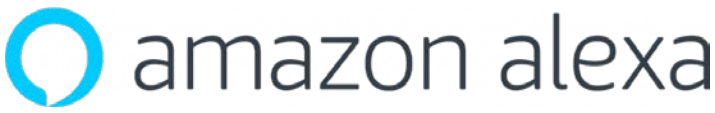
UK Speech Conference 2018

Trinity College Dublin

June 25th-26th

With Thanks to our Sponsors



Contents

Programme UK Speech 2018

Monday June 25th

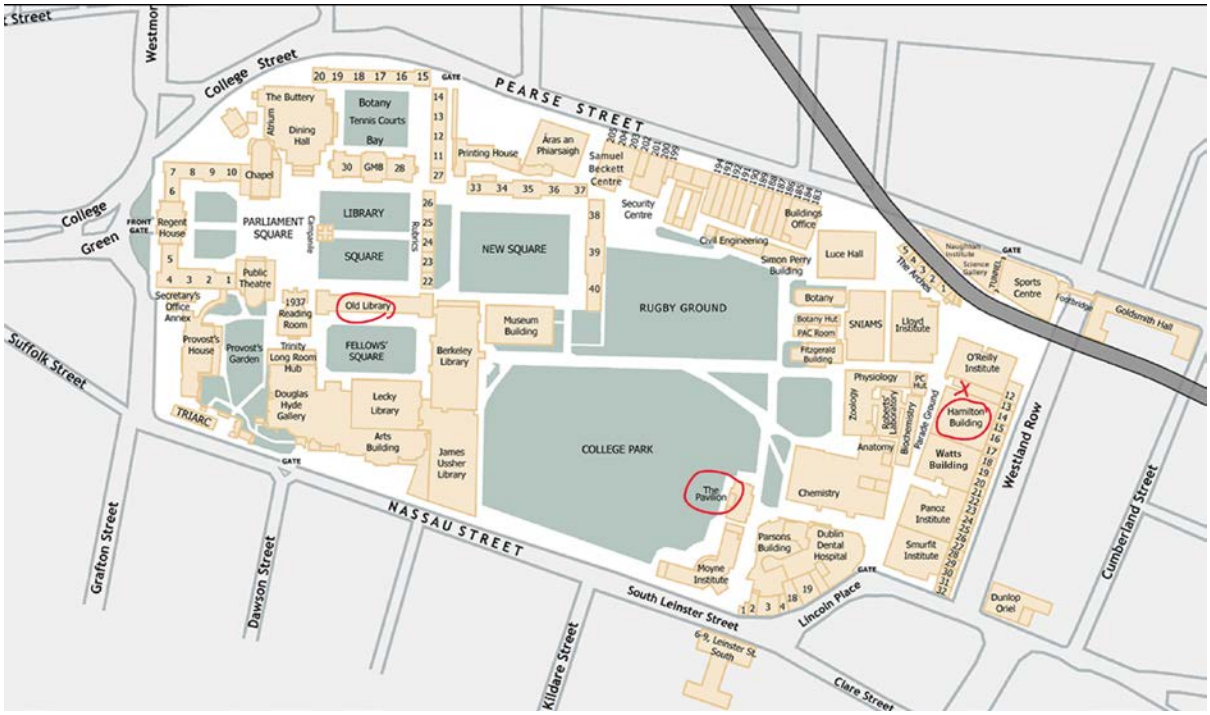
Time	Location	Session
11:00 – 12:00	Hamilton Ground Floor	Registration (continues until 13:00 for those not having lunch)
12:00 – 13:00	Hamilton Ground Floor	Lunch
13:00 – 13:15	Hamilton Lecture Room	Welcome Message
13:15 – 14:15	Hamilton – Joly Theatre	Keynote 1
14:15 – 15:30	Hamilton Ground Floor	Poster Session A
15:30 – 16:00	Hamilton Ground Floor	Coffee Break
16:00– 17:00	Hamilton – Joly Theatre	Oral session A
18:00 – 19:30	Trinity Old Library	Drinks reception
19:30 til late	The Pavillion Bar	Barbecue dinner

