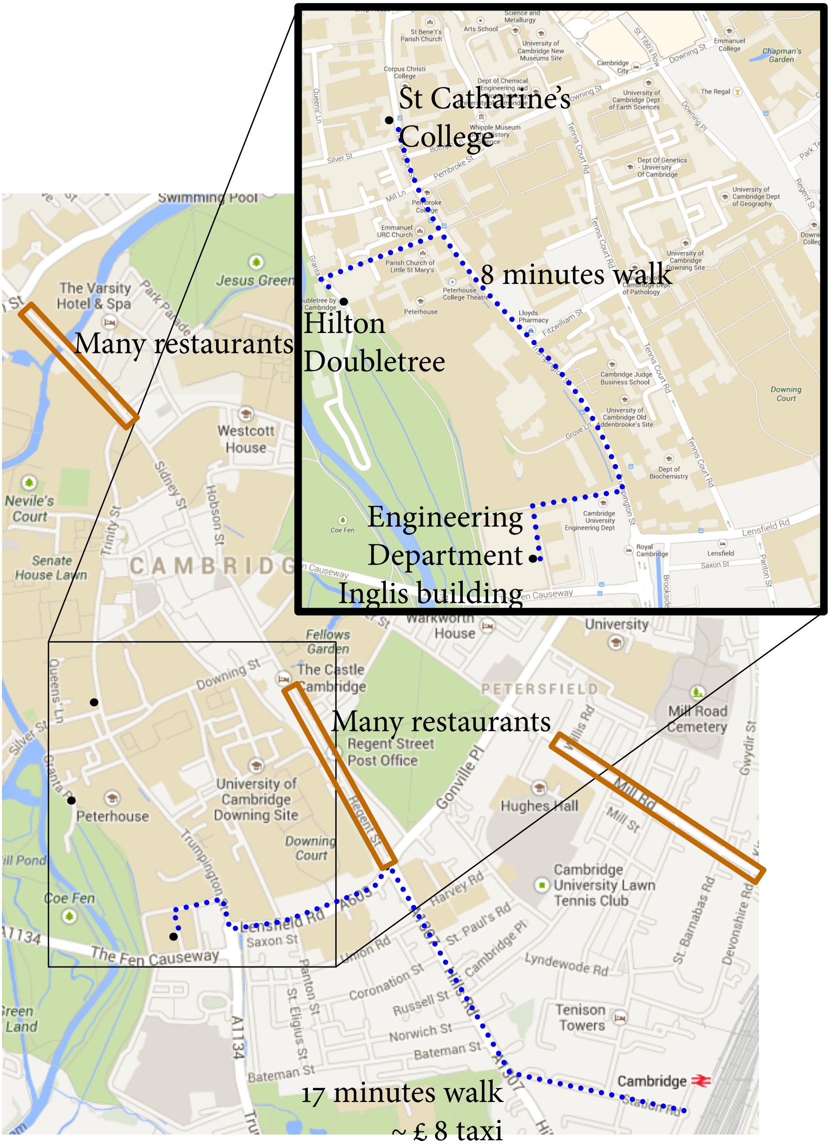
13:45–15:00 LR3 Posters

15:00–15:15 LR3 Tea

15:15–16:00 LT1 What's Happening in Accents & Dialects?

16:00–16:30 LT1
Closing comments/discussion

Map


St Catharine's College
8 minutes walk
Hilton Doubletree
Engineering Department
Inglis building
Many restaurants
Many restaurants
17 minutes walk
~ £ 8 taxi
The Varsity Hotel & Spa
Jesus Green
Swimming Pool
Westcott House
Nevile's Court
Senate House Lawn
CAMBRIDGE
Fellows Garden
The Castle Cambridge
Warkworth House
University
PETERSFIELD
Mill Road Cemetery
Regent Street Post Office
Hughes Hall
University of Cambridge Downing Site
Downing Court
Peterhouse
Coe Fen
The Fen Causeway
Lensfield Rd
Cambridge University Lawn Tennis Club
Tenison Towers
Cambridge

Tutorials

(Deep) Neural Networks for Speech Recognition
Steve Renals, University of Edinburgh
Neural networks have become a very hot topic in speech technology, with recent work on neural network acoustic and language modelling extending the state of the art, and attracting an extraordinary amount of interest. This tutorial will give an overview of current work in the area, making links with work done since the late 1980s, while showing what is new. I'll finish by talking about some current challenges that might drive work in neural networks for speech recognition.

An Introduction to Spoken Dialog Systems
Blaise Thomson, University of Cambridge
Science fiction has long assumed we will be able to build machines that we can interact with via speech, usually called spoken dialog systems. While most attempts at such machines have had limited success, recent advances have led to some programs like Apple's Siri and Google Now receiving a largely positive reception.

This tutorial will give an introduction to the key components required to build such a computer, as well as presenting some recent research trends in the area. We will discuss various statistical methods used for analysing meaning (semantics), tracking the state of the dialog, as well as making decisions about what to say.

Posters

Poster session 1: Tuesday 15:45–17:00

POSTER BOARD 1

The Sheffield Wargame Corpus (SWC)

Charles Fox, University of Sheffield
Yulan Liu, University of Sheffield
Erich Zwyssig, University of Edinburgh
Thomas Hain, University of Sheffield

Recognition of speech in natural environments is a challenging task, even more so if this involves conversations between several speakers. Work on meeting recognition has addressed some of the significant challenges, mostly targeting formal, business style meetings where people are mostly in a static position in a room. Only limited data is available that contains high quality near and far field data from real interactions between participants. In this poster we present a new corpus for research on speech recognition, speaker tracking and diarisation, based on recordings of native speakers of English playing a table-top wargame. The Sheffield Wargames Corpus (SWC) comprises 7 hours of data from 10 recording sessions, obtained from 96 microphones, 3 video cameras and, most importantly, 3D location data provided by a sensor tracking system. The corpus represents a unique resource, that provides for the first time location tracks of speakers that are constantly moving and talking. The corpus is available for research purposes, and includes annotated development and evaluation test sets. Baseline results for close-talking and far field sets are included in this paper.

