UK Speech

UK Speech Conference 2019

The University of Birmingham, UK
24–25 June

UNIVERSITY OF BIRMINGHAM

Contents

Schedule

Monday 24th June

Time	Details	Location
11:00 - 12:00	Registration	Atrium, Computer Science
12:00 - 13:00	Buffet Lunch	Atrium, Computer Science
13:00 - 13:15	Welcome Message	Room 124, Chemical Engineering
13:15 - 14:15	Keynote 1	Room 124, Chemical Engineering
14:15 - 15:15	Oral Session (A)	Room 124, Chemical Engineering
15:15 - 15:45	Tea / Coffee	Atrium, Computer Science
15:45 - 17:00	Poster Session (A)	Atrium, Computer Science
18:00 - 19:30	Drinks Reception	Lapworth Museum of Geology
19:30 - late	Dinner – Tapas and Wine	Cuore restaurant (Green Heart)

Tuesday 25th June

Time	Details	Location
9:00 - 10:00	Keynote 2	Room 124, Chemical Engineering
10:00 - 11:15	Poster Session (B)	Atrium, Computer Science
11:00 - 11:30	Tea / Coffee	Atrium, Computer Science
11:30 - 12:45	Poster Session (C)	Atrium, Computer Science
12:45 - 13:45	Lunch	Atrium, Computer Science
14:00 - 14:45	Keynote 3	Room 124, Chemical Engineering
14:45 - 15:45	Oral Session (B)	Room 124, Chemical Engineering
15:45 - 16:00	Final Remarks and Farewell	Room 124, Chemical Engineering

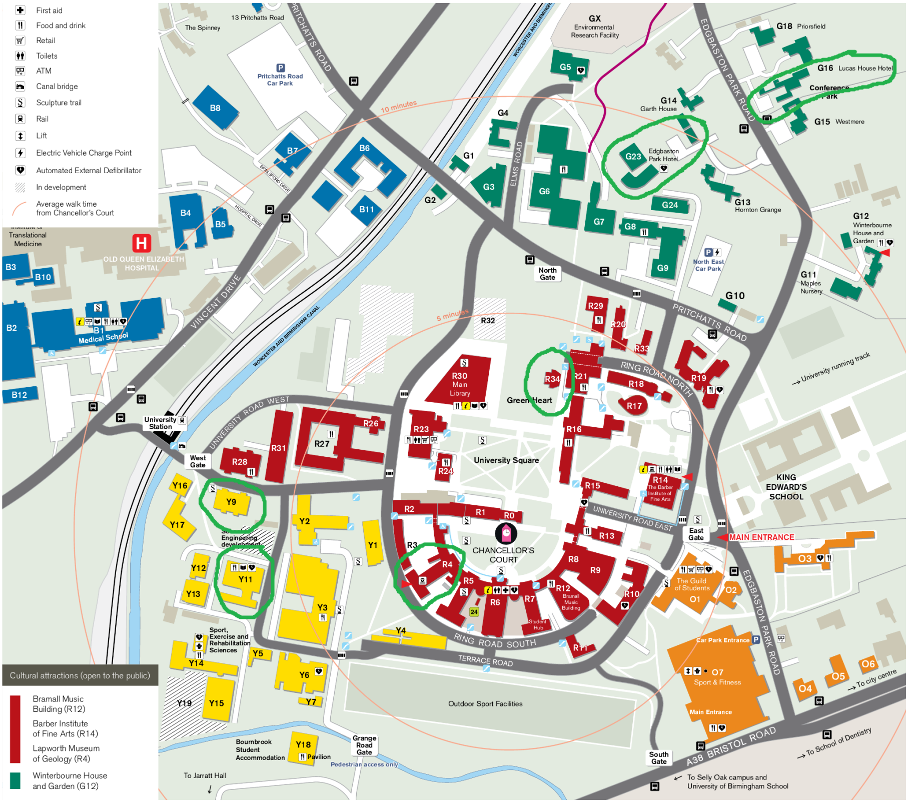
Map

All events will take place at the campus of The University of Birmingham (indicated in the map below by green circles):

Computer Science – building Y9 (yellow zone)
Chemical Engineering – building Y11 (yellow zone)

Lapworth Museum of Geology – building R4 (red zone)
Cuore restaurant – Green Heart R34 (red zone)

Edgbaston Park Hotel and Conference Centre – building G23 (green zone)
Lucas House Hotel – building G16 (green zone)



Social programme

Drinks Reception

Monday 18:00 – 19:30, Lapworth Museum of Geology

Dinner – Tapas and Wine

Monday 19:30 – late, Cuore restaurant

Lapworth Museum of Geology

The Lapworth Museum of Geology holds the finest and most extensive collections of fossils, minerals and rocks in the Midlands. Dating back to 1880, it is one of the oldest specialist geological museums in the UK.

The Museum is named after Charles Lapworth, the first Professor of Geology at Mason College, the forerunner of the University of Birmingham. Lapworth was one of the most important and influential geologists in the late 19th and early 20th Centuries. Located in the Universitys Grade II listed, Aston Webb Building, the museum retains its original Edwardian setting and interior.

A visit to the Lapworth Museum provides an insight into how the Earth formed and changed through time, and how life on earth developed and evolved.

https://www.birmingham.ac.uk/university/campus-destination/lapworth.aspx

Green Heart

A striking new parkland in the centre of the University of Birmingham's historic campus was completed in January 2019 following the completion of the new library in September 2016.

Measuring over 12 acres, the Green Heart opens up the centre of campus for students, staff and the local community to enjoy. It provides a unique space for performances, socialising, meeting and studying, while opening up views across the whole campus, as envisaged in the 1920s. The space also enhances the setting of those buildings which border the Green Heart, including the new library which opened in September 2016. It opens up new pedestrian and cycle routes, allowing students, staff and visitors to the campus to travel safely and with ease. Throughout the design process, the project team have also sought to create a sustainable, natural and environmentally friendly landscape; both for people and wildlife.

https://www.birmingham.ac.uk/university/building/green-heart/index.aspx

Keynote Talks

Keynote 1 – Monday 13:15 - 14:15

Session chair: Catherine Lai

Exploring core technologies for automated language teaching and assessment
Paula Buttery and Helen Yannakoudakis
ALTA Institute, Cambridge

Paula Buttery and Helen Yannakoudakis are members of the Automated Language Teaching and Assessment Institute (ALTA). This is an Artificial Intelligence institute that uses techniques from Machine Learning and Natural Language Processing to improve the experience of language learning online. ALTA carries out research that facilitates the creation of tools to promote the development of skills in Reading, Writing, Speaking and Listening for English language learners. In this talk, we will focus on core technologies for 1) automated assessment of learner language across these skills, and 2) automated generation of content for rapid expansion and diversification of (personalised) teaching and assessment materials. We will discuss how we can overcome some of the challenges we face in emulating human behaviour, and how we can visualise and inspect the internal 'marking criteria' and characteristics of automated models.

Keynote 2 – Tuesday 9:00 - 10:00

Session chair: Tom Merritt

Automated processing of pathological speech
Heidi Christensen
Department of Computer Science, University of Sheffield, UK

As speech technology is becoming increasingly pervasive in our lives, people with atypical speech and language are facing ever larger barriers to take full opportunity of this new technology. At the same time, recent advances in mainstream speech science and processing allows for increasingly sophisticated ways of addressing some of the specific needs that this population has. This talk will outline the major challenges faced by researcher in porting mainstream speech technology to the domain of healthcare applications; in particular, the need for personalised systems and the challenge of working in an inherently sparse data domain. Three areas in automatic processing of pathological speech will be covered: i) detection, ii) therapy/treatment and iii) facilitating communication. The talk will give an overview of recent state-of-the-art results and specific experiences from current projects in Sheffield's Speech and Hearing Group (SPandH).

Keynote 3 – Tuesday 14:00 - 14:45
Session chair: Martin Russell

The prospect of using accent recognition technology for forensic applications

Georgina Brown
University of Lancaster, UK

Forensic speech science is the forensic discipline concerned with speech recordings when they arise as pieces of evidence in a legal case or investigation. The most common type of task a forensic speech analyst is asked to conduct is forensic speaker comparison. This involves comparing multiple recordings in order to provide a view on whether or not the same speaker is featuring in these speech samples. In the UK, the most common way of approaching this task is to apply a comprehensive acoustic-phonetic analysis to these recordings. With the impressively low error rates produced by automatic speaker recognition systems, automatic speaker recognition is increasingly becoming an option for forensic speaker comparison cases. There is support for integrating such technologies into casework from the UK Forensic Science Regulator in order to boost the data-driven, repeatable and testable properties of forensic analyses (Tully, 2018). For numerous reasons, the integration of automatic speaker recognition into the UK forensic domain has been slow and work towards this is still ongoing. Forensic speaker comparison cases are not the only type of case encountered in practice. Rather than offering views on speaker identity, analysts may be asked to assess the characteristics of a speaker such as geographical background. Identifying a speakers accent could assist investigators in targeting their search for potential suspects (Watt, 2010). In view of the directions given by the UK Forensic Science Regulator, the present work has considered applying automatic accent recognition systems to these types of speaker profiling tasks (Brown 2016, 2018). This talk will discuss this research and will uncover the issues that arise.

References:

Brown, G. (2016), Automatic accent recognition systems and the effects of data on performance, *Odyssey: The Speaker and Language Recognition Workshop*, Bilbao, Spain, pp. 94–100.

Brown, G. (2018), Segmental content effects on text-dependent automatic accent recognition, *Odyssey: The Speaker and Language Recognition Workshop*, Les Sables d'Olonne, France, pp. 9–15.

Tully, G. (2018), Forensic Science Regulator Annual Report, Technical report, The UK Government, https://www.gov.uk/government/publications/forensic-science-regulator-annual-report-2018 .

Watt, D. (2010), The identification of the individual through speech, in C. Llamas & D. Watt, Eds, Language and Identities, Edinburgh University Press, Edinburgh, pp. 76–85.

Oral Session (A): Monday 14:15 – 15:15

Session chair: Simon King

Talk 1: 14:15 - 14:35
Interpretable Deep Learning Model for the Detection and Reconstruction of Dysarthric Speech
Daniel Korzekwa(1), Roberto Barra-Chicote(1), Bozena Kostek(2), Thomas Drugman(1) and Mateusz Lajszczak(1)
(1) Amazon TTS-Research,
(2) Gdansk University of Technology, Faculty of ETI, Poland

Talk 2: 14:35 - 14:55
Modern speech synthesis and its implications for speech sciences
Zofia Malisz(1), Gustav Eje Henter(1), Cassia Valentini-Botinhao(2), Oliver Watts(2), Jonas Beskow(1) and Joakim Gustafson(1)
(1) KTH Royal Institute of Technology, Stockholm, Sweden
(2) The University of Edinburgh, UK

Talk 3: 14:55 - 15:15
Continuous representations can support early phonetic learning
Yevgen Matusevych(1), Thomas Schatz(2), Sharon Goldwater(1) and Naomi Feldman(2)
(1) University of Edinburgh, UK
(2) University of Maryland, USA

Talk 1:

Interpretable Deep Learning Model for the Detection and Reconstruction of Dysarthric Speech

Daniel Korzekwa(1), Roberto Barra-Chicote(1), Bozena Kostek(2), Thomas Drugman(1) and Mateusz Lajszczak(1)
(1) Amazon TTS-Research
(2) Gdansk University of Technology, Faculty of ETI, Poland
Email: korzekwa@amazon.com

We present a novel deep learning model for the detection and reconstruction of dysarthric speech. We train the model with a multi-task learning technique to jointly solve dysarthria detection and speech reconstruction tasks. The model key feature is a low-dimensional latent space that is meant to encode the properties of dysarthric speech. It is commonly believed that neural networks are black boxes that solve problems but do not provide interpretable outputs. On the contrary, we show that this latent space successfully encodes interpretable characteristics of dysarthria, is effective at detecting dysarthria, and that manipulation of the latent space allows the model to reconstruct healthy speech from dysarthric speech. This work can help patients and speech pathologists to improve their understanding of the condition, lead to more accurate diagnoses and aid in reconstructing healthy speech for afflicted patients.

Talk 2:

Modern speech synthesis and its implications for speech sciences

Zofia Malisz(1), Gustav Eje Henter(1), Cassia Valentini-Botinhao(2), Oliver Watts(2), Jonas Beskow(1) and Joakim Gustafson(1)
(1) KTH Royal Institute of Technology, Stockholm, Sweden
(2) The University of Edinburgh, UK
Email: gustav.henter@ee.kth.se

Speech technology (e.g., speech synthesis) and speech sciences (e.g., phonetics) depend on an ongoing dialogue that benefits both fields. Insights into speech production, like source-filter separation, and perception, like the mel scale, were for example central in the development of classical formant-based synthesis technology and remain important also today. Speech sciences have also contributed towards advanced synthetic-speech evaluation methods. In return, milestones in phonetics such as evidence for categorical perception as well as advances like the motor theory of speech perception and acoustic cue analysis have relied on support from experiments on synthesised speech.

However, in recent decades the two fields have grown apart: Speech technologists have primarily pursued increasingly natural-sounding synthesis, relinquishing precise output control in the process. Speech scientists and phoneticians, meanwhile, have remained reliant on legacy synthesisers, since only these provide the careful output control necessary for phonetic studies. Unfortunately, a body of research has over the years identified substantial perceptual differences between natural speech and classical formant synthesis, casting doubt on speech-science findings from synthetic speech.

Recently, breakthroughs in deep learning have fuelled a rapid acceleration of speech-technology capabilities. In this work, we argue that modern speech synthesis with deep learning in fact has the potential to address both of the two key concerns of speech scientists – control and realism – by 1) bringing back precise control over synthetic-speech output and 2) significantly closing the perceptual gap between natural and synthetic speech. Both claims find support in recent research in speech-synthesis technology.

We supplement our two claims with an empirical evaluation contrasting classic rule-based formant synthesis (OVE III) against state-of-the-art synthesis methods, specifically speech-in-speech-out copy synthesis (MagPhase and Griffin-Lim), DNN-based statistical parametric text-to-speech (Merlin), and sequence-to-sequence neural TTS (DCTTS). The systems are compared in terms of subjective naturalness ratings as well as on a behavioural measure (response times in a lexical decision task). We find

that all modern methods vastly improve on formant synthesis naturalness and are rated above OVE III at least 99% of the time. Moreover, response times for copy-synthesis and Merlin are found not to differ notably from response times to natural speech, meaning that the troubling processing gap of older systems (including OVE III) is no longer evident.

In light of these findings and the parallel advances in synthesis control, the time is ripe for phonetics researchers to consider what modern speech-synthesis technology can do for their research problems.