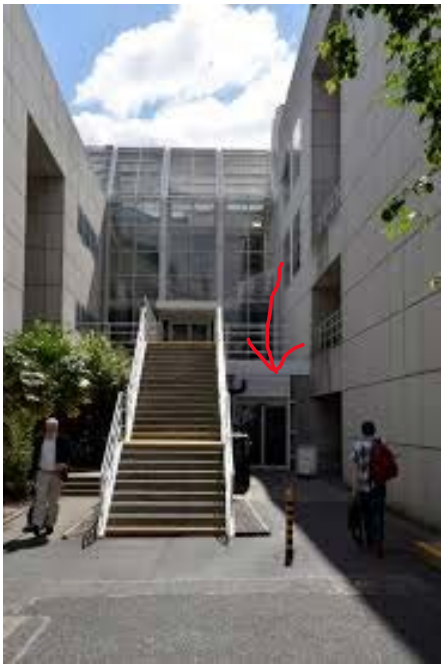
Tuesday June 26th

Time	Location	Session
08:40 – 09:00	Hamilton Ground Floor	Registration
09:00 – 10:00	Hamilton – Joly Theatre	Keynote 2
10:00 – 11:00	Hamilton Ground Floor	Poster Session B
11:00 – 11:30	Hamilton Ground Floor	Coffee Break
11:30 – 12:30	Hamilton – Joly Theatre	Oral session B
12:30 – 13:30	Hamilton Ground Floor	Lunch
13:30 – 14:15	Hamilton – Joly Theatre	Keynote 3
14:15 – 15:15	Hamilton Ground Floor	Poster session C
15:15 – 15:30	Hamilton – Joly Theatre	Final remarks and farewell

Note: Joly Theatre is on the first floor of the Hamilton Building.

Map and Directions





The Conference will be held in the Hamilton Building on the main campus of TCD. For registration, head to the Ground Floor of the building. This is the entrance marked “X” on the map above, and with an arrow on the picture to the left here.

Reception on Monday:

Long Room Library, marked as the “Old Library” on the map above

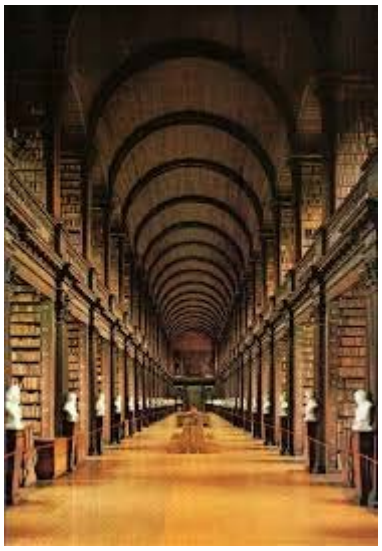
BBQ on Monday:

At the Pavillion Bar, marked on the map above. Usually called “The Pav” by the locals.

Reception at the Long Room Library & the Book of Kells

The Book of Kells Exhibition is a must-see on the itinerary of all visitors to Dublin. Located in the heart of Dublin City, a walk through the cobbled stones of Trinity College Dublin will bring visitors back to the 18th century, when the magnificent Old Library building was constructed and which displays the Book of Kells.

The main chamber of the Old Library is the Long Room; at nearly 65 metres in length, it is filled with 200,000 of the Library's oldest books and is one of the most impressive libraries in the world.



When built (between 1712 and 1732) it had a flat plaster ceiling and shelving for books was on the lower level only, with an open gallery. By the 1850s these shelves had become completely full; largely as since 1801 the Library had been given the right to claim a free copy of every book published in Britain and Ireland. In 1860 the roof was raised to allow construction of the present barrel-vaulted ceiling and upper gallery bookcases.

Marble busts line the Long Room, a collection that began in 1743 when 14 busts were commissioned from sculptor Peter Scheemakers. The busts are of the great philosophers and writers of the western world and also of men (and yes, they are all men) connected with Trinity College Dublin - famous and not so famous. The finest bust in the collection is of the writer Jonathan Swift by Louis Francois Roubiliac.

Other treasures in the Long Room include one of the few remaining copies of the 1916 Proclamation of the Irish Republic which was read outside the General Post Office on 24 April 1916 by Patrick Pearse at the start of the Easter Rising. The harp is the oldest of its kind in Ireland and probably dates from the 15th century. It is made of oak and willow with 29 brass strings. It is the model for the emblem of Ireland.

BBQ at the Pavillion Bar in TCD

The Pavilion Bar first opened in October 1961 and is the sports bar in Trinity College. Profits from the Pavilion go directly to support sport clubs through DUCAC. Having drinks on a sunny day outside “The Pav” as it is better known, is a great way to spend a sunny afternoon, or evening. Many students (and returning graduates) have spent evenings here watching the cricket on College Park, noting that they must one day learn the rules!