POSTER BOARD 2

Anthropomorphism and lexical alignment in human-computer dialogue

Benjamin R. Cowan, HCI Centre, University of Birmingham
Holly Branigan, Department of Psychology, University of Edinburgh
Russell Beale, HCI Centre, University of Birmingham

Our interlocutors affect our linguistic behaviours in discourse. A common observation is that people tend to converge, or align, linguistically in dialogue. This alignment is a key component of natural and successful communication. Recent research suggests that alignment at the lexical level can be influenced by our judgments of the abilities of our interlocutors as effective communication partners. Although the use of speech as an interaction modality in mainstream computing is rising, little is known about the influence interlocutor design may have on this alignment behaviour. The research presented uses a Wizard of Oz based referential communication task to explore how interlocutor design in human-computer dialogue in the form of voice anthropomorphism impacts lexical alignment. The results show a strong lexical alignment effect in

spoken human-computer dialogue, yet no significantly impact of interlocutor ability judgment on alignment levels. This gives support to the incorporation of lexical alignment in spoken dialogue system user models as well as suggesting that lexical alignment in human-computer dialogue may be influenced by priming rather than considered interlocutor modeling.

POSTER BOARD 3

Investigating the shortcomings of HMM synthesis

Thomas Merritt, University of Edinburgh
Simon King, University of Edinburgh

Despite years of improvement in the quality of HMM (Hidden Markov Model) synthesis, this type of synthetic speech still remains significantly less natural than speech output from good concatenative synthesis systems. This is commonly stated as being due to "over-smoothing" however to the best of our knowledge there has been no formal studies to support this. We will present a framework for separating each of the effects of modelling in turn to observe their independent effects.

POSTER BOARD 4

Investigation of multilingual deep neural networks for spoken term detection

Kate Knill, University of Cambridge
Mark Gales, University of Cambridge
Shakti Rath, University of Cambridge
Phil Woodland, University of Cambridge
Chao Zhang, University of Cambridge
Shi-Xiong Zhang, University of Cambridge

The development of high-performance speech processing systems for low-resource languages is a challenging area. 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 use bottleneck features, or hybrid systems, trained on multilingual data for speech-to-text (STT) systems. This poster presents an investigation into the application of these multilingual approaches to spoken term detection (STD). Experiments were run using the IARPA Babel limited language pack corpora (approx. 10 hours/language) with 4 languages for initial multilingual system development and an additional held-out target language. STT gains achieved through using multilingual bottleneck (BN) features in a Tandem configuration are shown to also apply to keyword search (KWS). Further improvements in both STT and KWS were observed by incorporating language questions into the Tandem GMM-HMM decision trees for the training set languages. Adapted hybrid systems performed slightly worse on average than the adapted Tandem systems. A language independent acoustic model test on the target language showed that retraining or adapting of the acoustic models to the target language is currently minimally needed to achieve reasonable performance.

POSTER BOARD 5

Infinite Support Vector Machines in Speech Recognition

Jingzhou Yang, University of Cambridge
Rogier van Dalen, University of Cambridge
Mark Gales, University of Cambridge

Generative feature spaces provide an elegant way to apply discriminative models in speech recognition, and system performance has been improved by adapting this framework. However, the classes in the feature space may be not linearly separable. Applying a linear classifier then limits performance. Instead of a single classifier, this paper applies a mixture of experts. This model trains different classifiers as experts focusing on different regions of the feature space. However, the number of experts is not known in advance. This problem can be bypassed by employing a Bayesian non-parametric model. In this paper, a specific mixture of experts based on the Dirichlet process, namely the infinite support vector machine, is studied. Experiments conducted on the noise-corrupted continuous digit task AURORA 2 show the advantages of this Bayesian non-parametric approach.

POSTER BOARD 6

Native Accent Classification via I-Vectors and Speaker Compensation Fusion

Andrea DeMarco, University of East Anglia
Stephen Cox, University of East Anglia

Accent classification is an interesting technology for speaker recognition. One approach to accent classification is to perform speech recognition and to then search for phonetic contrasts that determine accents - this relies on very accurate recognition. Here, we focus on classification of the accents of talkers without any annotation of their speech other than the accent label. Utterances are first represented as I-Vectors, to which discriminative transformations are applied to separate accent groups, ignoring ”noise” from variation in speakers and channels. We test a number of parameters for I-vector classifier systems:

1. Length-normalization
2. Universal Background Model (UBM) sizes
3. I-vector Factor dimensionality
4. Channel compensation methods (LDA, R-LDA, SDA, NCA)

POSTER BOARD 7

Improving Lightly Supervised Training for Broadcast Transcription

Y. Long, CUED
M.J.F. Gales, CUED
P. Lanchantin, CUED
X. Liu, CUED
M.S. Seigel, CUED
P.C. Woodland, CUED

This paper investigates improving lightly supervised acoustic model training for an archive of broadcast data. Standard lightly supervised training uses automatically derived decoding hypotheses using a biased language model. However, as the actual speech can deviate significantly from the original programme scripts that are supplied, the quality of standard lightly supervised hypotheses can be poor. To address this issue, word and segment level combination approaches are used between the lightly supervised transcripts and the original programme scripts which yield improved transcriptions. Experimental results show that systems trained using these improved transcriptions consistently outperform those trained using only the original lightly supervised decoding hypotheses. This is shown to be the case for both the maximum likelihood and minimum phone error trained systems.