Talk 3:

Continuous representations can support early phonetic learning

Yevgen Matusevych(1), Thomas Schatz(2), Sharon Goldwater(1) and Naomi Feldman(2)
(1) University of Edinburgh, UK
(2) University of Maryland, USA
Email: yevgen.matusevych@ed.ac.uk

Infants' speech perception becomes tailored to the native language over the first year of life (Werker and Tees, 1984). For example, American 10-12-month-olds discriminate English [r] and [l] better than Japanese infants do (Kuhl et al., 2006). This effect is commonly explained by phonetic category learning, yet no implemented model of such learning has been successfully demonstrated on realistic input data. Recent work presented a statistical learning model (DPGMM) that learned from raw unsegmented speech data and captured the discrimination pattern without using phonetic categories (Schatz et al., in submission). However, the DPGMM still used some categorical representations. Here we use a correspondence autoencoder (cAE; Kamper et al., 2015), a neural network that learns a continuous acoustic feature space without categorical representations from the same kind of data, using both low-level acoustic features and weak word-level supervision (under the assumption that infants rely on familiar words in phonetic learning; Feldman et al., 2013). We train the cAE on either English or Japanese speech corpus (considering two corpora per language) and test its ability to discriminate between English [r] and [l] using a machine ABX discrimination task (Schatz et al., 2013). The results show that the model captures the cross-linguistic differences in discrimination. Thus, purely continuous non-categorical representations are sufficient to explain some early perceptual changes. Since both DPGMM and cAE capture the infant-like pattern of cross-linguistic differences, pinning down the mechanisms of early phonetic learning requires testing the models on other phonetic contrasts and speech perception tasks.

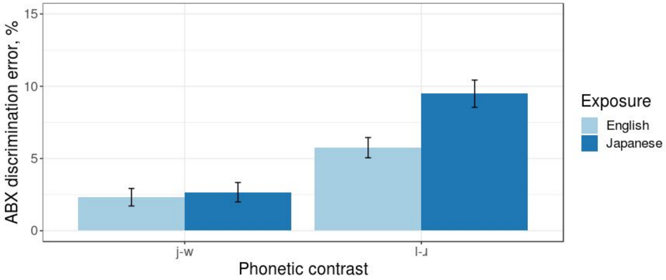


Figure 1. ABX discrimination error rate for our cAE model. Error bars show standard error over different surrounding phonetic contexts of the target sounds. Mirroring the experimental data from infants, the cAE model trained on Japanese data has a significantly higher error than the cAE model trained on English data on the target [l]–[ɹ] contrast (present in English, but not in Japanese), but not on the control [j]–[w] contrast (present in both languages), for which no cross-linguistic difference in discrimination is expected (Tsushima et al., 1994).

References

Feldman, N. H., Myers, E. B., White, K. S., Griffiths, T. L., & Morgan, J. L. (2013). Word-level information influences phonetic learning in adults and infants. *Cognition*, *127*, 427–438.

Kamper, H., Elsner, M., Jansen, A., & Goldwater, S. (2015). Unsupervised neural network based feature extraction using weak top-down constraints. In *Proceedings of ICASSP* (pp. 5818–5822).

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science*, *9*, F13–F21.

Schatz, T., Feldman, N., Goldwater, S., Cao, X., & Dupoux, E. (in submission). Early phonetic learning without phonetic categories – Insights from machine learning. PsychArXiv [Preprint.] https://doi.org/10.31234/osf.io/fc4wh

Schatz, T., Peddinti, V., Bach, F., Jansen, A., Hermansky, H., & Dupoux, E. (2013). Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline. In *INTERSPEECH 2013* (pp. 1–5).

Tsushima, T., Takizawa, O., Sasaki, M., Shiraki, S., Nishi, K., Kohno, M., Menyuk, P. & Best, C. (1994). Discrimination of English/rl/and/wy/by Japanese infants at 6-12 months: Language-specific developmental changes in speech perception abilities. In *3rd International Conference on Spoken Language Processing* (pp. 57–61).

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*, 49–63.

Oral Session (B): Tuesday 14:45 – 15:45

Session chair: Peter Jančovič

Talk 1: 14:45 - 15:05
Using generative modelling to produce varied intonation for speech synthesis
Zack Hodari, Oliver Watts and Simon King
Centre for Speech Technology Research, University of Edinburgh, UK

Talk 2: 15:05 - 15:25
Conversational systems: Why dialogue manager should consider context?
Margarita Kotti
Speech Technology Group, Toshiba Research Cambridge, UK

Talk 3: 15:25 - 15:45
Neural Network-Based Modeling of Phonetic Durations
Xizi Wei, Melvyn Hunt and Adrian Skilling
Apple Inc, UK

Talk 1:

Using generative modelling to produce varied intonation for speech synthesis

Zack Hodari, Oliver Watts and Simon King
The Centre for Speech Technology Research, University of Edinburgh, UK
Email: zack.hodari@ed.ac.uk, oliver.watts@ed.ac.uk, Simon.King@ed.ac.uk

Unlike human speakers, typical text-to-speech (TTS) systems are unable to produce multiple distinct renditions of a given sentence. This has previously been addressed by adding explicit external control. In contrast, generative models are able to capture a distribution over multiple renditions and thus produce varied renditions using sampling. Typical neural TTS models learn the average of the data because they minimise mean squared error. In the context of prosody, taking the average produces flatter, more boring speech: an "average prosody". A generative model that can synthesise multiple prosodies will, by design, not model average prosody. We use variational autoencoders (VAE) which explicitly place the most ”average” data close to the mean of the Gaussian prior. We propose that by moving towards the tails of the prior distribution, the model will transition towards generating more idiosyncratic, varied renditions. Focusing here on intonation, we investigate the trade-off between naturalness and intonation variation and find that typical acoustic models can either be natural, or varied, but not both. However, sampling from the tails of the VAE prior produces much more varied intonation than the traditional approaches, whilst maintaining the same level of naturalness.

Talk 2:

Conversational systems: Why dialogue manager should consider context?

Margarita Kotti

Speech Technology Group, Toshiba Research Cambridge, UK

margarita.kotti@crl.toshiba.co.uk

Abstract

Conversational systems is a thriving research area with applications, such as call-centers, tourist information, car navigation, education, banking, health services, and games. Commercial applications exist as well, such Microsoft's Cortana, Apple's Siri, and Amazon's Echo among others.

The PyDial case that we exploit here, is a Statistical Dialogue System (SDS) whose components are: natural language understanding, belief state tracking, policy manager, and natural language generation. In PyDial, context is not taken into account. This work: i) incorporates context information by taking into account past turns in a Toshiba patented featurisation manner; and ii) investigates two different neural network architectures, namely a DNN and a CNN, and in doing so verifying the importance of taking context into account.

The eNAC policy manager algorithm

Past turns are incorporated to the input of the dialogue manager, here eNAC. eNAC is an actor critic algorithm with a natural gradient. The update of the $\mathbf{w}$ weights of the value network is done to the same direction as θ weights of the policy network:

$$\nabla_{\boldsymbol{w}} A_{\mathbf{w}}(\mathbf{b}_t, a_t) = \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a_t|\mathbf{b}_t), \qquad (1)$$

where $\mathbf{b}$ is the belief state (BS) that policy π takes as input to produce the action a. A is the advantage function, i.e the difference of the state-action value function minus the state value function. If the objective function is $J(\boldsymbol{\theta})$, then: $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathbb{E}[\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a_t|\mathbf{b}_t) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a_t|\mathbf{b}_t)^T \mathbf{w}] = F(\theta) \cdot \mathbf{w}$ where $F(\theta)$ is the Fisher information matrix. It is true that $\mathbf{w} = F(\theta)^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ and $\mathbf{w} = \Delta \boldsymbol{\theta}_{\boldsymbol{NG}}$, where NG is the natural gradient. Once $\Delta \boldsymbol{\theta}_{\boldsymbol{NG}}$ has been found, the policy weights can be iteratively improved by $\theta' \leftarrow \theta + \beta \mathbf{w}$, where β is a step size.

DIP features

They are a Domain-Independent Parametrisation (DIP) of the BS. They map the standard BS to fixed feature space.

Experimental results

To incorporate past turns i) the standard BS representation is transformed to DIP features; ii) those DIPs are either concatenated (eNAC-flat) or stacked (eNAC-CNN) and then iii) fed to the policy manager that outputs a probability over the actions. The number of training dialogues is 1000 and of testing dialogues 100. Two figures-of-merit are used: the objective success and the number of turns. The optimal set of figures-of-merit is still an open problem for the research community.

The ontology used is Laptops11 and refers to Toshiba laptops. It has 11 requestable and 21 informable slots. Action space has 40 actions. Laptops11 has a standard BS size of 257. Those are mapped to a reduced set of 30 DIP features.

The "flat" case

In the "flat" case, the DIPs are concatenated, creating a BS of length 3*30=90. This is provided as input to eNAC-flat, as sketched in Fig. 1. Regarding the technical details, the network has 2 hidden layers with 50 and 20 neurons, the e-greedy policy starts with a value of 0.9 that linearly anneals to 0.5 after 1000 training episodes. The Adam optimiser is exploited with an initial learning rate of 0.007. The mini batch size is 6 and the capacity of the experience replay pool is 12. Results in Table 1.

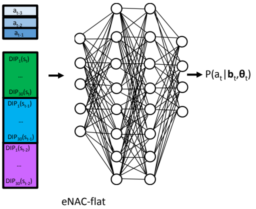


Figure 1: *eNAC-flat. The input to the system is concatenated turns in the form of DIP features & the respective actions.*

The "stacked" case

In this case, the DIPs are stacked one next to the other, creating a 2-dimensional BS of size 30(#DIP features)x3(#turns). This 30x3 "belief-state-in-context-image", is fed to a convolution layer with 50 filters of size 3x1 (so that exclusively a specific feature over turns is considered), then a ReLu activation function is applied and the output is flattened. From this point on, the NN follows the logic of the flat case. Hence, the network has 2 hidden layers with 130 and 50 neurons, the e-greedy policy starts with a value of 0.6 that linearly anneals to 0.5 after 1000 training episodes. The Adam optimiser is exploited with an initial learning rate of 0.05. The mini batch size is 64 and the capacity of the experience replay pool is 128. The architecture is sketched is Fig. 2. Results in Table 1.

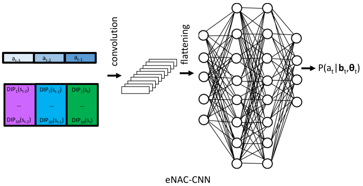


Figure 2: *eNAC-CNN. The input to the system is stacked turns in the form of DIP features & the respective actions.*

method	success	# turns
eNAC-flat	92%	6.91
eNAC-CNN	85%	6.60

Table 1: *eNAC-flat and eNAC-CNN results for 2 previous turns.*

Talk 3:

Neural Network-Based Modeling of Phonetic Durations

Xizi Wei[1]*, Melvyn Hunt, Adrian Skilling*

Apple Inc

xxw395@student.bham.ac.uk, {Melvyn_Hunt, askilling}@apple.com

Abstract

A feed-forward neural network (DNN)-based model has been developed to predict non-parametric distributions of durations of phonemes in specified phonetic contexts. It has been used to explore which factors influence durations most in (US) English. The factors included explicit phonetic context, proximity of a following pause, lexical stress, overall speaking rate, position in the syllable and word predictability from the language model. The first four of these were found to have most influence, with the explicit phonetic context being the most effective contributor to the prediction. We found that is useful to have information on at least three of the phonemes on each side of the phoneme whose duration is being predicted.

By noting outlier durations, manifested as phonemes whose durations appear to have very low probability according to the model, the model has been successfully used with text-to-speech (TTS) training speech to find departures from the script, abnormally pronounced words and misalignments. Only one male and one female speaker was examined in detail, but the relative contributions of the different factors and the overall prediction accuracy were remarkably similar for the two speakers, suggesting that the results are reasonably general, at least for professional voice talent. 30 hours of training speech was found to provide better accuracy than 10 hours.

Despite using much more speech to train the model, duration prediction is poorer with training speech for automatic speech recognition (ASR), mainly because the training corpus typically consists of single utterances from many speakers and is often noisy or casually spoken. Low-probability durations in ASR training material nevertheless mostly correspond to non-standard speech, with some having disfluencies. Children's speech is disproportionately present in these utterances, since children show much more variation in timing.

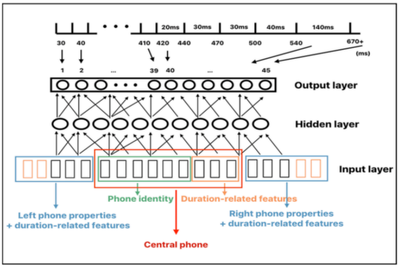


Figure 1: *An overview of the duration modeling.*

[1] The first author is a PhD candidate at the University of Birmingham. The work was carried out during her internship at Apple UK.

Poster Session (A): Monday 15:45 – 17:00

Session chairs: Mengjie Qian, Eva Fringi

POSTER 1
Data Science system for the quality assessment and monitoring of Neural Text-To-Speech on a large scale
A. Gabrys, D. Korzekwa, J. Rohnke, A. Ezzerg, R. Srikanth, G. Czachor and K. Viacheslav

POSTER 2
Computational cognitive assessment: investigating the use of an Intelligent Virtual Agent for the detection of early signs of dementia
Bahman Mirheidari, Daniel Blackburn, Ronan O'Malley, Traci Walker, Annalena Venneri, Markus Reuber and Heidi Christensen

POSTER 3
Disentangling Style Factors from Speaker Representations
Jennifer Williams and Simon King

POSTER 4
Automatic Grammatical Error Detection of Non-Native Spoken Learner English
Kate Knill(1), Mark J.F. Gales(1), Potsawee Manakul(1) and Andrew Caines(2)

POSTER 5
Exploring how phone classification neural networks learn phonetic information by visualising and interpreting bottleneck features
Linxue Bai(1), Philip Weber(2), Peter Jančovič(1) and Martin Russell(1)

POSTER 6
End-to-end speaker recognition using CNN-LSTM-TDNN
Xiaoxiao Miao(1,2) and Ian McLoughlin(1)

POSTER 7
Singing Voice Conversion with Generative Adversarial Networks
Berrak Sisman(1,2) and Haizhou Li(1)

POSTER 8
Towards the Understanding of Communicating Emotions for People with Dysarthria
Lubna Alhinti, Heidi Christensen and Stuart Cunningham

POSTER 9
Lip-Reading with Limited-Data Network
Adriana Fernandez-Lopez and Federico M. Sukno

POSTER 10
Developing Coherent Fallback Strategies for Open-domain Conversational Agents
Ioannis Papaioannou and Oliver Lemon

POSTER 11
Spontaneous conversational TTS from found data
Éva Székely, Gustav Eje Henter, Jonas Beskow and Joakim Gustafson

POSTER 12
Hierarchical RNNS for Waveform Level Speech Synthesis
Qingyun Dou, Moquan Wan, Gilles Degottex, Zhiyi Ma and Mark J.F. Gales

POSTER 13
Exploring the Trade-off between Acoustic and Language Modelling Constraints for Dysarthric Speech Recognition
Zhengjun Yue, Feifei Xiong, Heidi Christensen and Jon Barker

POSTER 14
On the Usefulness of Statistical Normalisation of Bottleneck Features for Speech Recognition
Erfan Loweimi, Peter Bell and Steve Renals

POSTER 15
Identification of geographical origin from accented speech
Wen Wu(1) and Martin Russell(2)

POSTER 16
Multitasking with Alexa: How Using Intelligent Personal Assistants Impacts Language-based Primary Task Performance
Justin Edwards(1), He Liu(1), Tianyu Zhou(1), Sandy Gould(2), Leigh Clark(1), Phillip Doyle(1) and Benjamin Cowan(1)

POSTER 1

Data Science system for the quality assessment and monitoring of Neural Text-To-Speech on a large scale

A. Gabrys, D. Korzekwa, J. Rohnke, A. Ezzerg, R. Srikanth, G. Czachor and K. Viacheslav

(1) Amazon.com, Gdansk, Pomeranian Voivodeship, Poland
(2) Amazon.com, Cambridge, Cambridgeshire, UK
Email: gabrysa@amazon.com

In this work, we describe the text-to-speech (TTS) evaluation platform. We present how we utilize Data Analysis, Speech Processing, Machine Learning, and Software Engineering to generate informative metrics on the quality of TTS voices. In this context, the metric is informative if it expedites the research and development of TTS technology. Metrics illustrated in this work allow us to identify areas in which improvement work lead to a tangible increase in final TTS quality. These metrics also help us to monitor the quality of production stage voices. We present how our platform collects the data. We describe triggers that execute the data analysis and the processes of reporting on TTS quality. To analyze the data, we use algorithms and machine learning models. We elaborate on them focusing on Neural TTS.