POSTER BOARD 8

Learning to imitate adult speech with the KLAIR virtual infant

Mark Huckvale, Speech, Hearing and Phonetic Sciences, University College London
Amrita Sharma, Speech, Hearing and Phonetic Sciences, University College London

Pre-linguistic infants need to learn how to produce spoken word forms that have the appropriate intentional effect on adult carers. One proposed imitation strategy is based on the idea that infants are innately able to match the sounds of their own babble to sounds of adults, while another proposed strategy requires only reinforcement signals from adults to improve random imitations. Here we demonstrate that knowledge gained from interactions between infants and adults can provide useful normalizing data that improves the recognisability of infant imitations. We use the KLAIR virtual infant toolkit to collect spoken interactions with adults, exploit the collected data to learn adult-to-infant mappings, and construct imitations of adult utterances using KLAIR's articulatory synthesizer. We show that speakers reinterpret and reformulate KLAIR's productions in terms of standard phonological forms, and that these reformulations can be used to train a system that generates infant imitations that are more recognisable to adults than a system based on babbling alone.

POSTER BOARD 9

Hybrid Acoustic Models for Distant and Multichannel Large Vocabulary Speech Recognition

Pawel Swietojanski, University of Edinburgh
Arnab Ghoshal, University of Edinburgh
Steve Renals, University of Edinburgh

We investigate the application of deep neural network-hidden Markov model hybrid acoustic models for far-field speech recognition of meetings recorded using microphone arrays. We show that on a large vocabulary distant speech recognition task the hybrid models achieve significantly better accuracy than conventional systems based on Gaussian mixture models (GMMs). We observe up to 8% absolute word error rate (WER) reduction from a discriminatively trained GMM baseline when using a single distant microphone, and between 6.4% to 3.7% absolute WER reduction when using beamforming on various combinations of array channels. By training the networks on audio from multiple channels, we find the networks can recover significant part of accuracy difference between the single distant microphone and beamformed configurations. Finally, we show that it is possible to improve the accuracy of a network recognising speech from a single distant microphone to the level of a multi-microphone setup by suitably constraining the learning using data from other microphones.

POSTER BOARD 10

Cross-domain Paraphrasing For Improving Language Modelling Using Out-of-domain Data

Xunying Liu, Cambridge University
Mark Gales, Cambridge University
Phil Woodland, Cambridge University

In natural languages the variability in the underlying linguistic generation rules significantly alters the observed surface word sequence they create, and thus introduces a mismatch against other data generated via alternative realizations associated with, for example, a different domain. Hence, direct modelling of out-of-domain data can result in poor generalization to the in- domain data of interest. To handle this problem, this paper investigated using cross-domain paraphrastic language models to improve in-domain language modelling (LM) using out-of- domain data. Phrase level paraphrase models learned from each domain were used to generate paraphrase variants for the data of other domains. These were used to both improve the context coverage of in-domain data, and reduce the domain mismatch of the out-of-domain data. Significant error rate reductions of 0.6%-0.8% absolute were obtained on two state-of-the-art LVCSR tasks using a cross-domain paraphrastic multi-level LM trained on a billion words of domain mixed data. Consistent improvements on the in-domain data context coverage were also obtained.

POSTER BOARD 11

IDLAK - Parametric Text to Speech for Kaldi

Mathew Aylett, University of Edinburgh and CereProc Ltd.

Kaldi, the open source ASR system, has been influential in developing new statistical modelling techniques. In order to harness these new developments, as well as in order to offer an alternative to th HTS system based on HTK which has a more liberal licensing environment, a parametric speech synthesis system is being developed within the Kaldi framework, called Idlak. In this poster we will give an overview of the design of this new TTS system, current progress, an examples of the output of different Idlak modules.

POSTER BOARD 12

Using statistical language models and edit distance metrics for prediction and error correction in a novel interface for mathematical text

Dilaksha Attanayake, School of Computing and Information Systems Kingston University
Gordon Hunter, School of Mathematics Kingston University
Eckhard Pfluegel, School of Computing and Information Systems Kingston University
James Denhold-Price, School of Mathematics Kingston University

Editing mathematical text is a tedious and rather error-prone process. It is even difficult for people with disabilities such as visual impairments and/or limited (or no) use of their hands. In this poster, we discuss the development, implementation and initial evaluation of a system designed to address these issues by allowing the creation and editing of mathematical text via spoken and/or typed natural language commands. In order to have this system easier to use, we have incorporated predictive text (using statistical language models) and error correction (using edit-distance metric) features whilst still allowing the user to have the final choice. We also present an initial evaluation of the system using example mathematical formulae, translated into natural language, containing controlled artificially introduced errors.

POSTER BOARD 13

Room Geometry Estimation from a Single Channel Acoustic Impulse Response

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

For a 2D rectangular room of unknown dimensions and with unknown source and microphone positions, the times of arrival of reflections can be described in terms of image source positions. Adopting a microphone-centred co-ordinates system, it is shown that to satisfy certain combinations of arrival times imposes constraints on the possible room geometry: a second-order reflection from adjacent walls determines the source-microphone distance; a second-order reflection from opposite walls in a given

dimension determines the source displacement in that dimension as a function of the source-receiver distance. Given a subset of time differences of arrival, the extent to which the geometry can be determined is related to these constraints. The geometry estimation is further posed as a least squares optimisation problem whose results verify the analytical results.

POSTER BOARD 14

Modelling reverberation compensation effects in time-forward and time-reversed rooms

Amy V Beeston, University of Sheffield
Guy J Brown, University of Sheffield

Human listeners can perceptually compensate for the effects of reverberation in rooms. Recent work suggests that listeners can achieve constancy for some properties of the signal envelope [Kuwada et al. (2012) Front. Neural Circuits 6, 42], which may help to explain the increase in likelihood that a reverberant test word will be correctly identified when it is presented following a similarly reverberated speech carrier [Brandewie and Zahorik (2010) J. Acoust. Soc. Am. 128 (1), 291-299]. These compensation effects are not observed when the reverberation characteristics of the room are time-reversed [Watkins (2005) J. Acoust. Soc. Am. 118 (1) 249-262; Longworth-Reed et al. (2008) J. Acoust. Soc. Am. 125 (1) EL13-EL19]. Rather than presenting the listener with decay tails following offsets, reflected energy precedes the direct sound in time-reversed reverberation, causing ramps prior to onsets. Objective measures of reverberation typically quantify the temporal modulation reduction imposed by a room (e.g., speech transmission index, modulation transfer function). In their standard implementation, however, such measures do not depend on the time-direction of the reverberation and thus cannot explain the pattern of results observed in human listener data. A computational auditory model of compensation for reverberation is presented. The model is based on an efferent feedback loop which monitors and controls the dynamic range of the simulated auditory nerve response resulting from peripheral processing in the afferent pathway. Using a reverberation-detection metric that examines the energy present during tails in the simulated auditory nerve signal, the model displays a qualitative match to human results on a categorical perception task in time-forward and time-reversed rooms.