POSTER 2

Computational cognitive assessment: investigating the use of an Intelligent Virtual Agent for the detection of early signs of dementia

Bahman Mirheidari, Daniel Blackburn, Ronan O'Malley, Traci Walker, Annalena Venneri, Markus Reuber and Heidi Christensen

University of Sheffield, UK

Email: bmirheidari2@sheffield.ac.uk

The ageing population has caused a marked increased in the number of people with cognitive decline linked with dementia. Thus, current diagnostic services are overstretched, and there is an urgent need for automating parts of the assessment process. In previous work, we demonstrated how a stratification tool built around an Intelligent Virtual Agent (IVA) eliciting a conversation by asking memory-probing questions, was able to accurately distinguish between people with a neuro-degenerative disorder (ND) and a functional memory disorder (FMD). In this paper, we extend the number of diagnostic classes to include healthy elderly controls (HCs) as well as people with mild cognitive impairment (MCI). We also investigate whether the IVA may be used for administering more standard cognitive tests, like the verbal fluency tests. A four-way classifier trained on an extended feature set achieved 48% accuracy, which improved to 62% by using just the 22 most significant features (ROC-AUC:82%).

POSTER 3

Disentangling Style Factors from Speaker Representations

Jennifer Williams and Simon King
Centre for Speech Technology Research (CSTR)
School of Informatics, University of Edinburgh
j.williams@ed.ac.uk and Simon.King@ed.ac.uk

Our goal is to separate out speaking style from speaker identity in utterance-level representations of speech such as *i*-vectors and *x*-vectors. We adopt a working definition of style to be: *how a speaker adapts their speaking manner according to the speaking context*. We first show that both types of vectors contain information not only about speaker but also about speaking style (for the IViE data set) or emotion (for the IEMOCAP data set), even when projected into a low-dimensional space. To disentangle these factors, we use an autoencoder in which the latent space is split into two subspaces, $z1$ and $z2$. The entangled information about speaker and style/emotion is pushed apart by the use of auxiliary classifiers that take one of the two latent subspaces as input and that are jointly learned with the autoencoder.

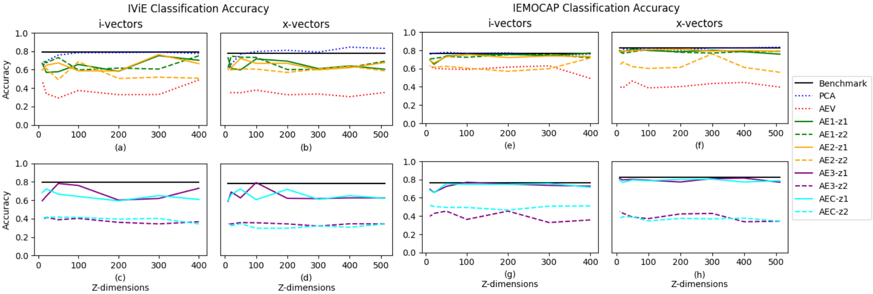


Figure 1: Classification accuracy results before disentanglement (top) and after (bottom), with benchmarks constant for comparison. The benchmarks use raw *i*-vectors or *x*-vectors respectively as input and are shown in the plots as a constant horizontal line indicating classification accuracy without any compression or disentanglement. On IViE: 79% and 78%. For IEMOCAP: 76% and 82%.

We evaluate how well the latent subspaces separate the factors by using them as input to separate style/emotion classification tasks, as shown in Figure 1. Overall, the $z2$ space has lost information about style and emotion. On the other hand, the $z1$ space has preserved it through a range of latent dimensions, while continuing to classify style/emotion close to benchmark. We have demonstrated that two types of utterance-level representation invented for speaker identification, *i*-vectors and *x*-vectors, contain information that is predictive of style and emotion. This finding suggests the existence of *style factors* that are separate from channel and other speaker-invariant characteristics. Disentangling such factors would be highly useful in many speech applications including speech-to-speech translation, speech synthesis, and speaker identification.

References

T. Asami, R. Masumura, H. Masataki, and S. Sakauchi, “Read and Spontaneous Speech Classification Based on Variance of GMM Supervectors,” in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous, “Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-To-End Speech Synthesis,” *arXiv preprint arXiv:1803.09017*, 2018.

K. Akuzawa, Y. Iwasawa, and Y. Matsuo, “Expressive Speech Synthesis via Modeling Expressions with Variational Autoencoder,” *arXiv preprint arXiv:1804.02135*, 2018.

POSTER 4

Automatic Grammatical Error Detection of Non-Native Spoken Learner English

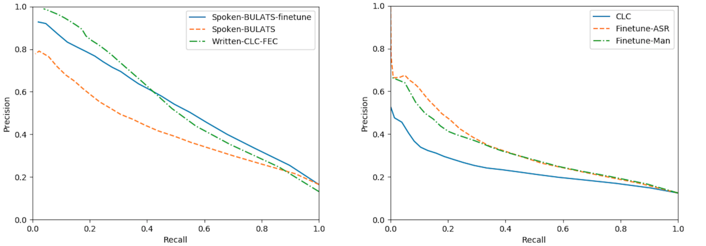
Kate Knill[1], *Mark J.F. Gales*[1], *Potsawee Manakul*[1], *Andrew Caines*[2]

[1]ALTA Institute / Engineering Department
[2]ALTA Institute / Computer Science and Technology Department
Cambridge University, UK

{kate.knill,mjfg,pm574}@eng.cam.ac.uk, apc38@cam.ac.uk

Abstract

Automatic language assessment and learning systems are required to support the global growth in English language learning. These systems must be capable of providing reliable and meaningful feedback to help learners develop their skills. This paper considers the question of detecting "grammatical" errors in non-native spoken English as a first step to providing feedback on a learner's use of English. This is a challenging problem. When speaking spontaneously even native speakers generally don't speak in full sentences, they hesitate, repeat themselves etc. These effects are accentuated in learner speech. This paper presents initial investigations into applying a state-of-the-art deep learning based grammatical error detection (GED) system [2, 3] designed for written texts to free speaking English learner tasks. Learners across the full range of proficiency levels and with a mix of first languages (L1s) are considered. This presents a number of challenges. Free speech contains disfluencies that disrupt the spoken language flow but are not grammatical errors. The lower the level of the learner the more these both will occur which makes the underlying task of automatic transcription harder. The baseline written GED system is seen to perform less well on manually transcribed spoken language. When the GED model is fine-tuned to free speech data from the target domain the spoken system is able to match the written performance as shown in Figure 1(a). A fully automatic system will use ASR to transcribe the learner's speech. When the GED is run on ASR transcriptions, however, the ability to detect grammatical errors is seen to be much lower (Figure 1(b))., even though a state-of-the-art non-native learner English ASR system was used.



a) Manual Transcriptions   b) Spoken-BULATS tests with ASR Transcriptions

Figure 1: *Precision-recall curves for written CLC-FCE-public and spoken BULATS tests with a CLC trained GED system, and fine-tuned to the BULATS data.*

1. References

[1] K. Knill, M. Gales, P. Manakul, and A. Caines, "Automatic grammatical error detection of non-native spoken learner english," in *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019.

[2] M. Rei and H. Yannakoudakis, "Compositional Sequence Labeling Models for Error Detection in Learner Writing," in *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (ACL-2016)*, 2016.

[3] M. Rei, G. K. Crichton, and S. Pyysalo, "Attending to characters in neural sequence labeling models," in *Proc. of the 26th International Conference on Computational Linguistics (COLING-2016)*, 2016.

This paper reports on research supported by Cambridge Assessment, University of Cambridge. Thanks to Cambridge English Language Assessment for supporting this research and providing access to the BULATS data. Full paper presented at ICASSP 2019 [1].

POSTER 5

Exploring how phone classification neural networks learn phonetic information by visualising and interpreting bottleneck features

Linxue Bai(1), Philip Weber(2), Peter Jančovič(1) and Martin Russell(1)
(1) University of Birmingham, UK
(2) Aston University, UK
Email: lxb190@bham.ac.uk, p.weber1@aston.ac.uk

Neural networks have a reputation for being "black boxes", into which it has been suggested that techniques from user interface development, and visualisation in particular, could give insight. We explore 9-dimensional bottleneck features (BNFs) that have been shown in our earlier work to represent speech well in the context of speech recognition, and 2-dimensional BNFs, extracted directly from bottleneck neural networks. The 9-dimensional BNFs obtained from a phone classification neural network are visualised in 2-dimensional space using linear discriminant analysis (LDA) and t-distributed stochastic neighbour embedding (t-SNE). The 2-dimensional BNF space is analysed in regard to phonetic features. A back-propagation method is used to create "cardinal" features for each phone under a particular neural network. The visualisations of both 9-dimensional and 2-dimensional BNFs show distinctions between most phone categories. In particular, the 2-dimensional BNF space seems to be a union of phonetic category-related subspaces that preserve local structures within each subspace, where the organisation of phones appears to correspond to phone production mechanisms. By applying LDA to the features of higher dimensional non-bottleneck layers, we observe a triangular pattern which may indicate that silence, friction and voicing are the three main properties learned by the neural networks.

POSTER 6

End-to-end speaker recognition using CNN-LSTM-TDNN

Xiaoxiao Miao(1)(2) and Ian McLoughlin(1)

(1) University of Kent, UK, (2) Institute of Acoustics, University of Chinese Academy of Sciences, China

Email: xm39@kent.ac.uk

Recently, end-to-end methods that map utterances to fixed-dimensional embeddings have emerged as the state-of-the-art in speaker recognition (SRE). In this paper, we aim to improve traditional DNN x-vector SRE performance by employing Convolutional and Long Short Term Memory-Recurrent (CLSTM) Neural Networks to combine the benefits of convolutional neural network front-end feature extraction and a recurrent neural to capture longer temporal dependencies. Experimental results using the speakers in the wild dataset show that CLSTM can significantly outperform traditional DNN i-vector or x-vector implementations.

POSTER 7

Singing Voice Conversion with Generative Adversarial Networks

Berrak Sisman(1,2) and Haizhou Li(1)

(1) National University of Singapore,
(2) CSTR, The University of Edinburgh, UK
Email: berraksisman@u.nus.edu

Singing voice conversion (SVC) is a task to convert the source singer's voice to sound like that of the target singer, without changing the lyrical content. So far, most of the voice conversion studies mainly focus only on the speech voice conversion that is different from singing voice conversion. We note that singing conveys both lexical and emotional information through words and tones. It is one of the most expressive components in music and a means of entertainment as well as self expression. In this paper, we propose a novel singing voice conversion framework, that is based on Generative Adversarial Networks (GANs). The proposed GAN-based conversion framework, that we call SINGAN, consists of two neural networks: a discriminator to distinguish natural and converted singing voice, and a generator to deceive the discriminator. With GAN, we minimize the differences of the distributions between the original target parameters and the generated singing parameters. To our best knowledge, this is the first framework that uses generative adversarial networks for singing voice conversion. In experiments, we show that the proposed method effectively converts singing voices and outperforms the baseline approach.

POSTER 8

Towards the Understanding of Communicating Emotions for People with Dysarthria

Lubna Alhinti, Heidi Christensen and Stuart Cunningham
The University of Sheffield, UK
Email: laalhinti1@sheffield.ac.uk

People with speech disorders may rely on augmentative and alternative communication (AAC) technologies to help them communicate. However, the limitations of the current AAC technologies act as barriers to the optimal use of these technologies in daily communication settings. The ability to communicate effectively relies on a number of factors that are not limited to the intelligibility of the spoken words. In fact, non-verbal cues play a critical role in the correct comprehension of messages and having to rely on verbal communication only, as is the case with current AAC technology, may contribute to problems in communication. This is especially true for people's ability to express their feelings and emotions, which are communicated to a large part through non-verbal cues. This paper focuses on understanding more about the non-verbal communication ability of people with dysarthria, with the overarching aim of our research being to improve AAC technology by allowing people with dysarthria to better communicate emotions. Preliminary survey results are presented that gives an understanding of how people with dysarthria convey emotions, what emotions that are important for them to get across, what emotions that are difficult for them to convey, and whether there is a difference in communicating emotions when speaking to familiar versus unfamiliar people.

POSTER 9

Lip-Reading with Limited-Data Network

Adriana Fernandez-Lopez and Federico M. Sukno
UPF
Email: adriana.fernandez@upf.edu

The development of Automatic Lip-Reading (ALR) systems is currently dominated by Deep Learning (DL) approaches. However, DL systems generally face two main issues related to the amount of data and the complexity of the model. To find a balance between the amount of available training data and the number of parameters of the model, in this work we introduce an end-to-end ALR system that combines CNNs and LSTMs and can be trained without large-scale databases. To this end, we propose to split the training by modules, by automatically generating weak labels per frames, termed visual units. These weak visual units are representative enough to guide the CNN to extract meaningful features that when combined with the context provided by the temporal module, are sufficiently informative to train an ALR system in a very short time and with no need for manual labeling. The system is evaluated in the well-known OuluVS2 database to perform sentence-level classification. We obtain an accuracy of 91.38% which is comparable to state-of-the-art results but, differently from most previous approaches, we do not require the use of external training data.