Poster session 2: Wednesday 11:00–12:15

POSTER BOARD 1

An experimental comparison of multiple vocoder types

Qiong Hu, University of Edinburgh, U.K.
Korin Richmond, University of Edinburgh, U.K.
Junichi Yamagishi, University of Edinburgh & National Institute of Informatics, Tokyo, Japan
Javier Latorre, Toshiba Research Europe Ltd, Cambridge, U.K.

This paper presents an experimental comparison of a broad range of the leading vocoder types which have been previously described. We use a reference implementation of each of these to create stimuli for a listening test using copy synthesis. The listening test is performed using both Lombard and normal read speech stimuli, and with two types of question for comparison. Multi-dimensional Scaling (MDS) is conducted on the listener responses to analyse similarities in terms of quality between the vocoders. Our MDS and clustering results show that the vocoders which use a sinusoidal synthesis approach are perceptually distinguishable from the source-filter vocoders. To help further interpret the axes of the resulting MDS space, we test for correlations with standard acoustic quality metrics and find one axis is strongly correlated with PESQ scores. We also find both speech style and the format of the listening test question may influence test results. Finally, we also present preference test results which compare each vocoder with the natural speech.

POSTER BOARD 2

Combining perceptually-motivated spectral shaping with loudness and duration modification for intelligibility enhancement of HMM-based synthetic speech in noise

Cassia Valentini-Botinhao, Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK
Junichi Yamagishi, Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK & National Institute of Informatics, Tokyo, Japan
Simon King, Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK
Yannis Stylianou, Institute of Computer Science, Foundation of Research and Technology Hellas, Crete, Greece

This paper presents our entry to a speech-in-noise intelligibility enhancement evaluation: the Hurricane Challenge. The system consists of a Text-To-Speech voice manipulated through a combination of enhancement strategies, each of which is known to be individually successful: a perceptually-motivated spectral shaper based on the Glimpse Proportion measure, dynamic range compression, and adaptation to Lom-

bard excitation and duration patterns. We achieved substantial intelligibility improvements relative to unmodified synthetic speech: 4.9 dB in competing speaker and 4.1 dB in speech-shaped noise. An analysis conducted across this and other two similar evaluations shows that the spectral shaper and the compressor (both of which are loudness boosters) contribute most under higher SNR conditions, particularly for speech-shaped noise. Duration and excitation Lombard-adapted changes are more beneficial in lower SNR conditions, and for competing speaker noise.

POSTER BOARD 3

docuMeet: meeting transcription and summarisation system

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

Meetings are an essential part of business life in large and small organisations. In a multimedia information age a desire to capture the important decisions and the associated processes in meetings is only natural. However, recording meeting minutes is expensive on human resource, and hence minutes are rarely generated or even integrated into the knowledge base of an organisation. DocuMeet is a research project funded by European Union. It aims to develop a user-friendly SW/HW platform that allows organisations to better manage their meetings, and to easily document, disseminate, search and implement the conclusions of each meeting. The scope of the project encompasses components in hardware and software. Hardware components include an appropriate and efficient audio recording platform as well as the design of an efficient control device. Automatic speech recognition and further downstream processes such as summarisation are implemented as cloud services. This poster aims to describe the hardware/software design of the DocuMeet meeting management system, with specific focus on components under development in the Speech and Hearing Group at the University of Sheffield. The ASR system includes personalisation and robustness aspects. The output of recognition is enhanced by punctuation and capitalisation systems and entered into a summarisation unit. The system will allow track agendas as well as speakers across meetings. All recognition processes will be performed off-line. We include a description of the system architecture and report on development results for various component technologies.

POSTER BOARD 4

Language modelling for TED talks recognition

Fergus McInnes, University of Edinburgh
Arnab Ghoshal, University of Edinburgh
Siva Reddy Gangireddy, University of Edinburgh
Qiang Huang, University of Edinburgh
Steve Renals, University of Edinburgh

This poster presents our recent work on language modelling for TED talks recognition, part of the University of Edinburgh's system for the IWSLT-2013 evaluation. The poster will include the use of cross-entropy filtering to select out-of-domain data to include in the language model, the use of recurrent neural network language models in combination with n-gram language models, and the incorporation of a Bayesian topic model.

POSTER BOARD 5

Speech Enhancement from Additive Noise and Channel Distortion — a Corpus-Based Approach

Ji Ming, Queen's University Belfast
Danny Crookes, Queen's University Belfast

In this paper, we study a new approach to recovering clean speech from additive noise and channel distortion, and demonstrate its application to robust speech recognition. The new approach consists of two parts: a clean, wideband speech corpus providing examples of the speech to recover, and a longest matching segment method for locating the matching corpus speech given noisy speech with additive noise and channel distortion. We assume no specific knowledge about the noise and channel characteristics, but only that they change slower than the speech. Lengthening the segments being compared reduces the uncertainty of the unknown, slowly-varying background noise and channel effect in forming the match, and hence the uncertainty of the matching corpus speech. We have evaluated the new approach on the Aurora 4 database containing speech with additive noise and combined additive noise and channel distortion. The reconstructed speech from the new approach was passed to a speech recogniser for recognition. The recogniser was trained using the WSJ0 training set which also served as the corpus for speech recovery. The new approach resulted in an absolute word error reduction from 44.4% to 19.0% for the speech with additive noise, and from 63.8% to 25.5% for the speech with both additive noise and channel distortion, a performance comparable to that of the state-the-art speech recognition systems based on model/feature adaptation.

POSTER BOARD 6

Asynchronous Factorisation of Speaker and Background with Feature Transforms in Speech Recognition

Oscar Saz, University of Sheffield
Thomas Hain, University of Sheffield

This poster presents a novel approach to separate the effects of speaker and background conditions by application of feature–transform based adaptation for Automatic Speech Recognition (ASR). So far factorisation has been shown to yield improvements in the case of utterance-synchronous environments. In this paper we show successful separation of conditions asynchronous with speech, such as background music. Our work takes account of the asynchronous nature of the background, by estimation of condition-specific Constrained Maximum Likelihood Linear Regression (CMLLR) transforms. In addition, speaker adaptation is performed, allowing to factorise speaker and background effects. Equally, background transforms are used asynchronously in the decoding process, using a modified Hidden Markov Model (HMM) topology which applies the optimal transform for each frame. Experimental results are presented on the WSJCAMo corpus of British English speech, modified to contain controlled sections of background music. This addition of music degrades the baseline Word Error Rate (WER) from 10.1% to 26.4%. While synchronous factorisation with CMLLR transforms provides 28% relative improvement in WER over the baseline, our asynchronous approach increases this reduction to 33%.

POSTER BOARD 7

Topic model features in neural network language models

Siva Reddy Gangireddy, University of Edinburgh
Qiang Huang, University of Edinburgh
Steve Renals, University of Edinburgh
Fergus McInnes, University of Edinburgh
Junichi Yamagishi, University of Edinburgh

In this paper we investigate the use of statistical topic models to provide longer-term context for neural network language models (NNLMs). Latent Dirichlet Allocation (LDA) is used to compute the context vector, by considering the long-span context of the current word. The computed LDA features and the previous n-1 words are used as inputs to an NNLM, optimizing the width of the long-span context with respect to perplexity. Experiments on the Penn Tree Bank Corpus subset of the Wall Street Journal text data, we observed 21% relative improvement with LDA features over a Kneser-Ney smoothed n-gram baseline, and 27% after interpolation.

POSTER BOARD 8

Identification of Gender from Children's Speech by Computers and Humans

Saeid Safavi, University of Birmingham
Peter Jancovic, University of Birmingham
Martin Russell, University of Birmingham
Michael Carey, University of Birmingham

This paper presents results on gender identification (GI) for children's speech, using the OGI Kids corpus and GMM-UBM and GMM-SVM systems. Regions of the spectrum containing important gender information for children are identified by conducting GI experiments over 21 frequency sub-bands. Results show that the frequencies below 1.8 kHz and above 3.8 kHz are most useful for GI for older children, while the frequencies above 1.4 kHz are most useful for the youngest children. The effect of using age-independent and age-dependent gender modelling (including the effects of puberty on boys voices) is explored. The application of intersession variability compensation is explored but experiments showed only little improvement. Experiments on human GI were also conducted and the results show that the humans do not achieve the performance of the machine.

POSTER BOARD 9

Lightly Supervised Automatic Subtitling of Weather Forecasts

Joris Driesen, University of Edinburgh
Steve Renals, University of Edinburgh

There are many examples of large archives of acoustic data that are only imperfectly transcribed: audio books, lectures, television broadcasts, etc. Since the training of acoustic models for automatic speech recognition requires verbatim orthographic transcriptions, none of this data can readily be used for this purpose. To unlock these resources, one has to resort to light supervision methods, which essentially use the imperfect transcriptions to determine a usable subset of the data. In this work, we discuss and compare two such methods, one based on forced alignments using finite state automata, the other based on free alignments using a biased language model. We compare both of the selected methods on a large set of weather reports, using their subtitles as approximate transcriptions. Finally, the best training set obtained is used to create a hybrid deep neural network-based recognition system, which is shown to perform extremely well on three separate test sets.

POSTER BOARD 10

Unsupervised and lightly-supervised learning for rapid construction of TTS systems in multiple languages from 'found' data: evaluation and analysis

Oliver Watts, University of Edinburgh, UK
Adriana Stan, Technical University of Cluj-Napoca, Romania
Rob Clark, University of Edinburgh, UK
Yoshitaka Mamiya, University of Edinburgh, UK
Mircea Giurgiu, Technical University of Cluj-Napoca, Romania
Junichi Yamagishi, University of Edinburgh, UK
Simon King, University of Edinburgh, UK

We present techniques for building text-to-speech front-ends in a way that avoids the need for language-specific expert knowledge, but instead relies on universal resources (such as the Unicode character database) and unsupervised learning from unannotated data to ease system development. The acquisition of expert language-specific knowledge and expert annotated data is a major bottleneck in the development of corpus-based TTS systems in new languages. The methods presented here side-step the need for such resources as pronunciation lexicons, phonetic feature sets, part of speech tagged data, etc. The paper explains how the techniques introduced are applied to the 14 languages of a corpus of 'found' audiobook data. Results of an evaluation of the intelligibility of the systems resulting from applying these novel techniques to this data are presented.

POSTER BOARD 11

Pairwise audio comparison for visualisation of mispronunciation.

Amy Beeston, University of Sheffield
Mauro Nicolao, University of Sheffield
Thomas Hain, University of Sheffield

Assessment of pronunciation quality in learners of foreign languages is still a challenging task, especially with adolescent students. As part of a research project on language learning for a public domain learning platform, the work presented in this poster explores methods for pronunciation quality assessment for Dutch learners of English. The project is supported by a large corpus of classroom recordings, and aims to develop training tools for language learners. Our initial work concentrates on comparison of teacher/student utterance pairs in a phone-by-phone manner, considering differences in their energy contour and spectral shape. To test our baseline system, pairwise comparisons were drawn from an artificial syllable dataset to systematically simulate 'good' pronunciation (all phones correct), 'bad' pronunciation (all phones incorrect), and 'close' pronunciation (based on an inventory of typical Dutch phone substitutions). As anticipated, all acoustic analysis methods resulted in lower difference values for good- than for bad-pronunciation pairs. Differences for the common Dutch mistakes were also small since these mispronunciations mainly constituted substitutions within the

same phone-class. System performance was found to be broadly consistent with the correctness (or otherwise) of the reference pronunciation, using objective assessment metrics for cross-correlation, agreement, and normalised cross entropy.