POSTER 10

Developing Coherent Fallback Strategies for Open-domain Conversational Agents

Ioannis Papaioannou and Oliver Lemon

Heriot-Watt University, UK

Email: i.papaioannou@hw.ac.uk

We first describe the problem of maintaining conversational coherence in open-domain dialogue systems such as Alana (a socialbot developed for the Amazon Alexa Challenge in 2017 and 2018). A particular issue is how to maintain coherence when the system has to fallback or recover from an error or simply avoid a dead-end in a conversation. We then present the current coherence fallback strategy implemented in the Alana system, and its performance. Finally we present directions for future work on learning the coherence strategy from data. We explore how to cast the problem as a Reinforcement Learning task, where coherence decisions may be optimised to improve conversation ratings, length,and explicit user feedback during conversations. We show initial results of fallback strategy optimisation using Reinforcement Learning.

POSTER 11

Spontaneous conversational TTS from found data

Éva Székely, Gustav Eje Henter, Jonas Beskow and Joakim Gustafson
KTH Royal Institute of Technology, Stockholm, Sweden
Email: gustav.henter@ee.kth.se

Most of human speech occurs in spontaneous conversation, making it an important goal to replicate such speech with text-to-speech (TTS). Using spontaneous conversational speech data in synthesis is however a challenge due to disfluencies, syntactic differences from written language, and general high variability. Moreover, building synthesisers from genuine spontaneous conversations found in the wild (as opposed to conversations elicited and recorded in the lab) brings further complications such as overlapping speech, lack of transcriptions, and no control over recording conditions. Taken together, these challenges mean that synthesis of conversational spontaneous speech from found data has seldom, if ever, been attempted before.
We have previously proposed to address some of the above issues by using deep learning to automatically identify and extract single-speaker breath groups (segments of speech bookended by breaths). In this study we build several Tacotron 2 voices on a corpus of 9 hours of clean single-speaker US English breath groups from a conversational podcast and transcribed using off-the-shelf ASR. Our findings from listening tests on these voices include:
1) Phonetic instead of graphemic input improved pronunciation accuracy, as did transfer learning from a larger read-speech corpus.
2) If filler tokens are left untranscribed, the stochastic synthesis will spontaneously insert filled pauses (FPs) into the output with an FP distribution broadly similar to that in the training corpus. With filler tokens transcribed, FPs are only synthesised when requested. Thus control over output FPs is possible but optional.
3) The presence of filled pauses improved perceived speaker authenticity when synthesising a sequence of extemporaneous prompts.
4) More fluent conversational TTS can be achieved by omitting disfluent utterances from the training corpus.
5) When speaking spontaneous prompts (from public speeches as well as causal conversation), our new voices were preferred over both read-speech synthesis from found data and spontaneous-speech synthesis from a small, carefully transcribed, lab-recorded corpus of spontaneous conversational speech.

POSTER 12

Hierarchical RNNS for Waveform Level Speech Synthesis

Qingyun Dou, Moquan Wan, Gilles Degottex, Zhiyi Ma and Mark J.F. Gales
Cambridge University, Engineering Department, UK
Email: qd212@cam.ac.uk

Speech synthesis technology has a wide range of applications such as voice assistants. In recent years waveform-level synthesis systems have achieved state-of-the-art performance, as they overcome the limitations of vocoder-based synthesis systems. A range of waveform-level synthesis systems have been proposed; this paper investigates the performance of hierarchical Recurrent Neural Networks (RNNs) for speech synthesis. First, the form of network conditioning is discussed, comparing linguistic features and vocoder features from a vocoder-based synthesis system. It is found that compared with linguistic features, conditioning on vocoder features requires less data and modeling power, and yields better performance when there is limited data. By conditioning the hierarchical RNN on vocoder features, this paper develops a neural vocoder, which is capable of high quality synthesis when there is sufficient data. Furthermore, this neural vocoder is flexible, as conceptually it can map any sequence of vocoder features to speech, enabling efficient synthesizer porting to a target speaker. Subjective listening tests demonstrate that the neural vocoder outperforms a high quality baseline, and that it can change its voice to a very different speaker, given less than 15 minutes of data for fine tuning.

POSTER 13

Exploring the Trade-off between Acoustic and Language Modelling Constraints for Dysarthric Speech Recognition

Zhengjun Yue, Feifei Xiong, Heidi Christensen and Jon Barker

Dept. of Computer Science, University of Sheffield, UK

Email: z.yue@sheffield.ac.uk

There has been much recent interest in building speech recognition systems for people with severe speech impairments, i.e., dysarthria. Research is progressing from isolated word recognition to more challenging connected speech scenarios. However, the datasets that are commonly used are typically designed for tasks other than ASR development (e.g., assessment). As such, they feature much overlap in the prompts used in the training and test set. Previous dysarthric acoustic modelling research has neglected this issue. Using unfairly designed language models (LMs) has potentially produced misleading, unrealistically optimistic results for continuous speech recognition. We investigate the impact of LM design using the widely used TORGO corpus, which is one of few dysarthric speech databases. In particular, we combine state-of-the-art acoustic models (AMs) with a range of LMs trained with out-of-domain (OOD) data originating from LibriSpeech. We build LMs over a range of vocabulary sizes and examine the trade-off between out-of-vocabulary (OOV) rate and recognition confusions for speakers with varying degrees of dysarthria. Although the result is on average 24.28% worse than that using the TORGO LM, specifically 37.23% for speakers with severe dysarthria, it could be a more realistic baseline LM for further exploration. It is found that in general, the greater the severity, the less complexity the LM is required to have for the best results, and that the quality of the AM also has obvious effect on the constraint. Thus not only is the choice of AM important and speaker dependent, the optimal LM complexity is also highly speaker dependent, highlighting the need to design speaker-dependent LMs alongside speaker-dependent acoustic models when considering highly variable atypical speech, for instance dysarthric speech.

POSTER 14

On the Usefulness of Statistical Normalisation of Bottleneck Features for Speech Recognition

Erfan Loweimi, Peter Bell and Steve Renals

Centre for Speech Technology Research (CSTR), School of Informatics, University of Edinburgh
{e.loweimi, peter.bell, s.renals}@ed.ac.uk

Abstract

DNNs play a central role in the state-of-the-art ASR systems. They can extract features and build probabilistic models for acoustic and language modelling. Despite their huge practical success, the level of theoretical understanding about them has remained shallow. This has triggered an expanding body of work (e.g. [1]) aiming at deciphering the DNNs as black boxes. This paper [2] investigates DNNs from a statistical standpoint.

To this end, we scrutinise the effect of activation functions on the distribution of the pre-activations (z) and activations (y). We carry out such statistical study analytically and compare the results with the results of empirical experiments. It is shown that under normal, zero-mean assumption for z ($z \mathrel{\dot\sim} \mathcal{N}(z; 0, \sigma_z^2)$) the distribution of y when using tanh activation ($y = \tanh(z)$) takes the following form

$$\begin{aligned} P_Y^{\tanh}(y) &\approx \frac{1}{1-y^2}\mathcal{N}(\frac{1}{2}\log\frac{1+y}{1-y}; 0, \sigma_z^2)) \\ &= \frac{1}{1-y^2}\frac{1}{\sqrt{2\pi}\sigma_z}\big(\frac{1+y}{1-y}\big)^{-\frac{1}{8\sigma_z^2}\log\frac{1+y}{1-y}}. \qquad (1) \end{aligned}$$

This study, among others, shows why the pre-activation (z) should be used as a feature for ASR, not the activation (y). It is demonstrated that the distribution of the pre-activations in the bottleneck layer can be well fitted with a diagonal GMM with a few Gaussians. This makes them a perfect choice for GMM-HMM systems. Figs. 1 and 2 illustrate the statistical properties of z and y when $\tanh$ and ReLU are used, respectively.

We also show how and why the ReLU activation function promotes sparsity. Histograms of the ReLU activations illustrates that there is a concentration of activation values around positive zero (0^+) (Fig. 2(d)). An important advantage of this observation is boosting the sparsity which makes the network more biologically plausible and also brings about some mathematical advantages from modelling and learning viewpoints. We believe the sparsity provided by ReLU is explainable as follows: to get the network operate in the non-linear mode, the operating point of the units should be around positive zero because before zero ReLU blocks information and after zero it acts like a linear system. Therefore, the sparsity of ReLU is due to the coincidence of zero activations with the *only* region where ReLU shows the desirable non-linear behaviour.

Motivated by the benign statistical properties of the pre-activations, the usefulness of post-processing the bottleneck (BN) feature through some statistical normalisation techniques was also investigated. In this regard, methods such as mean(-variance) normalisation, Gaussianisation, and histogram equalisation (HEQ) were employed and up to 2% (absolute) WER reduction achieved in the Aurora-4 task (Table 1).

Acknowledgements Supported by EPSRC Project EP/R012180/1 (SpeechWave).

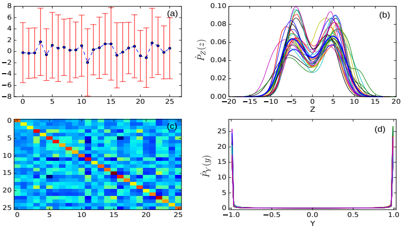


Figure 1: *Statistical analysis of preactivation (Z) and activation (Y) for all nodes in the bottleneck layer when* $\tanh$ *is used. (a) Error bar of preactivations ($\mu_z \pm \sigma_z$), (b) distribution of Z, (c) covariance matrix of Z, (d) distribution of Y.*

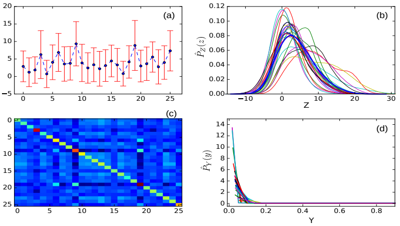


Figure 2: *Statistical analysis of preactivation (Z) and activation (Y) for all nodes in the bottleneck layer when* ReLU *is used. (a) Error bar of preactivations ($\mu_z \pm \sigma_z$), (b) distribution of Z, (c) covariance matrix of Z, (d) distribution of Y.*

Table 1: *WER for Aurora-4 (LDA-MLLT).*

Feature	A	B	C	D	*Ave4*
BN	3.87	7.96	21.80	32.72	16.58
BN+MN	3.64	7.66	21.02	32.20	16.13
BN+MVN	4.07	8.31	20.34	33.04	16.44
BN+Gauss	4.15	8.12	20.18	32.67	16.28
BN+HEQ	3.96	7.43	19.76	30.87	15.50
BN+PCA	3.75	7.88	21.56	32.46	16.41
BN+DCT	3.77	7.77	21.76	32.49	16.44

1. References

[1] E. Loweimi, P. Bell, and S. Renals, “On learning interpretable cnns with parametric modulated kernel-based filters,” in *INTERSPEECH*, 2019.

[2] ——, “On the usefulness of statistical normalisation of bottleneck features for speech recognition,” in *ICASSP*, May 2019, pp. 3862–3866.

POSTER 15

Identification of geographical origin from accented speech

Wen Wu(1) and Martin Russell(2)

(1) Dept. of Electronic, Electrical & Systems Engineering, School of Engineering, UK

(2) School of Computer Science, University of Birmingham, UK

Email: m.j.russell@bham.ac.uk

This paper investigates whether it is possible to identify the geographical origin of an individual from a sample of his or her accented speech, focussing on British English speakers who have lived in the same location for all of their lives. The problem is novel and challenging, because of the non-linear relationship between the acoustic and geographical spaces. The study uses the ABI-1 and ABI-2 speech corpora, comprising speech from approximately 20 individuals from each of 27 locations in the British Isles, i-vector representations of speech, because of their proven utility for speaker modelling, and a neural network to implement the acoustic-to-geographical mapping. Three approaches are investigated: (i) direct estimation of grid coordinates, (ii) linear interpolation based on the posterior probabilities of a range of ?reference? accents, and non-linear interpolation using a second neural network. The results demonstrate good performance for regional accents that are included in the training set, but very poor performance for those that are not. Further investigation shows that this is due to unexpected systematic acoustic differences between different parts of the corpora.

POSTER 16

Multitasking with Alexa: How Using Intelligent Personal Assistants Impacts Language-based Primary Task Performance

Justin Edwards(1), He Liu(1), Tianyu Zhou(1), Sandy Gould(2), Leigh Clark(1), Phillip Doyle(1) and Benjamin Cowan(1)

(1) University College Dublin, Ireland
(2) University of Birmingham, UK
Email: justin.edwards@ucdconnect.ie

Intelligent personal assistants (IPAs) are supposed to help us multitask. Yet the impact of IPA use on multitasking is not clearly quantified, particularly in situations where primary tasks are also language based. Using a dual task paradigm, our study observes how IPA interactions impact two different types of writing primary tasks; copying and generating content. We found writing tasks that involve content generation, which are more cognitively demanding and share more of the resources needed for IPA use, are significantly more disrupted by IPA interaction than less demanding tasks such as copying content. We discuss how theories of cognitive resources, including multiple resource theory and working memory, explain these results. We also outline the need for future work how interruption length and relevance may impact primary task performance as well as the need to identify effects of interruption timing in user and IPA led interruptions.

Poster Session (B): Tuesday 10:00 – 11:15

Session chairs: Xizi Wei, Yikai Peng

POSTER 1
On Learning Interpretable CNNs with Parametric Modulated Kernel-based Filters
Erfan Loweimi, Peter Bell and Steve Renals

POSTER 2
Non-native Speaker Verification for Spoken Language Assessment: Malpractice Detection in Speaking Tests
Linlin Wang, Yu Wang and Mark J. F. Gales

POSTER 3
Mapping Perceptions of Humanness in Intelligent Personal Assistant Interactions
Philip R Doyle, Justin Edwards, Odile Dumbleton, Leigh Clark and Benjamin R Cowan

POSTER 4
Natural Language Processing Applied to Empathy Agent for People with Mental Health Problem
Feifei Xiong(1), Fuschia Sirois(2), Katherine Easton(3,7), Abigail Millings(2), Matthew Bennion(3,7), Paul Radin(4), Ian Tucker(5), Rafaela Ganga(6) and Heidi Christensen(1,7)

POSTER 5
Speech Synthesis and Dramatic Performance: You have to Suffer Darling
Matthew P. Aylett(1), Benjamin R. Cowan(2) and Leigh Clark(2)

POSTER 6
Deep Scattering End-to-End Architectures for Speech Recognition
Iyalla John Alamina, David Wilson and Andrew Crampton

POSTER 7
Improving the intelligibility of speech playback in everyday scenarios
Carol Chermaz(1), Cassia Valentini-Botinhao(1), Henning Schepker(2) and Simon King(1)

POSTER 8
Disfluency Detection for Spoken Learner English
Yiting Lu, Mark Gales, Kate Knill, Potsawee Manakul and Yu Wang

POSTER 9
Lattice inspired semisupervised training of end to end speech recognition
Andrea Carmantini, Peter Bell and Steve Renals

POSTER 10
EFFECT OF DATA REDUCTION ON SEQUENCE-TO-SEQUENCE NEURAL TTS
Javier Latorre, Jakub Lachowicz, Jaime Lorenzo-Trueba, Thomas Merritt, Thomas Drugman, Srikanth Ronanki, Klimkov Viacheslav

POSTER 11
Diligently Delete Entry: Determining Errors in Non-Native Spontaneous Speech
John Sloan, Emma O'Neill and Julie Carson-Berndsen

POSTER 12
The University of Birmingham 2019 Spoken CALL Shared Task Systems: Exploring the importance of word order in text processing
Mengjie Qian(1), Peter Jančovič(1) and Martin Russell(2)

POSTER 13
Using Video Information to Improve Automatic Speech Recognition in the Distant Microphone Scenario
Jack Deadman and Jon Barker

POSTER 14
Exploring Generalizability of Automatic Phoneme Recognition Models
Emir Demirel(1), Sven Ahlback(2) and Simon Dixon(1)

POSTER 15
An investigation of auditory models to objectively analyze speech synthesis
Sébastien Le Maguer, Marie-Caroline Villedieu and Naomi Harte

POSTER 16
The effects of expressional feature transplant on singing synthesis
Christopher G. Buchanan, Matthew P. Aylett, and David A. Braude

POSTER 1

On Learning Interpretable CNNs with Parametric Modulating Kernel-based Filters

Erfan Loweimi, Peter Bell and Steve Renals

Centre for Speech Technology Research (CSTR), School of Informatics, University of Edinburgh
{e.loweimi, peter.bell, s.renals}@ed.ac.uk

Abstract

We investigate the problem of direct waveform modelling using modulated kernel-based filters in a convolutional neural network (CNN) framework, building on SincNet [1], a CNN employing the cardinal sine (sinc) function to implement learnable ideal (brick-wall) bandpass filters.

To this end, the general problem of learning a filterbank consisting of kernel-based baseband filters modulating a carrier is studied [2]. Each filter is characterised by the kernel parameter(s) as well as the carrier frequency which determines the centre frequency of the corresponding passband filter. The parameters are learned through backpropagation.

Compared to standard CNNs, such models have fewer parameters, learn faster and require less training data. Furthermore, they benefit from some implicit regularisation due to imposing constraint on the hypothesis space which can potentially improve the generalisation. In addition, such parametric models are more amenable to human interpretation, paving the way to embedding some perceptual prior knowledge in the network.

In this paper, we develop a general formulation for filterbank learning in a convolutional layer with parametric kernel-based filters. SincNet is a special case in which the kernel is the sinc function. Having derived the general formulation, we investigate the replacement of the rectangular filters of SincNet with triangular, gammatone and Gaussian filters. The corresponding networks are called Sinc^{2}Net, GammaNet and GaussNet. They lead to a more biologically plausible models and result in a reduction to the phone error rate (Table 2).

We also explore the properties of the filters learned for TIMIT phone recognition from both perceptual and statistical standpoints. We find that the filters in the first layer, which directly operate on the waveform, are in accord with the prior knowledge utilised in designing and engineering standard filters such as mel-scale triangular filters. That is, the networks learn to be more discriminative in perceptually significant spectral neighbourhoods (Fig. 1) and also where the data centroid is located, and the variance and entropy are highest (Fig. 2). For GammaNet, the mean of the learned order value is 4.3 (Table 1) which correlates well with Cochlea filters order, namely 4.

Finally, we consider the optimal frame length for direct waveform modelling using kernel-based filters. As Table 3 illustrates, the optimal frame length for all kernels is about 200 ms which is considerably larger than the conventional 25 ms used in Fourier-based front-ends. This observation poses the question that why 200 ms is optimal for such models. Although further exploration using other databases and tasks is warranted, possible answers include learning some kind of temporal masking or optimal syllable modelling, noting that the mean syllable length in English is 200 ms.

Acknowledgements: Supported by EPSRC Project EP/R012180/1 (SpeechWave). We benefited from discussions with Zoran Cvetkovic (KCL).

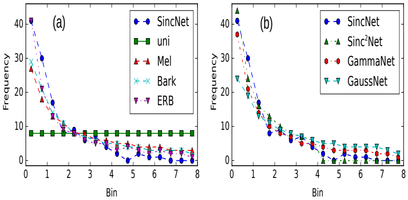


Figure 1: *Histogram of the centre frequencies (in kHz) of the kernel-based filters vs those of filterbanks designed using perceptual scales. (a) conventional filters, (b) kernel-based filters.*

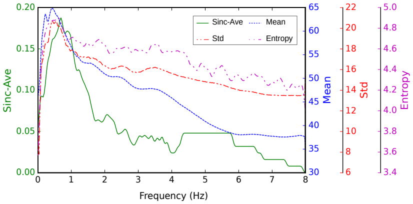


Figure 2: *TIMIT Mean/Std/Entropy for each bin vs SincNet average frequency response. All TIMIT training data is used.*

Table 1: *Statistics of the GammaNet learned filters order.*

	Mean	Median	Std	Min	Max
GammaNet	4.39	4.30	0.97	1.73	6.80

Table 2: *TIMIT PER for different kernels (200 ms).*

	MLP	CNN	Sinc	Sinc^2	Gamma	Gauss
PER	18.5	18.2	17.6	16.9	17.2	17.0

Table 3: *TIMIT PER for different frame lengths (ms).*

	25	50	100	200	300	400
CNN	30.0	21.7	18.8	18.2	18.6	19.0
SincNet	27.7	20.6	17.6	17.4	17.6	17.7
Sinc^{2}Net	27.1	20.7	17.3	16.9	17.4	17.7

1. References

[1] M. Ravanelli and Y. Bengio, "Speaker and speech recognition from raw waveform with SincNet," in *IEEE ICASSP*, 2019.

[2] E. Loweimi, P. Bell, and S. Renals, "On learning interpretable cnns with parametric modulated kernel-based filters," in *INTERSPEECH*, 2019.

POSTER 2

Non-native Speaker Verification for Spoken Language Assessment: Malpractice Detection in Speaking Tests

Linlin Wang, Yu Wang and Mark J. F. Gales

ALTA Institute / Engineering Department, Cambridge University, UK

Email: lw519@cam.ac.uk

Automatic spoken English assessment systems are becoming increasingly popular with the high demand around the world for learning of English as a second language. One challenge for these systems is to ensure the integrity of a candidate's score by detecting malpractice, which can take a range of forms. This work is focused on detecting when a candidate attempts to impersonate another in a speaking test, closely related to speaker verification, but applied in the specific domain of spoken language assessment. Deep learning based approaches have been successfully applied to a range of native speaker verification tasks with speaker representations extracted by advanced neural network models. In this work, these approaches are explored for non-native spoken English data, mainly taken from the BULATS test, which assesses English language skills for business. Though built with only limited data, systems trained on just BULATS data outperformed systems trained on the standard large speaker verification corpora of VoxCeleb. However, experimental results on large scale test sets with millions of trials have shown that, by adapting both the PLDA model and the deep speaker representations, the VoxCeleb-based systems yield lower EERs. Breakdown of impostor trials across different first languages and grades is then analysed, which shows that inter-L1 impostors are more challenging for speaker verification systems, though the grade does also influence performance.

POSTER 3

Mapping Perceptions of Humanness in Intelligent Personal Assistant Interactions

Philip R Doyle, Justin Edwards, Odile Dumbleton, Leigh Clark and Benjamin R Cowan

HCI, UCD Voysis Ltd, Ireland

Email: philip.doyle1@ucdconnect.ie

Humanness is core to speech interface design. Yet little is known about how users conceptualise perceptions of humanness and how people define their interaction with speech interfaces through this. To map these perceptions 21 participants held dialogues with a human and two Intelligent Personal Assistant interfaces, and then reflected and compared their experiences using the repertory grid technique. Analysis of the constructs show that perceptions of humanness are multidimensional, focusing on eight key themes: partner knowledge set, interpersonal connection, linguistic content, partner performance and capabilities, conversational interaction, partner identity and role, vocal qualities and behavioural affordances. Through these themes, it is clear that users define the capabilities of speech interfaces differently to humans, seeing them as more formal, fact based, impersonal and less authentic. Based on the findings, we dis- cuss how the themes help to scaffold, categorise and target research and design efforts, considering the appropriateness of emulating humanness.

POSTER 4

Natural Language Processing Applied to Empathy Agent for People with Mental Health Problem

Feifei Xiong[1], Fuschia Sirois[2], Katherine Easton[3,7], Abigail Millings[2], Matthew Bennion[3,7], Paul Radin[4], Ian Tucker[5], Rafaela Ganga[6], and Heidi Christensen[1,7]

[1]Department of Computer Science, University of Sheffield
[2]Department of Psychology, University of Sheffield
[3]School of Health and Related Research, University of Sheffield
[4]Nottinghamshire Healthcare NHS Foundation Trust
[5]School of Psychology, University of East London
[6]Institute of Cultural Capital, Liverpool John Moores University
[7]Centre for Assistive Technology and Connected Healthcare

Automated methods for answering natural language questions is an active field involving natural language processing and machine learning techniques. Great progress has been seen in designing useful and usable agent-based human-computer interact systems. The potential benefits of these natural language processing advances have also been transferred to health-related purposes, e.g., e-therapies. Recent studies have shown that e-therapy, delivering maximum impact with minimal cost in healthcare, is effective for a wide variety of psychological and emotional needs and can improve the treatment outcome over medication alone. On the other hand, such intelligent agent must be also endowed with the capability to understand service users and their intentions, motivations and feelings, often referred to as *empathy*, which often goes overlooked. This study aims to apply natural language processing technique to a mobile-based empathy agent (EA) for delivering empathetic peer-led support via smartphone for service users suffering from mental illness. This is powered by a Peer Support Community of service user experts in providing empathetic peer support, leading to a peer-to-peer support among all engaged service users.

The proposed EA will deliver an intelligent personalised human-centred mental health advisor that provides empathetic advice according to service user mental health related query. The EA will allow for service user to provide queries using free text, and automatically select empathetic and personalised responses, drawing from the EA engine trained using the rated response bank. Natural language processing technique is applied to accomplish this task with three main modules: (i) analysis of the service user queries posed in natural language; (ii) analysis of knowledge derived from the response bank (associated to the specific-domain queries); (iii) response retrieval and extraction that can satisfy the information needs of service users. To this end, a bank of 530 supportive responses categorized into 7 groups from the Peer Support Community is firstly generated within 5 typical request scenarios. An online survey is conducted to rate the response bank, in which 205 participators, with lived experience of mental health issues, were invited to evaluate the responses using a rating scale from 1 to 7 in terms of how appropriate, how empathetic and how useful.

Moreover, the proposed EA is designed to receive the rating feedback from service users, so that the pairs of query and answer can be fine-tuned to ensure that the EA performance improves its level of empathetic responding over time. Initial evaluation of the EA system will be launched via a web-interface with invited participants from the Peer Support Community to provide their ratings of appropriateness, usefulness, and particularly degree of perceived empathy.

POSTER 5

Speech Synthesis and Dramatic Performance: You have to Suffer Darling

Matthew P. Aylett, Benjamin R. Cowan, Leigh Clark

CereProc Ltd., Edinburgh, UK
University College Dublin, Ireland
matthewa@cereproc.com, benjamin.cowan@ucd.ie, leigh.clark@ucd.ie

Abstract

Siri, Ivona, Google Home, and most speech synthesis systems have voices which are based on imitating a neutral citation style of speech and making it sound natural. But, in the real world, darling, people have to act, to perform! In this paper we will talk about speech synthesis as performance, why the uncanny valley is a bankrupt concept, and how academics can escape from studying corporate speech technology as if it's been bestowed by God.

Figure 1: *Wax work of Jack Nicholson. Looks like Jack but doesn't perform like Jack and is very, very creepy. "Jack Nicholson figure at Madame Tussauds Hollywood" by lorenjavier is licensed under 🅭🅯⊜2.0*

Col Jessep: *I'll answer the question. You want answers?*
LTJG Kaffee: *I think I'm entitled to them.*
Col Jessep: *You want answers?!*
LTJG Kaffee: *I want the truth!*
Col Jessep: *You can't handle the truth!*
- A Few Good Men[1]

[1] https://www.youtube.com/watch?v=5j2F4VcBmeo

POSTER 6

Deep Scattering End-to-End Architectures for Speech Recognition

Iyalla John Alamina, David Wilson and Andrew Crampton
University of Huddersfield, UK
Email: john.alamina@hud.ac.uk, d.r.wilson@hud.ac.uk

This work explores the prospects of deep recurrent end-to-end architectures applied to speech recognition. Complementary aspects of developing speech recognition systems are eliminated by focusing on end-to-end speech units as a two-step process requiring a Connectionist Temporal Character Classification (CTCC) model and Language Model (LM) rather than a three-step process requiring an Acoustic model (AM), LM and phonetic dictionary. A two-step process rather than a three-step process is particularly desirable for low resource languages as resources are required to build only two models instead of three models. Our Bi-directional Recurrent neural network (Bi-RNN) end-to-end system, is augmented by features derived from a deep scattering network as opposed to the standard Mel Cepstral (MFCC) features used in state of the art acoustic models. These specialised deep scattering features, consumed by the Bi-RNN, model a light-weight convolution network. This work shows that it is possible to build a speech model from a combination of deep scattering features and a Bi-RNN. There has been no record of deep scattering features being used in end-to-end bi-RNN speech models as far as we are aware.

POSTER 7

Improving the intelligibility of speech playback in everyday scenarios

Carol Chermaz(1), Cassia Valentini-Botinhao(1), Henning Schepker(2) and Simon King(1)

(1) The Centre for Speech Technology Research, The University of Edinburgh, UK
(2) Dept. Medical Physics and Acoustics and Cluster of Excellence Hearing4all, University of Oldenburg, Germany
Email: c.chermaz@sms.ed.ac.uk

Speech playback is common in everyday life: from radio to TV, from laptops to PA systems in public spaces. The intelligibility of the message being played might be compromised by noise and reverberation, which are present to some degree in every real-world situation. NELE (Near End Listening Enhancement) algorithms try to tackle the issue by modifying the signal before it is played back, in order to make it more intelligible for the listener. Such technologies are often tested in lab-controlled conditions (e.g. against "speech shaped noise"), which might yield inaccurate predictions on their performance. For this reason, we simulated two representative scenarios with real binaural noise recordings and reverberation, and we tested a selection of state of the art NELE algorithms (plus unmodified speech). The algorithms we chose feature different approaches: noise-dependent and noise-independent strategies, with or without a specific compensation for reverberation. The results we obtained from a listening test with N=30 normal hearing listeners suggest that different strategies might be more suitable for different environments; however, realistic listening conditions prove to be harder in respect to lab controlled noise, which is reflected in the psychometric curves we obtained for plain speech in our realistic environments (in comparison to previous studies which featured controlled noise). Realistic noise and reverberation are possibly much harder to harness in comparison to lab noise, but we believe that a more reliable prediction of speech intelligibility might be worth the extra effort.