POSTER BOARD 12

Detecting Summarization Hot Spots in Meetings Using Group Level Involvement and Turn-Taking Features

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

In this paper we investigate how participant involvement and turn-taking features relate to extractive summarization of meeting dialogues. In particular, we examine whether automatically derived measures of group level involvement, like participation equality and turn-taking freedom, can help detect where summarization relevant meeting segments will be. Results show that classification using turn-taking features performed better than the majority class baseline for data from both AMI and ICSI meeting corpora in identifying whether meeting segments contain extractive summary dialogue acts. The feature based approach also provided better recall than using manual ICSI involvement hot spot annotations. Turn-taking features were additionally found to be predictive of the amount of extractive summary content in a segment. In general, we find that summary content decreases with higher participation equality and overlap, while it increases with the number of very short utterances. Differences in results between the AMI and ICSI data sets suggest how group participatory structure can be used to understand what makes meetings easy or difficult to summarize.

POSTER BOARD 13

Adaptation to Regional Accented Speech Using Limited Data for Automatic Speech Recognition

Maryam Najafian, University of Birmingham
Martin Russell, University of Birmingham
Saeid Safavi, University of Birmingham
Abualsoud Hanani, Birzeit University

This poster investigates how regional accents of British English affect automatic speech recognition (ASR) performance. Given a small amount of speech (48s) from a new speaker, is it better to apply speaker adaptation, or to identify the speaker's accent and use accent-dependent ASR? Three alternative approaches to accent dependent acoustic modelling are investigated, namely using the 'true' accent model, choosing the model using an automatic accent identification (AID) system, and building a model using data from the N closest speakers in 'AID feature space', all based on 48s of speech from the test speaker. In fact, all three methods give similar performance, which is significantly better than the performance obtained with the baseline, accent-independent

model. The results show relative reductions in ASR error rate of 37% and 44% for accent-dependent models built using MAP and MLLR adaptation, respectively, compared with the baseline system. It is also been shown that using the 48s of speech to identify an appropriate accent-dependent model outperforms using the same 48s of speech for speaker-adaptation, by 35.8% for MAP- and 7.6% for MLLR-based speaker adaptation.

POSTER BOARD 14

Deep Neural Network Approach for the Dialog State Tracking Challenge

Matthew Henderson, Cambridge University
Blaise Thomson, Cambridge University
Steve Young, Cambridge University

While belief tracking is known to be important in allowing statistical dialog systems to manage dialogs in a highly robust manner, until recently little attention has been given to analysing the behaviour of belief tracking techniques. The Dialogue State Tracking Challenge has allowed for such an analysis, comparing multiple belief tracking approaches on a shared task. Recent success in using deep learning for speech research motivates the Deep Neural Network approach presented here. The model parameters can be learnt by directly maximising the likelihood of the training data. The paper explores some aspects of the training, and the resulting tracker is found to perform competitively, particularly on a corpus of dialogs from a system not found in the training.

POSTER BOARD 15

homeService: Voice-enabled assistive technology in the home using cloud-based automatic speech recognition

H. Christensen, University of Sheffield
I. Casanueva, University of Sheffield
S. Cunningham, University of Sheffield
P. Green, University of Sheffield
T. Hain, University of Sheffield

We report on the development of a system which will bring personalised state-of-the-art automatic speech recognition into the homes of people who require voice-controlled assistive technology. The ASR will be sited remotely ('in-the-cloud') and run over a broadband link. This will enable us to adapt the system to the userfbs requirements and improv the accuracy and range of the system while it is in use. We outline a methodology for this: the 'Virtuous Circle'. A case study indicates that we can obtain acceptable performance by adapting speaker-independent recognisers with 10 examples of each word in a 30-word command-and-control vocabulary. We explain the idea of a PAL - a Personal Adaptive Listener - which we intend to develop out of this study.

Poster session 3: Wednesday 13:45–15:00

POSTER BOARD 1

Automatic selection of voice donors for the voice reconstruction of individuals with vocal disorders.

Christophe Veaux, CSTR University of Edinburgh
Junichi Yamagishi, CSTR University of Edinburgh
Simon King, CSTR University of Edinburgh

When individuals lose the ability to produce their own speech, due to degenerative diseases such as motor neurone disease (MND) or Parkinson's, they lose not only a functional means of communication but also a display of their individual and group identity. In order to build personalized synthetic voices, attempts have been made to capture the voice before it is lost, using a process known as voice banking. But, for some patients, the speech deterioration frequently coincides or quickly follows diagnosis. Using HMM-based speech synthesis and speaker adaptation, it is now possible to build personalized synthetic voices with minimal data recordings and even disordered speech. When the speech has begun to deteriorate, the adapted voice model can be further modified in order to compensate for the disordered characteristics found in the patient's speech. This technique is called voice reconstruction. We present in this poster the latest advances in the voice reconstruction technique.

POSTER BOARD 2

An overview of research at Nuance

Gary Cook, Nuance Communications

As the leading supplier of speech technologies Nuance Communications undertakes research in numerous speech technology areas as well as in natural language understanding. This poster will provide an overview of some of the areas in which Nuance is currently active, and includes examples of how the research has led to enhanced products and user experiences. A more detailed description will be given of Nuance's voicemail-to-text platform and the speech technologies deployed that have been developed in Cambridge UK.