POSTER 8

DISFLUENCY DETECTION FOR SPOKEN LEARNER ENGLISH

Yiting Lu, Mark Gales, Kate Knill, Potsawee Manakul, Yu Wang
Department of Engineering
University of Cambridge

May 24, 2019

Disfluencies often present in spontaneous speech and can make spoken language more challenging than written text. A standard disfluency structure [1] looks like:

$$\text{I want a flight [} \underbrace{\textit{to Denver}}_{\text{reparandum}}\ \underbrace{\textit{uh I mean}}_{\text{interregnum}}\ + \underbrace{\text{to Atlanta}}_{\text{repair}}\text{]}$$

Disfluency detection helps to make speech transcriptions more text-like and allows advanced text processing techniques to be applied to automatic spoken language processing. In this study, we are interested in applying disfluency detection (DD) to non-native spoken English for computer-assisted language learning (CALL). A bi-directional LSTM based sequence tagging DD model was used [2]. The model achieved state-of-the-art performance for sequence-labeling approaches on the Switchboard corpus. CALL needs ASR to transcribe the data. On both Switchboard and non-native English learner spoken corpora, DD performance drops on these errorful transcriptions. The performance of a downstream grammatical error detection (GED) [3] task on the non-native corpora is helped, however, by using automatic DD versus no DD.

Corpus	Test	DD-F_1	DD-processing	GED-$F_{0.5}$
NICT-JLE	REF	79.8	none	36.5
			auto	**43.7**
			man	49.2
BULATS	REF	64.0	none	38.3
			auto	**41.4**
			man	42.0
	ASR	44.6	non	23.7
			auto	**24.1**
			man	24.4

Table 1: Disfluency detection (DD) & Grammatical error detection (GED) performance on non-native data.

On NICT-JLE, GED $F_{0.5}$ gained 12.7 by manually removing disfluencies, and running automatic disfluency removal achieved a 7.2 absolute gain. A proprietary BULATS corpus was used to extend the investigation to transcriptions generated by an ASR system. ASR transcriptions were produced using a joint stacked hybrid DNN and LSTM system with an overall WER of 25.6%. DD is significantly disrupted by ASR errors. Automatic disfluency removal improved GED $F_{0.5}$ performance from the baseline of 23.7 to 24.1 and 38.3 to 41.4 on ASR and manual transcriptions, respectively. Future work will explore more advanced neural network models to improve cross-domain application. We also seek to improve the performance on ASR transcriptions by combining ASR confidence scores in DD training.

References

[1] Elizabeth Ellen Shriberg. *Preliminaries to a theory of speech disfluencies*. PhD thesis, Uni. of California at Berkeley, 1994.

[2] Vicky Zayats, Mari Ostendorf, and Hannaneh Hajishirzi. Disfluency detection using a bidirectional LSTM. In *Proc. INTERSPEECH 2016*, pages 2523–2527, 2016.

[3] K. Knill, M. Gales, P. Manakul, and A. Caines. Automatic grammatical error detection of non-native spoken learner english. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2019.

POSTER 9

Lattice inspired semisupervised training of end to end speech recognition

Andrea Carmantini, Peter Bell and Steve Renals
University of Edinburgh, UK
Email: A.carmantini@ed.ac.uk

End-to-end models require large amounts of data to obtain good performance. For most domain and languages, this data is not easily obtained. In our work, we investigate semisupervised training as a solution to the adaptation to less well resourced domain. By interpreting the beam search of an attentional sequence to sequence model as a lattice, we calculate approximated state occupancy probabilities for unsupervised data, then use the generated vectors as adaptation targets. Our method gave us ∼14% relative improvement in WER when using 20 hours of data.

POSTER 10

EFFECT OF DATA REDUCTION ON SEQUENCE-TO-SEQUENCE NEURAL TTS

Javier Latorre, Jakub Lachowicz, Jaime Lorenzo-Trueba, Thomas Merritt, Thomas Drugman, Srikanth Ronanki, Klimkov Viacheslav

Amazon

Recent speech synthesis systems based on sampling from autoregressive neural networks models can generate speech almost undistinguishable from human recordings. However, these models require large amounts of data. This paper shows that the lack of data from one speaker can be compensated with data from other speakers. The naturalness of Tacotron2-like models trained on a blend of 5k utterances from 7 speakers is better than that of speaker dependent models trained on 15k utterances, but in terms of stability multi-speaker models are always more stable. We also demonstrate that models mixing only 1250 utterances from a target speaker with 5k utterances from another 6 speakers can produce significantly better quality than state-of-the-art DNN-guided unit selection systems trained on more than 10 times the data from the target speaker.

POSTER 11

Diligently Delete Entry: Determining Errors in Non-Native Spontaneous Speech

John Sloan, Emma O'Neill and Julie Carson-Berndsen
University College Dublin, Ireland
Email: Julie.Berndsen@ucd.ie

Emotional Response Language Education (ERLE) is a personalised, e-learning platform which allows learners to interact with the system via speech or text. Speech input is recognised by an ASR system (Google's speech to text API) and presented back to the learner for confirmation that the utterance is indeed what s/he wished to say before it becomes part of the conversation with ERLE. Learners may make a number of attempts to rectify pronunciation or grammatical errors if the ASR does not output what they were trying to say and sometimes they even edit the text output to correct it. As a result, the system is collecting a corpus of non-native spontaneous speech from adult, L2 English language learners which has essentially been labelled by the learners themselves. An initial analysis of some of this data has been carried out with the aim of determining whether errors are due to pronunciation, to change of choice of vocabulary or grammar, or to the language model. The utterance "I feel like I need to study more diligently for the final exam" was what one learner wanted to say, but the ASR recognised "I feel like I need to study more delete entry for the final exam" which points to a pronunciation error. On the other hand, the change from "was" to "am" in the two utterances "I was wondering" and "I am wondering" points to an intentional change of grammar. This poster will present the types of error found, the inter-annotator agreement and the way in which they can be used to provide more useful feedback to the learner.

POSTER 12

The University of Birmingham 2019 Spoken CALL Shared Task Systems: Exploring the importance of word order in text processing

Mengjie Qian(1), Peter Jančovič(1) and Martin Russell(2)

(1) Dept. of Electronic, Electrical & Systems Engineering, University of Birmingham, UK

(2) School of Computer Science, University of Birmingham, UK

Email: mxq486@bham.ac.uk

This paper describes the systems developed by the University of Birmingham for the 2019 Spoken CALL Shared Task (ST) challenge. The task is automatic assessment of grammatical and semantic aspects of English spoken by German-speaking Swiss teenagers. Our system has two main components: automatic speech recognition (ASR) and text processing (TP). We use the ASR system that we developed for 2018 ST challenge. This is a DNN-HMM system based on sequence training with the state-level minimal Bayes risk criteria. It achieved word-error-rates (WER) of 8.89% for the ST2 test set and 11.13% for the ST3 test set. This paper focuses on development of the TP component. In particular, we explore machine learning (ML) approaches which preserve different degrees of word order. The ST responses are represented as vectors using Word2Vec and Doc2Vec models and the similarities between ASR transcriptions and reference responses are calculated using Word Movers Distance (WMD) and Dynamic Programming (DP). A baseline rule-based TP system obtained a Dfull score of 5.639 and 5.313 for the ST2 and ST3 test set, respectively. The best ML-based TP, consisting of a Word2Vec model trained on the ST data, DP-based similarity calculation and a neural network, achieved D_{full} score of 7.244 and 5.777 for ST2 and ST3 test sets, respectively.

POSTER 13

Using Video Information to Improve Automatic Speech Recognition in the Distant Microphone Scenario

Jack Deadman and Jon Barker
The University of Sheffield, UK
Email: jdeadman1@sheffield.ac.uk

One of the most difficult challenges in Automatic Speech Recognition (ASR) is recognising speech from people who are far away from the microphone. The difficulty comes from the reverberation and competing sound sources in the containing room, corrupting the desired signal. The work in this Ph.D. will explore how the additional modality of video information can be used to improve ASR systems. In this work, the CHiME5 dataset is used as the main data source. The dataset consists of a series of dinner parties with 3 distinct stages (cooking, dining and after-dinner socialising). The parties are recorded using Microsoft Kinect devices which have a 1080p camera and a 4-channel linear microphone array.

The video information can be used in the preprocessing stage of the ASR pipeline. People-tracking techniques have been deployed to monitor the movement patterns of the people in the scenes. Using this information, microphone beamforming algorithms can be directed to enhance the signal in directions where speakers appear to be active whilst suppressing audio in competing directions. Adaptive beamformers such as MVDR and GEV are being explored.

The work also seeks to explore how video information can be used in improving demixing the multiple sound sources, especially during a period where speech overlaps. We seek to develop novel techniques to integrate the video information into source separation. There have been several very recent advances in single channel source separation that exploit deep learning. The most promising techniques include Deep Clustering, SpeakerBeam and TasNet. In all of these approaches, there are clear opportunities to improve performance by exploiting video-based speaker localisation, speaker identity or speech information. We will be exploring extensions of these approaches using a combination of real and simulated multispeaker data.

The initial aim will be to optimise and evaluate source separation with respect to standard speech enhancement objectives (SNR, SDR, etc), but the eventual goal is to optimise the signal enhancement with respect to the ASR objective function in an end-to-end system. Initial ideas in this direction will be presented.

POSTER 14

Exploring Generalizability of Automatic Phoneme Recognition Models

Emir Demirel(1), Sven Ahlback(2) and Simon Dixon(1)

(1) Queen Mary University of London, UK
(2) Doremir Music Research AB
Email: e.demirel@qmul.ac.uk

Human speech and singing voice are both produced by the same sound source, the vocal organ. Despite its growing popularity in the last decades, phoneme / word recognition in singing voice has not been widely investigated as it is in the speech domain. According to prior research, one of the major differences between speech and singing is the duration of vowels. This can be interpreted as difference in pronunciation of the voiced phonemes. Phoneme recognition in singing is still not a solved problem due to complex spectral characteristics of the sung vowels. In this study, we tackle this problem using recent and traditional Automatic Speech Recognition (ASR) models that are trained on different speech corpora. To observe the influence of pronunciation, we hold experiments on 'NUS Sung and Read Lyrics Corpus', which consists of lyrics-level utterances both pronounced as speech and singing. We perform the experiments using the Kaldi ASR Toolkit and explore different topologies in the Kaldi PyTorch extension. We further analyze the recognition and the alignment results on both singing and speech, and address the problems to achieve a better recognition result in singing. We have obtained some initial results where we observed a decrease in recognition performance when context dependency is added to the feature space. This indicates that it is necessary to include domain specific information in the phoneme recognition pipeline when applied to singing voice.

POSTER 15

An investigation of auditory models to objectively analyze speech synthesis

Sébastien Le Maguer, Marie-Caroline Villedieu, Naomi Harte

ADAPT Centre, Sigmedia Lab, EE Engineering, Trinity College Dublin, Dublin, Ireland

Speech synthesis is a domain attracting a lot of attention, as recently improvements have yielded very high quality speech. However, while the quality is improving, the models are getting more and more complicated. In addition to this, protocols to analyze what is captured by the models remains sparse.

The main objective analysis protocols which are used in speech synthesis evaluation nowadays focus mainly on the use of distances like the Mel-Cepstral Distortion (MCD) for the filter part, or the Root Mean Square Error (RMSE) for the Fundamental Frequency (F0) and the duration. Model-based evaluation, such as the use of Gaussian mixture model (GMM)[1] or Hidden Markov Model (HMM)[2], has also been introduced. However, the main idea underlying these protocols is to quantify what information is lost or changed compared to the original data. This is done without making any assumption about what information is important for the human listener. In order to move towards to an objective evaluation protocol which suitably embeds the human listener, we propose to simulate a listener using an Auditory Nerve (AN) model[3]. Then we use the output of such a model, a neurogram, to achieve a GMM based evaluation as well as a dedicated neurogram distance used successfully in speech intelligibility measurement: the Neurogram Similarity Index Measure (NSIM).

In order to investigate the feasibility of using AN models and neurograms in an objective analysis context, we conducted a set of experiments following three main hypothesis. First we used the Blizzard Challenge 2013 results[4] to examine the consistency of the results achieved using the proposed analysis methodology with subjective evaluation results. Then, we applied the proposed protocol to analyze the evolution of HMM models considering the enrichment of the linguistic descriptive feature set. Finally, we replicated the HMM analysis using deep neural network (DNN) models.

The results show that a naive approach to using neurograms to evaluate Text-To-Speech (TTS) doesn't correlate strongly with the subjective evaluation results. However, they seem to capture different properties than the evaluation based on the use of the spectrogram representation. Therefore, in the future we want to further investigate the use of AN models in speech synthesis evaluation, as this approach could bring an interesting new way to analyze speech synthesis production.

1. References

[1] S. L. Maguer, N. Barbot, and O. Boeffard, "Evaluation of contextual descriptors for hmm-based speech synthesis in french," in *Eighth ISCA Workshop on Speech Synthesis*, 2013.

[2] T. H. Falk and S. Moller, "Towards signal-based instrumental quality diagnosis for text-to-speech systems," *IEEE Signal Processing Letters*, vol. 15, pp. 781–784, 2008.

[3] I. C. Bruce, Y. Erfani, and M. S. Zilany, "A phenomenological model of the synapse between the inner hair cell and auditory nerve: Implications of limited neurotransmitter release sites," *Hearing research*, vol. 360, pp. 40–54, 2018.

[4] S. King and V. Karaiskos, "The blizzard challenge 2013," 2013.

POSTER 16

The effects of expressional feature transplant on singing synthesis

Christopher G. Buchanan, Matthew P. Aylett, David A. Braude

CereProc Ltd., Edinburgh, UK
{chrisb,dave,matthewa}@cereproc.com

Abstract

Professional quality singing in modern contemporary music is easily recognisable by most listeners, with notable performers able to exhibit a combination of singing techniques and melodic expression closely associated with their identity. Whilst pitch accuracy is one metric that mirrors a listener's positive impression of a performer, other aspects of pitch and expression usually employed by professionals to embellish their performance can also be considered. In this paper we explore the effect of "sustained segment transplant", a process targeting regions of vocal melodies containing melodic expressions such as vibrato, and its relevance to the perception of skill in the singing voice. We extract and analyse sustained segments from performers of varying ability, include these as auxiliary features in music score-derived markup before feeding into both a source-filter based vocoder and our CereVoice singing synthesiser. Our hypothesis is that the transfer of expressive features of a professional vocalist (donor) to a non-professional (patient) can improve the perceived quality of the patient's singing in copy-vocoded and end-to-end synthesis, and test this via subjective listening test. Results show the two singers' performances were perceived as significantly different in singing quality. This distinction extends to processed examples supporting the use of singing synthesis as a new methodology to support the analysis of singing quality. However, both tuning and sustained segment transfer modification methods did not show an improvement in the patient's sung performance, and the degradational nature of vocoding and end-to-end synthesis corrupts these expressional pitch features such that they become imperceptible to a human listener.

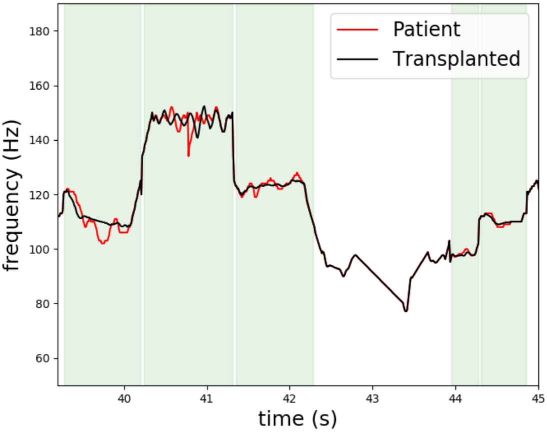


Figure 1: *A section of the patient's F0 track before and after the transplant procedure.*

Poster Session (C): Tuesday 11:30 – 12:45

Session chairs: Guy Coop, Eva Fringi

POSTER 1
Speaker-independent classification of phonetic segments from raw ultrasound in child speech
Manuel Sam Ribeiro, Aciel Eshky, Korin Richmond and Steve Renals

POSTER 2
Time Domain Multi-device Speech Separation
Jisi Zhang and Jon Barker

POSTER 3
What Makes a Good Conversation?: Challenges in Designing Truly Conversational Agents
Leigh Clark(1), Nadia Pantidi(2), Orla Cooney(1), Philip Doyle(3), Diego Garaialde(1), Justin Edwards(1), Brendan Spillane(4), Emer Gilmartin(4), Christine Murad(5), Cosmin Munteanu(6), Vincent Wade(4) and Benjamin R. Cowan(1)

POSTER 4
Bi-directional Lattice Recurrent Neural Networks for Confidence Estimation
Anton Ragni, Qiujia Li, Preben Ness and Mark Gales

POSTER 5
Windowed Attention Mechanisms for Speech Recognition
Shucong Zhang, Erfan Loweimi, Peter Bell and Steve Renals

POSTER 6
Automatically Discovering the Special Relationship between Modalities in Audio-Visual Speech Recognition
George Sterpu, Christian Saam and Naomi Harte

POSTER 7
Speaker Diarization using Odd-Even Mel-Frequency Cepstral Coefficients
Ahmed Isam Ahmed(1), John P. Chiverton(1), David L. Ndzi(2) and Victor M. Becerra(1)

POSTER 8
Sequence-to-sequence neural TTS: an assessment of the contribution of various ingredients
Oliver Watts(1), Gustav Eje Henter(2), Jason Fong and Cassia Valentini-Botinhao(1)

POSTER 9
Unaccompanied sung speech recognition: a state-of-the-art ASR baseline
Gerardo Roa Dabike and Jon Barker

POSTER 10
Sequence Teacher-Student Training of Acoustic Models For Automatic Free Speaking Language Assessment
Yu Wang, Jeremy Wong, Mark Gales, Kate Knill and Anton Ragni

POSTER 11
In Other News: A Bi-style Text-to-speech Model for Synthesizing Newscaster Voice with Limited Data
Nishant Prateek, Mateusz ajszczak, Roberto Barra-Chicote, Thomas Drugman, Jaime Lorenzo-Trueba, Thomas Merritt, Srikanth Ronanki, Trevor Wood

POSTER 12
Datasets, voices, and ethics: Update on Grassroot Wavelengths
David A. Braude, Matthew P. Aylett and Skaiste Butkute

POSTER 13
Transfer Learning for Personalised Dysarthric Speech Recognition
Feifei Xiong, Jon Barker, Zhengjun Yue and Heidi Christensen

POSTER 14
"Sorry, I didnt catch that" How do Training Corpora influence Algorithmic Bias in Automatic Speech Recognition?
Meghan Avery(1) and Martin Russell(2)

POSTER 1

Speaker-independent classification of phonetic segments from raw ultrasound in child speech

Manuel Sam Ribeiro, Aciel Eshky, Korin Richmond and Steve Renals

The Centre for Speech Technology Research, University of Edinburgh, UK

Email: sam.ribeiro@ed.ac.uk

Ultrasound tongue imaging (UTI) provides a convenient way to visualize the vocal tract during speech production. UTI is increasingly being used for speech therapy, making it important to develop automatic methods to assist various time-consuming manual tasks currently performed by speech therapists. A key challenge is to generalize the automatic processing of ultrasound tongue images to previously unseen speakers. In this work, we investigate the classification of phonetic segments (tongue shapes) from raw ultrasound recordings under several training scenarios: speaker-dependent, multi-speaker, speaker-independent, and speaker-adapted. We observe that models underperform when applied to data from speakers not seen at training time. However, when provided with minimal additional speaker information, such as the mean ultrasound frame, the models generalize better to unseen speakers. This paper was presented and published in the proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

POSTER 2

Time Domain Multi-device Speech Separation

Jisi Zhang and Jon Barker
University of Sheffield, UK
Email: jzhang132@sheffield.ac.uk, j.p.barker@sheffield.ac.uk

Recent speech separation techniques have achieved a significant improvement by integrating neural networks and clustering algorithms. Most of these methods address the separation problem in the time-frequency domain, by assigning time-frequency bins to individual sources. A recently proposed end-to-end single channel speech separation method, TasNet, operates directly on the time domain signal and outperforms previous time-frequency domain approaches by a large margin. Extracting features from raw signals can jointly make use of magnitude and phase information, which have been shown useful for this separation problem. Also, reconstruction errors caused by inverse Fourier Transform can be reduced with the end-to-end approach. This work aims to extend TasNet to microphone array and multiple microphone array situations. First, a convolutional spatial encoder has been introduced to the single channel TasNet for extracting spatial features from each single microphone array. The spectral features and spatial features are then concatenated as input to a seperator network for generating mask functions. In the multiple devices case, the features extracted from each array are subsequently combined using a temporal convolutional network consisting of dilated convolutions, which is able to solve sampling rate mismatch and asynchronization problems. The network is trained and evaluated on a simulated database (spatialized wsj0-mix) which uses room simulation to model distant-microphone recordings of utterances spoken simultaneously in a reverberant environment. Experiments have shown using only 2 channels in a single array can lead to a signal to distortion ratio (SDR) improvement of 10.7 dB. Further results using multiple devices will be presented at the meeting.

POSTER 3

What Makes a Good Conversation?: Challenges in Designing Truly Conversational Agents

Leigh Clark(1), Nadia Pantidi(2), Orla Cooney(1), Philip Doyle(3), Diego Garaialde(1), Justin Edwards(1), Brendan Spillane(4), Emer Gilmartin(4), Christine Murad(5), Cosmin Munteanu(6), Vincent Wade(4) and Benjamin R. Cowan(1)

(1) University College Dublin, Ireland
(2) University College Cork, Ireland
(3) Voysis Ltd.
(4) Trinity College Dublin, Ireland
(5) University of Toronto, Canada
(6) University of Toronto, Mississauga
Email: leigh.clark@ucd.ie

Conversational agents promise conversational interaction but fail to deliver. Efforts often emulate functional rules from human speech, without considering key characteristics that conversation must encapsulate. Given its potential in supporting long-term human-agent relationships, it is paramount that HCI focuses efforts on delivering this promise. We aim to understand what people value in conversation and how this should manifest in agents. Findings from a series of semi-structured interviews show people make a clear dichotomy between social and functional roles of conversation, emphasising the long-term dynamics of bond and trust along with the importance of context and relationship stage in the types of conversations they have. People fundamentally questioned the need for bond and common ground in agent communication, shifting to more utilitarian definitions of conversational qualities. Drawing on these findings we discuss key challenges for conversational agent design, most notably the need to redefine the design parameters for conversational agent interaction.

POSTER 4

Bi-directional Lattice Recurrent Neural Networks for Confidence Estimation

Anton Ragni, Qiujia Li, Preben Ness and Mark Gales

University of Cambridge, UK

Email: ar527@cam.ac.uk

The standard approach to mitigate errors made by an automatic speech recognition system is to use confidence scores associated with each predicted word. In the simplest case, these scores are word posterior probabilities whilst more complex schemes utilise bi-directional recurrent neural network (BiRNN) models. These neural network approaches are highly flexible and have shown promising results in confidence estimation. The standard BiRNNs are fundamentally limited to processing sequential input such as 1-best hypotheses. A number of upstream and downstream applications, however, rely on confidence scores assigned not only to 1-best hypotheses but to all words found in confusion networks or lattices. These include but are not limited to speaker adaptation, semi-supervised training and information retrieval. To make improved confidence scores more generally available, this work shows how recurrent models such as BiRNNs can be extended from 1-best sequences to confusion network and lattice structures. Experiments are conducted using one of the Cambridge University submissions to the IARPA OpenKWS 2016 competition. The results above show that confusion network and lattice-based BiRNNs can provide a significant improvement in confidence estimation.

POSTER 5

Windowed Attention Mechanisms for Speech Recognition

Shucong Zhang, Erfan Loweimi, Peter Bell and Steve Renals
University of Edinburgh, UK
Email: s1603602@sms.ed.ac.uk

The usual attention mechanisms used for encoder-decoder models do not constrain the relationship between input and output sequences to be monotonic. To address this we explore windowed attention mechanisms which restrict attention to a block of source hidden states. Rule-based windowing restricts attention to a (typically large) fixed-length window. The performance of such methods is poor if the window size is small. In this paper, we propose a fully-trainable windowed attention and provide a detailed analysis on the factors which affect the performance of such an attention mechanism. Compared to the rule-based window methods, the learned window size is significantly smaller yet the model?s performance is competitive. On the TIMIT corpus this approach has resulted in a 17% (relative) performance improvement over the traditional attention model. Our model also yields comparable accuracies to the joint CTC-attention model on the Wall Street Journal corpus.

POSTER 6

Automatically Discovering the Special Relationship between Modalities in Audio-Visual Speech Recognition

George Sterpu, Christian Saam and Naomi Harte

Sigmedia Lab, School of Engineering, ADAPT Centre, Trinity College Dublin, Ireland
Email: sterpug@tcd.ie

This work analyses the recently proposed Audio-Visual Speech Recognition (AVSR) strategy AV Align, previously shown to improve the speech recognition accuracy by up to 30% in noisy conditions over an audio system alone on the laboratory recorded TCD-TIMIT dataset. Since AV Align explicitly models the alignment between the auditory and visual modalities of speech, we examine the cross-modal alignment patterns under multiple conditions, exposing the difficulty of learning visual representations in an end-to-end framework given a dominant audio modality. To address this problem, we propose to apply a secondary loss function aimed at learning to regress two lip-related Facial Action Units directly from the visual representations. We find that the proposed enhancement effectively nudges the system to discover monotonic cross-modal alignments on the largest publicly available AVSR dataset LRS2. Furthermore, we report performance improvements of up to 30% on this challenging dataset when capitalising on the visual modality, without making use of additional pretraining data required by alternative AVSR methods. We also report a direct comparison with the more popular Watch, Listen, Attend, and Spell architecture, showing the superiority of AV Align. This result reinforces the suitability of learning cross-modal correlations in AVSR.

POSTER 7

Speaker Diarization using Odd-Even Mel-Frequency Cepstral Coefficients

Ahmed Isam Ahmed[1], John P. Chiverton[1], David L. Ndzi[2] and Victor M. Becerra[1]

[1] School of Energy and Electronic Engineering, University of Portsmouth,
Portsmouth, UK, PO1 3DJ
ahmed.ahmed5@myport.ac.uk, john.chiverton@port.ac.uk, victor.becerra@port.ac.uk
[2] School of Computing, Engineering and Physical Sciences, University of the West of Scotland,
Paisley, UK, PA1 2BE
david.ndzi@uws.ac.uk

In Ahmed et al. (2019), we introduced Odd-Even Mel-Frequency Cepstral Coefficients (OE-MFCC) as an improvement over the widely used MFCC. This new acoustic feature extraction focuses on the role of the filter bank analysis as opposed to the Discrete Cosine Transform (DCT). In OE-MFCC, the conventional filter bank is split into odd and even non-overlapping filters subsets. In the correlation matrix of each of the odd and even filters' energies, the residual correlation is lower than the case of the full filters set. The DCT is applied to the filters' energies of each subset separately and the cepstral coefficients are concatenated. Our previous paper evaluated the performance of OE-MFCC in speaker verification. The evaluation is extended here to assess the performance of OE-MFCC in speaker diarization. This speaker recognition modality attempts to identify who spoke and when in a multi-speaker conversation. The binary key based diarization system (Anguera and Bonastre, 2011; Delgado et al., 2015) is used in the evaluation here. In this system, a sequence of acoustic feature vectors is represented by a 896 dimensional vector of binary values. The diarization process is started with 16 uniform clusters. Segments of 3 seconds length are assigned to the clusters based on maximum Jaccard coefficient value among their binary keys. Then, the most similar clusters are merged. This is repeated until only one cluster remains. The Within Cluster Sum of Squares (WCSS) is used to select the best number of clusters. In that case, one cluster hypothetically represents one speaker. Then, final re-segmentation takes place as described in Delgado et al. (2015). The RT-05S dataset is used in the evaluation. The Diarization Error Rate (DER) is 32% using 23 dimensional MFCC and 29% using 20 dimensional OE-MFCC. The number of filters in the filter bank is 24 and the 0^{th} order coefficients are discarded. The results reported here indicate that OE-MFCC has the potentials to improve the performance of speaker recognition in general.

References

Ahmed, A. I., Chiverton, J. P., Ndzi, D. L., Becerra, V. M., 2019. Speaker recognition using PCA-based feature transformation. Speech Communication 110, 33 – 46.

Anguera, X., Bonastre, J.-F., 2011. Fast speaker diarization based on binary keys. In: Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. IEEE, pp. 4428–4431.

Delgado, H., Anguera, X., Fredouille, C., Serrano, J., 2015. Fast single-and cross-show speaker diarization using binary key speaker modeling. IEEE/ACM Transactions on Audio, Speech, and Language Processing 23 (12), 2286–2297.