POSTER BOARD 3

Statistical Parametric Speech Synthesis Using Deep Neural Networks

Heiga Zen, Google
Andrew Senior, Google
Mike Schuster, Google

Conventional approaches to statistical parametric speech synthesis typically use decision tree-clustered context-dependent hidden Markov models (HMMs) to represent probability densities of speech parameters given texts. Speech parameters are generated from the probability densities to maximize their output probabilities, then a speech

waveform is reconstructed from the generated parameters. This approach is reasonably effective but has a couple of limitations, e.g. decision trees are inefficient to model complex context dependencies. This paper examines an alternative scheme that is based on a deep neural network (DNN). The relationship between input texts and their acoustic realizations is modeled by a DNN. The use of the DNN can address some limitations of the conventional approach. Experimental results show that the DNN-based systems outperformed the HMM-based systems with similar numbers of parameters.

POSTER BOARD 4

Cochannel Speech Segregation Using Visually-derived Binary Masks

Faheem Khan, School of Computing Sciences University of East Anglia Norwich UK
Ben Milner, School of Computing Sciences University of East Anglia Norwich UK

This paper is concerned with the problem of cochannel speech separation and exploits visual speech information to aid the separation process. Audio from a mixture of speakers is received from a single microphone and to supplement this, video from each speaker in the mixture is also captured. The visual features are used to create a time-frequency binary mask that identifies regions where the target speaker dominates. These regions are retained and form the estimate of the target speaker's speech. Experimental results compare the visually-derived binary masks with ideal binary masks which shows a useful level of accuracy. The effectiveness of the visually-derived binary mask for speaker separation is then evaluated through estimates of speech quality and speech intelligibility and shows substantial gains over the original mixture.

POSTER BOARD 5

Gaining confidence in ASR Hypotheses

Matthew Seigel, University of Cambridge
Phil Woodland, University of Cambridge

The hypothesized transcriptions generated by speech recognisers may contain errors. It is therefore valuable to have a measure of just how much confidence should be placed in these hypotheses being correct. The process of obtaining such measures is known as confidence estimation (CE). A summary of work in which a principled approach making use of conditional random field (CRF) models for CE is presented. This framework makes it possible to combine multiple (potentially correlated) information sources, in order to estimate confidence scores for ASR hypotheses. The problem of estimating both word and sub-word-level confidence scores is addressed. The sub-word-level task is investigated both as a means for improving word-level scores, and sub-word-level scores directly. A novel approach using hidden-state CRFs to model regions of confidence is shown to yield further substantial performance improvements on the word and sub-word level. An approach in which assigning scores to detections of key terms is cast as a confidence estimation problem is also taken. Applying the CRF-based framework within this setting resulted in substantial improvements in detection scores. This

is due to the direct score normalisation technique made possible by the model, as well as the effective use of informative predictor features extracted during search. Preliminary approaches in recent work investigating techniques to account for deletions in the CE process will also be presented.

POSTER BOARD 6

Real-Time Incremental HMM-based Speech Synthesis using MAGE

Rasmus Dall, University of Edinburgh
Maria Astrinaki, Universite de Mons
Alexis Moinet, Universite de Mons
Junichi Yamagishi, University of Edinburgh & NII
Simon King, University of Edinburgh
Nicolas d'Alessandro, Universite de Mons

HMM-based speech synthesis is an 'offline' process. That is, the text to be synthesised is decided and synthesised 'offline'. However in many practical applications of TTS there is a need for 'online' processing and synthesis of the speech. This work demonstrates how this may be achieved using the MAGE platform to synthesise in real time and incrementally using the HTS engine.

POSTER BOARD 7

Kernelized Log Linear Models For Continuous Speech Recognition

Shi-Xiong Zhang, University of Cambridge
M.J.F. Gales, University of Cambridge

Large margin criteria and discriminative models are two effective improvements for HMM-based speech recognition. This paper proposed a large margin trained log linear model with kernels for CSR. To avoid explicitly computing in the high dimensional feature space and to achieve the nonlinear decision boundaries, a kernel based training and decoding framework is proposed in this work. To make the system robust to noise a kernel adaptation scheme is also presented. Previous work in this area is extended in two directions. First, most kernels for CSR focus on measuring the similarity between two observation sequences. The proposed *joint* kernels defined a similarity between two observation-label sequence pairs on the sentence level. Second, this paper addresses how to efficiently employ kernels in large margin training and decoding with lattices. To the best of our knowledge, this is the first attempt at using large margin kernel-based log linear models for CSR. The model is evaluated on a noise corrupted continuous digit task: AURORA 2.0.

POSTER BOARD 8

Late Integration of Features for Acoustic Emotion Recognition

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

It is widely accepted that the ability to understand emotion or affect from speech is central to the design of more natural human-computer interfaces. This is a relatively young field. There is little consensus as to the optimum features or classifier structure for emotion recognition. This work explores the Hidden Markov Model (HMM) recognition of four affective dimensions: activation; expectation; power; and valence. A variety of features are used, some of which have never before been applied to emotion recognition. Finally, these different features are combined discriminatively to achieve a competitive performance on the AVEC 2011 affect classification task.

POSTER BOARD 9

Improved feature processing for Deep Neural Networks

Shakti Rath, University of Cambridge
Daniel Povey, Johns Hopkins University
Karel Vesel, Brno University of Technology
Jan Cernocky, Brno University of Technology

In this paper, we investigate alternative ways of processing MFCC-based features to use as the input to Deep Neural Networks (DNNs). Our baseline is a conventional feature pipe-line that involves splicing the 13-dimensional front-end MFCCs across 9 frames, followed by applying LDA to reduce the dimension to 40 and then further decorrelation using MLLT. Confirming the results of other groups, we show that speaker adaptation applied on the top of these features using feature-space MLLR is helpful. The fact that the number of parameters of a DNN is not strongly sensitive to the input feature dimension (unlike GMM-based systems) motivated us to investigate ways to increase the dimension of the features. In this paper, we investigate several approaches to derive higher-dimensional features and verify their performance with DNN. Our best result is obtained from splicing our baseline 40-dimensional speaker adapted features again across 9 frames, followed by reducing the dimension to 200 or 300 using another LDA. Our final result is about 3% absolute better than our best GMM system, which is a discriminatively trained model.