POSTER 8

Sequence-to-sequence neural TTS: an assessment of the contribution of various ingredients

Oliver Watts(1), Gustav Eje Henter(2), Jason Fong and Cassia Valentini-Botinhao(1)

(1) Edinburgh University, UK
(2) KTH Royal Institute of Technology, Stockholm, Sweden
Email: owatts@inf.ed.ac.uk

Sequence-to-sequence neural networks with attention mechanisms have recently been widely adopted for text-to-speech. Compared with older, more modular statistical parametric synthesis systems, sequence-to-sequence systems feature three prominent innovations: 1) They replace substantial parts of traditional fixed front-end processing pipelines (like Festival's) with learned text analysis; 2) They jointly learn to align text and speech and to synthesise speech audio from text; 3) They operate autoregressively on previously-generated acoustics. Performance improvements have been reported relative to earlier systems which do not contain these innovations. It would be useful to know how much each of the various innovations contribute to the improved performance. We here propose one way of associating the separately-learned components of a representative older modular system, specifically Merlin, with the different sub-networks within recent neural sequence-to-sequence architectures, specifically Tacotron 2 and DCTTS. This allows us to swap in and out various components and subnets to produce intermediate systems that step between the two paradigms; subjective evaluation of these systems then allows us to isolate the perceptual effects of the various innovations. We report on the design, evaluation, and findings of such an experiment.

POSTER 9

Unaccompanied sung speech recognition: a state-of-the-art ASR baseline

Gerardo Roa Dabike, Jon Barker

Speech and Hearing Research Group (SPandH), University of Sheffield, Sheffield, UK
{groadabike1, j.p.barker}@sheffield.ac.uk

Abstract

Automatic sung speech recognition is a relatively understudied and challenging task that has been held back by a lack of extensive and freely available datasets. Previous systems have reported poor performances when addressing this problem by using spoken speech in conjunction with traditional adaptation techniques [1, 2] or by using force alignment techniques to annotate non-annotated datasets [3]. In this research, we processed a new annotated karaoke dataset DAMP Sing! [4], released in early 2018, to construct a replicable baseline system for un-accompaniment recreational singing [5]. The DAMP Sing! corpus is organised by country of origin of the recording - 30 countries in total; this allowed the division of the training set into three different sizes using the country information for data augmentation; DSing1, DSing3 and DSing30. The smallest training dataset, DSing1, was constructed using 80% of the recordings from Great Britain, the remaining 20% was reserved for development and test sets (10% each). The DSing3 training set uses recordings from Australia and USA to augment DSing1, and the largest training set, DSing30, uses all the English language recordings from all 30 countries. For the development and test sets, high-quality utterance alignments and the transcriptions have been generated using human annotators. A baseline ASR system has been constructed using a TDNN-F acoustic model and state-of-the-art lattice-free MMI training techniques implemented in Kaldi toolkit [6]. The baseline obtains a best WER of 19.7%, similar than human performance in some experiments [7]. Significantly, our experiments show that training with the larger DSing30 dataset produces best results despite it being dominated by non-native English.

Now that a solid baseline has been established, we are exploring the benefits of using non-conventional musical features, such as pitch and beat tracking, to improve the performance of the recogniser. Later in the project, we will tackle the challenge of singer enhancement, i.e. isolating singing from musical accompaniment, with a view to then joining the source separation and speech recognition stages into a single DNN-based system.

References

[1] Mesaros, A. and Virtanen, T. (2010). “Automatic recognition of lyrics in singing”. In *Eurasip Journal on Audio, Speech, and Music Processing*, volume 2010.

[2] Tsai, C. P., Tuan, Y. L., and Lee, L. S. (2018). “Transcribing Lyrics from Commercial Song Audio: The First Step Towards Singing Content Processing. ”. In *ICASSP 2018.*

[3] Kruspe, A.M. (2016) “Retrieval of Textual Song Lyrics from Sung Inputs”. In *INTERSPEECH 2016.*

[4] Smule Sing! 300x30x2 Dataset, *https://ccrma.stanford.edu/damp/*, accessed September 2018.

[5] Roa Dabike, G and Barker, J.(2019). “Automatic lyric transcription from Karaoke vocal tracks: Resources and a Baseline System”. Unpublished paper submitted to *INTERSPEECH 2019.*

[6] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). “The Kaldi Speech Recognition Toolkit”. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.* IEEE Signal Processing Society.

[7] Collister, L. B., and Huron, D. (2008). “Comparison of Word Intelligibility in Spoken and Sung Phrases”. *Empirical Musicology Review*, 3(3), 109125.

POSTER 10

Sequence Teacher-Student Training of Acoustic Models For Automatic Free Speaking Language Assessment

Yu Wang, Jeremy Wong, Mark Gales, Kate Knill, Anton Ragni

University of Cambridge, Engineering Dept., Trumpington St., Cambridge, CB2 1PZ, U.K.

With increasing global demand for learning English as a second language, there has been considerable interest in methods of automatic assessment of spoken language proficiency for use in interactive electronic learning tools as well as for auto-marking candidates for formal qualifications, especially for free speaking English tests. A high performance automatic speech recognition (ASR) system is an important constituent component of an automatic language assessment system. The ASR system is required to be capable of recognising non-native spontaneous English speech and to be deployable under real-time conditions. The performance of ASR systems can often be significantly improved by leveraging upon multiple systems that are complementary, such as an ensemble. Ensemble methods, however, can be computationally expensive, often requiring multiple decoding runs, which makes them impractical for deployment. One approach to making the decoding with ensembles practical is to compress the ensemble into a single model using teacher-student training. Standard teacher-student training trains the student to replicate the average performance of the ensemble. In speech recognition systems this is normally implemented by propagating the average frame posterior from the teachers to the student, ignoring the sequential nature of speech. This can limit the ability of the student to replicate the performance of the ensemble. To address this problem sequence teacher-student training has recently been proposed where the hypothesis posterior distribution is propagated from the teacher ensemble to the student. In this work sequence teacher-student training is used to train an acoustic model for non-native English speech recognition. This method allows a single student model to emulate the performance of an ensemble of teachers but without the need for multiple decoding runs, thereby allowing for real-time applications. Adaptations of the student model to speakers from different first languages (L1s) and grades are also explored.

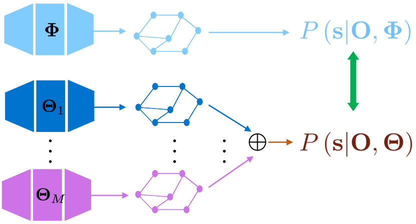

$\mathbf{\Phi}$
$\mathbf{\Theta}_1$
$\mathbf{\Theta}_M$
$P(\mathbf{s}|\mathbf{O},\mathbf{\Phi})$
$\oplus$
$P(\mathbf{s}|\mathbf{O},\mathbf{\Theta})$

Figure 1: Sequence teacher-student training. The state sequence posteriors of the student model, $\mathbf{\Phi}$, is given by $P(\mathbf{s}|\mathbf{O},\mathbf{\Phi})$ and the state sequence posteriors of the combined teacher ensemble, $\mathbf{\Theta}$, is given by $P(\mathbf{s}|\mathbf{O},\mathbf{\Theta})$. The goal of teacher-student training is to train the single student model to emulate the combined ensemble by minimising the KL-divergence between $P(\mathbf{s}|\mathbf{O},\mathbf{\Theta})$ and $P(\mathbf{s}|\mathbf{O},\mathbf{\Phi})$.

POSTER 11

In Other News: A Bi-style Text-to-speech Model for Synthesizing Newscaster Voice with Limited Data

Nishant Prateek, Mateusz ajszczak, Roberto Barra-Chicote, Thomas Drugman, Jaime Lorenzo-Trueba, Thomas Merritt, Srikanth Ronanki, Trevor Wood

Amazon

Neural text-to-speech synthesis (NTTS) models have shown significant progress in generating high-quality speech, however they require a large quantity of training data. This makes creating models for multiple styles expensive and time-consuming. In this paper different styles of speech are analysed based on prosodic variations, from this a model is proposed to synthesise speech in the style of a newscaster, with just a few hours of supplementary data. We pose the problem of synthesising in a target style using limited data as that of creating a bi-style model that can synthesise both neutral-style and newscaster-style speech via a one-hot vector which factorises the two styles. We also propose conditioning the model on contextual word embeddings, and extensively evaluate it against neutral NTTS, and neutral concatenative-based synthesis. This model closes the gap in perceived style-appropriateness between natural recordings for newscaster-style of speech, and neutral speech synthesis by approximately two-thirds.

POSTER 12

Datasets, voices, and ethics: Update on Grassroot Wavelengths

David A. Braude, Matthew P. Aylett, Skaiste Butkute

CereProc Ltd., Edinburgh, UK

{dave,matthewa,skaiste}@cereproc.com

Abstract

Grassroot Wavelengths (GW) is an H2020 project for developing communities through local radio with the RootIO platform (Fig. 1). One of the big issues with community radio stations is ensuring there is enough content. Within the GW project we have been exploring the use of TTS. Here we present TTS outputs from the first half of the project.

We have created a a public domain "Living Audio Dataset" (LADs), a platform for sharing and more importantly *crowd-building* audio data in a structured format. LADs contains tools for helping develop new languages as well as recording scripts and audio data. Thus far LADs has recordings in English (RP), Dutch (Holland), Irish (non-native), and Russian (Moscow). LADs is available via GitHub.

As part of the project new functionality was added to the Idlak fork of Kaldi. Firstly we have added a normaliser, for which the language resources have been developed to process English, Dutch, Irish, and Russian. We also have written a Python wrapper for Idlak. Using this wrapper we wrote a RESTful server for TTS, which uses Idlak-Tangle voices for synthesis.

During a recording session in Ireland for a locally accented (Bere Island) English voice we faced a number of challenges, 1. A lack of understanding from our unpaid volunteer concerning what speech synthesis was and what it was for. 2. A serious concern over loss of control over the use of their voice. 3. The lack of an appropriate procedure for ensuring informed consent. This raises the important distinction between voices recorded from professional speakers for commercial use and unpaid volunteers for community projects and research.

Finally we developed some commercial voices to help local partners. We have released an Irish voice, and are in the process of recording and releasing a Romanian voice. Like all CereProc voices, they are free to use for academic research.

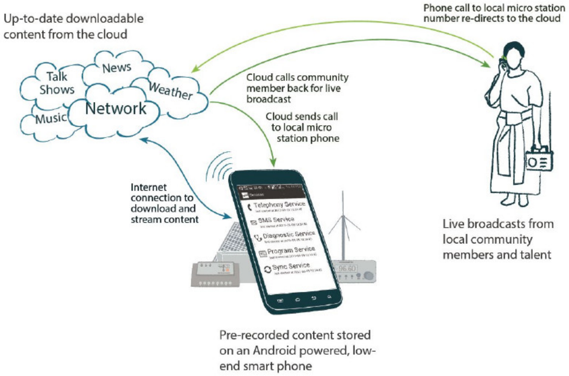

Up-to-date downloadable content from the cloud
Talk Shows
News
Weather
Network
Music
Phone call to local micro station number re-directs to the cloud
Cloud calls community member back for live broadcast
Cloud sends call to local micro station phone
Internet connection to download and stream content
Telephony Service
SMS Service
Diagnostic Service
Program Service
Sync Service
Live broadcasts from local community members and talent
Pre-recorded content stored on an Android powered, low-end smart phone

Figure 1: *RootIO's technical stack. TTS will be integrated into the RootIO system to generate automatic radio program content, and IVR prompts*

Acknowledgements

This research is funded by the European Union's Horizon 2020 research and innovation programme under grant ID No. 780890.

POSTER 13

Transfer Learning for Personalised Dysarthric Speech Recognition

Feifei Xiong, Jon Barker, Zhengjun Yue, Heidi Christensen

Department of Computer Science, University of Sheffield, UK

{f.xiong,j.p.barker,z.yue,heidi.christensen}@sheffield.ac.uk

Dysarthric speech recognition is a challenging research field as the available in-domain data is very sparse and it is difficult to collect more. In this work, we investigate transfer learning which attempts to utilise the out-of-domain data to improve personalised speech recognition models for speakers with dysarthria. A neural network model trained solely on out-of-domain data is adapted onto the specific domain utilising the limited available data from target dysarthric speakers.

First of all, a systematic experiment is conducted to analyse various impacts arising from the transferred layers and the amount of target data. Experimental results using UASpeech corpus show that the linear components in hidden layers play the most important role in transfer learning for an improved modelling of dysarthric speech. In comparison to the conventional speaker-dependent training and data combination strategy, transfer learning achieves 4% and 1.7% absolute word error rate reduction on average, respectively. Furthermore, results show that the best performance for speakers with dysarthria severity from moderate-severe to severe comes from data combination from other dysarthric speakers. This indicates that if the target domain is too dissimilar to source domain, a brute-force transfer might be not the best option, which also motivates the second part of this work as follows.

In order to further improve the transferability towards the target domain particularly in the cases of moderate and severe dysarthria, an utterance-based data selection strategy is proposed based on the entropy of posterior probability that is shown to be followed using a Gaussian distribution. It actively selects the potentially beneficial data for either increasing the in-domain training pool or constructing an intermediate domain for incremental transfer learning. Results show that the proposed utterance-based data selection outperforms the selection schemes based on the similarity measure in terms of the perceptual speech intelligibility or the final recognition performance. More specifically with the utterance-based data selection, data combination outperforms incremental learning for speakers with severe dysarthria, resulting in an averaged 2.1% absolute word error rate reduction compared to the base transferred model from the first part of this work. On the other hand, for moderate-severe and moderate groups, incremental learning is superior in general.

POSTER 14

“Sorry, I didnt catch that” How do Training Corpora influence Algorithmic Bias in Automatic Speech Recognition?

Meghan Avery(1) and Martin Russell(2)

(1) Department of Liberal Arts and Natural Sciences, University of Birmingham, UK

(2) School of Computer Science, University of Birmingham, UK

Automatic Speech Recognition is a rapidly developing form of data-driven learning, but despite the infiltration of ASR into everyday life, there has been anecdotal evidence to show that this technique does not work consistently for different social groups. This paper evaluates this bias more rigorously and finds significant gaps in performance for different social demographics, ruling that ASR performs more accurately for standard US and UK English, and much less effectively for non-native accents and minority accents such as Scottish and New Zealand. The paper then investigates how far this bias can be attributed to training corpora by analysing both open-source and proprietary training sets. It is concluded that open-source corpora suffer from demographically unlabelled and unbalanced data which propagates bias in ASR; proprietary corpora cannot be directly investigated, but from analysing home speaker system usage statistics, an imbalance can be inferred. In the future, the development of spontaneous speech corpora has the potential to exacerbate this issue. It is acknowledged that no corpus will realistically be able to incorporate all accents, but through sensible corpus design and labelling corpora demographically, performance gaps in ASR can be mitigated significantly.

Notes

Organizing committee:	Martin Russell Peter Jančovič Mengjie Qian Xizi Wei Yikai Peng Guy Coop Eva Fringi
UK Speech committee:	Catherine Lai Tom Merritt