POSTER BOARD 10

Objective Speech Quality Metrics for VoIP: Comparing ViSQOL, PESQ and POLQA

Andrew Hines, Trinity College Dublin, Ireland
Naomi Harte, Trinity College Dublin, Ireland
Jan Skoglund, Google, Inc.
Anil Kokaram, Google, Inc.
Peter Pocta, University of Zilina, Slovakia
Hugh Melvin, NUI Galway, Ireland

The Virtual Speech Quality Objective Listener (ViSQOL) is an objective speech quality model. It is an intrusive, signal based full-reference metric. ViSQOL aims to predict the speech quality for the end listener whether the quality loss is due to ambient noise or transmission channel degradations. Three experiments and presented: clockdrift and warp conditions; playout adjustment conditions; noise and VoIP channel degradations. The results are compared with the ITU-T objective models for speech quality: PESQ and POLQA. The results show ViSQOL and POLQA have a higher correlation with subjective listener tests for warping and playout adjustments. POLQA was shown to have lower correlation with subjective scores than the other metrics for the NOIZEUS corpus.

POSTER BOARD 11

Identifying New Bird Species from Differences in Birdsong

Naomi Harte, Trinity College Dublin

The analysis of birdsong has increased in the speech processing community of late. Much of the reported research has concentrated on the identification of bird species from their songs or calls. A lesser reported topic is the analysis of birdsongs from subspecies of the same bird. Birdsong when combined with other biometrics, such as DNA or morphological measurements, can help determine whether historical divisions of birds into various subspecies is valid. More importantly, it can help build a case for declaring a new species of bird, with important consequences for bird conservation.

POSTER BOARD 12

Combining a Vector Space Representation of Linguistic Context with a Deep Neural Network for Text-To-Speech Synthesis

Heng Lu, University of Edinburgh
Simon King, University of Edinburgh
Oliver Watts, University of Edinburgh

Conventional statistical parametric speech synthesis relies on decision trees to cluster together similar contexts, resulting in tied-parameter context-dependent hidden Markov models (HMMs). However, decision tree clustering has a major weakness: it use hard division and subdivides the model space based on one feature at a time, frag-

menting the data and failing to exploit interactions between linguistic context features. These linguistic features themselves are also problematic, being noisy and of varied relevance to the acoustics. We propose to combine our previous work on vector-space representations of linguistic context, which have the added advantage of working directly from textual input, and Deep Neural Networks (DNNs), which can directly accept such continuous representations as input. The outputs of the network are probability distributions over speech features. Maximum Likelihood Parameter Generation is then used to create parameter trajectories, which in turn drive a vocoder to generate the waveform. Various configurations of the system are compared, using both conventional and vector space context representations and with the DNN making speech parameter predictions at two different temporal resolutions: frames, or states. Both objective and subjective results are presented.

POSTER BOARD 13

An Investigation of Single-Microphone Automatic Meeting Transcription

Takuya Yoshioka, NTT Communication Science Laboratories
Mark J. F. Gales, Cambridge University

We present our work on single-microphone automatic meeting transcription. Previous work on automatic meeting transcription has often employed beamforming techniques to obtain less distorted speech. While effective, the beamforming requires dedicated multi-microphone systems to be installed in target rooms. By contrast, single-microphone transcription systems can work with commodity sound capturing devices (such as digital voice recorders and even smartphones) and will have much wider applications. The present work evaluates how effective various automatic speech recognition components are for this task, including denoiding, dereverberation, deep neural network-based acoustic modelling, speaker adaptive training, and use of filter-bank features, in both single-pass and multi-pass decoding setups. The experimental results show that all of these components provide meaningful improvement in word error rates. We also compare the performance of our single-microphone transcription system with a multiple microphone system using BeamformIt to compare our single-microphone front-end to this well-known beamforming algorithm.

POSTER BOARD 14

Uniform Concatenative Excitation Model for Synthesising Speech without Voiced/Unvoiced Classification

J. P. Cabral, Trinity College Dublin, Ireland

In general, speech synthesis using the source-filter model of speech production requires the classification of speech into two classes (voiced and unvoiced) which is prone to errors. For voiced speech, the input of the synthesis filter is an approximately periodic excitation, whereas it is a noise signal for unvoiced. This paper proposes an excitation model which can be used to synthesise both voiced and unvoiced speech, thus over-

coming the problem of degradation in speech quality caused by those classification errors. Basically this model consists of representing two contiguous segments of the residual signal pitch-synchronously. The first segment is represented by the original residual in a fraction of the period around the pitch-mark (obtained using an epoch detector), in order to capture the most important aspects of the residual during voiced speech. Instead, the remaining part of the period is modelled by a set of parameters of the amplitude envelope of the residual waveform and its energy. The technique for synthesising the excitation combines these shaping parameters with a novel method for regeneration of the residual waveform and a method to mix a periodic signal with noise based on the Harmonic plus Noise model. Besides producing high-quality speech, this technique is computationally fast.

POSTER BOARD 15

An Empirical Study on Meeting Transcription

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

Meeting transcription is regarded as one of the most challenging tasks in speech recognition. A large proportion of previous research has focused on "scenario" meetings, in which discussion topics are specified and participants know they are being recorded, as this is the style of data usually made available. In this paper actual meetings, in this case discussion on speech recognition and speech synthesis research, are examined and the performance of state-of-art ASR system evaluated. Furthermore, this work concentrates on meeting data recorded using a single microphone array as this is felt to closely resemble the form of data collection for practical applications. A range of ASR systems are contrasted, using various acoustic models, such as Tandem and Hybrid systems, and language models, such as feed-forward and recurrent neural network language models. From the results, the real meeting data is significantly more difficult than scenario meeting data. Experimental results show that the word error rate is about 30% on scenario AMI meeting, while the word error rate is over 55%. This indicates the importance of using "real" data for final system evaluation, and any subsequent processing stage optimization.

Notes

Organizing committee	Finance & Website	Local arrangements
Arnab Ghoshal	Mark Huckvale	Rogier van Dalen
Naomi Harte		Zoi Roupakia
Peter Jancovic		
Rogier van Dalen		