Keynote 1

Children's Speech Recognition – from the Lab to the Living Room

Patricia Scanlon, Soapbox Labs **Monday 13.15-14.15**

Voice is predicted to replace typing, clicks, touch and gesture as the dominant way to interface with technology in all aspects of our lives across homes, cars, office, schools. However, voice interfaces designed and built for adults using adult speech data do not perform well for children and performance deteriorates the younger the child. This is due to the fact that children's voices differ from adults both physically and behaviours and behaviourally, these differences increase the younger the child.

Deep learning approaches to speech recognition have increased performance in recent years but require significant volumes of data to achieve such improvements. While large volumes of varied adult speech data-sets are publicly available to buy or license, in sharp contrast, only small limited children's speech data-sets are available. Children's speech is notoriously difficult to collect, particularly under 8 years old. Publicly available children's speech datasets are typically recorded in clean, quiet, high quality headset microphones and in highly controlled conditions. This causes significant problems, as speech technology systems built with such data require that a child's environment mimic conditions in order for the system to work effectively.

SoapBox Labs have been working, on the problem of children's speech recognition, since 2013, and have built a children's voice technology platform for children aged 4-12, which is licensed to third parties to voice-enable their products for use with children. Our high accuracy platform and uses deep learning techniques and has been built using thousands our hours of proprietary, high-quality, real-world, uncontrolled and varied speech data from young children across the globe. SoapBox Labs is currently scaling our platform to multiple new languages.

Application areas include voice control and conversational engagement for home assistants skills, gaming, AR/VR, toys, robotics as well as educational assessment for reading and language learning tutors/assistants.

Globally there is a growing concern about data privacy. Scrutiny is likely to continue and there will be further focus on children's voice data. SoapBox Labs also helps companies take a proactive approach to privacy for children's voice data and ensure full US COPPA and EU GDPR compliance with our patent-pending privacy-by-design approach.

About the Speaker:

With over 20 years' experience in artificial intelligence and speech recognition technology, Dr. Patricia Scanlon has spent the majority of her career working on the commercialisation of research innovations. Currently founder and CEO of SoapBox Labs, building voice technology for kids, which has raised over €3 million in funding to date. Scanlon holds a PhD in Speech Recognition technology and Machine Learning systems and has held positions with Columbia University in New York, IBM T.J. Watson Research Centre, Trinity College Dublin and Nokia Bell Labs.

Keynote 2

Spoken Language Processing: Are We Nearly There Yet?

Roger Moore, Sheffield University **Tuesday 9.00-10.00**

Maybe, maybe not!

About the Speaker:

Prof. Moore (http://www.dcs.shef.ac.uk/~roger) has over 40 years' experience in Speech Technology R&D and, although an engineer by training, much of his research has been based on insights from human speech perception and production. As Head of the UK Government's Speech Research Unit from 1985 to 1999, he was responsible for the development of the Aurix range of speech technology products and the subsequent formation of 20/20 Speech Ltd. Since 2004 he has been Professor of Spoken Language Processing at the University of Sheffield, and also holds Visiting Chairs at Bristol Robotics Laboratory and University College London Psychology & Language Sciences. He was President of the European/International Speech Communication Association from 1997 to 2001, General Chair for INTERSPEECH-2009 and ISCA Distinguished Lecturer during 2014-15. Prof. Moore is the current Editor-in-Chief of Computer Speech & Language and he was recently awarded the 2016 LREC Antonio Zampoli Prize for "Outstanding Contributions to the Advancement of Language Resources & Language Technology Evaluation within Human Language Technologies".

Keynote 3

Deep Learning for End-to-End Audio-Visual Speech Recognition

Stavros Petridis, Imperial College London **Tuesday 13.30-14.15**

Decades of research in acoustic speech recognition have led to systems that we use in our everyday life. However, even the most advanced speech recognition systems fail in the presence of noise, e.g., giving voice commands to your mobile phone in the street does not work so well as in a quiet room. This problem can be (partially) addressed by using visual information, e.g., monitoring the lip movements which are not affected by the noise. Recent advances in deep learning have made it straightforward to extract information from the mouth region and combine it naturally with the acoustic signal in order to enhance the performance of speech recognition. In this talk, we will see how deep learning has made this possible and also present a few relevant applications like end-to-end speech-driven facial animation

About the Speaker:

Stavros is a Research Fellow at the intelligent behaviour understanding group (iBUG) at Imperial College London working on multimodal recognition of human behaviour. He studied electrical and computer engineering at the Aristotle University of Thessaloniki, Greece and completed the MSc degree in Advanced Computing at Imperial College London. He also did his PhD in Computer Science at the same university. Stavros has been a visiting researcher at the image processing group at University College London, at the Robotics Institute, Carnegie Mellon University and at the affect analysis group at the University of Pittsburgh. He has worked in a wide range of human behaviour understanding problems like emotion recognition, age / gender recognition, nonlinguistic vocalisation (e.g. laughter) recognition, native speaker recognition and face re-identification. He is currently working ondeep learning models for audio-visual fusion, audio-visual speech recognition and facial expression recognition.

Oral Session A - Monday

Korin Richmond, University of Edinburgh	**16.00-16.20**
Seeing speech: ultrasound imaging for child speech therapy.	
Matthew Roddy, Trinity College Dublin	**16.20-16.40**
Multimodal Continuous Turn-Taking Prediction Using Multiscale RNNs	
Christos Christodoulopoulos, , Amazon	**16.40-17.00**
Expanding Alexa's Knowledge Base: Relation extraction from unstructured text	

Seeing speech: ultrasound imaging for child speech therapy

Korin Richmond, University of Edinburgh

Abstract

It is estimated up to 6.5% of children (or two children in every classroom) in Britain suffer from a Speech Sound Disorder, defined as difficulty in producing one or more native language speech sounds. This can make it difficult for children to communicate normally, impacting self-esteem and leading to recognised risk of poor integration and educational attainment. Current speech therapy methods have little technological support, relying upon therapist and child client "ears". Attractively, a medical ultrasound scanner offers the potential to visualise and monitor what is going on inside the client's mouth. Here I will give an overview of our "Ultrax" project and its ongoing work to apply machine learning and signal processing techniques to develop ultrasound imaging as a useful technology to support child speech therapy (http://www.ultrax-speech.org).

Multimodal Continuous Turn-Taking Prediction Using Multiscale RNNs

Matthew Roddy[1]*, Gabriel Skantze*[2]*, Naomi Harte*[1]

[1]ADAPT Centre, School of Engineering, Trinity College Dublin, Ireland
[2]Department of Speech Music and Hearing, KTH, Stockholm, Sweden

roddym@tcd.ie

Abstract

In human conversational interactions, turn-taking exchanges can be coordinated using cues from multiple modalities (e.g. linguistic, acoustic, visual). To design spoken dialog systems (SDSs) that can conduct fluid interactions it is desirable to incorporate cues from these separate modalities into turn-taking models. We propose that there is an appropriate temporal granularity at which modalities should be modeled. To test this hypothesis, we develop a multiscale RNN architecture in which modalities can be flexibly modeled at separate timescales. Our results show that when using acoustic and linguistic features to model turn-taking, modeling linguistic features at a variable temporal rate yields significant improvements over using a uniform temporal rate.

1. Method

Our approach builds on the work of Skantze in [1] where turn-taking is modeled in a continuous manner using a single LSTM neural network. The main objective of continuous turn-taking models is to predict speech probability scores for discrete frames within a future window. Speech features are extracted from speakers in a dyadic conversation on a frame-by-frame basis. These features are concatenated into vectors that are fed into an LSTM at each timestep of 50ms. The LSTM is trained to output a vector of speech activity predictions for 60 future timesteps (3 seconds). These continuous predictions represent generalized turn-taking behaviors observed in the training data and can be applied to make decisions that are not just limited to end-of-turn detection. This continuous approach contrasts with more traditional end-of-turn models, where decisions are made at utterance end-points, and are limited in their ability to model naturalistic turn-taking behavior that often involves fast turn-switches and overlapping speech.

A drawback to the approach proposed in [1] is that all the features must be input to the LSTM at a fixed temporal rate. Therefore long-term dependencies that exist in linguistic features are lost due to the vanishing gradient problem. Our multiscale architecture (shown in Fig. 1) addresses this problem by using separate LSTMs for each modality. The modalities are then fused in a master LSTM that operates at the prediction rate (50ms). This allows the modalities to be treated at flexible timescales. In particular this allows the lingustic LSTM to only be updated when there is a new word. Using this approach we achieve significant improvements over the state of the art reported on the HCRC Map-Task Corpus.

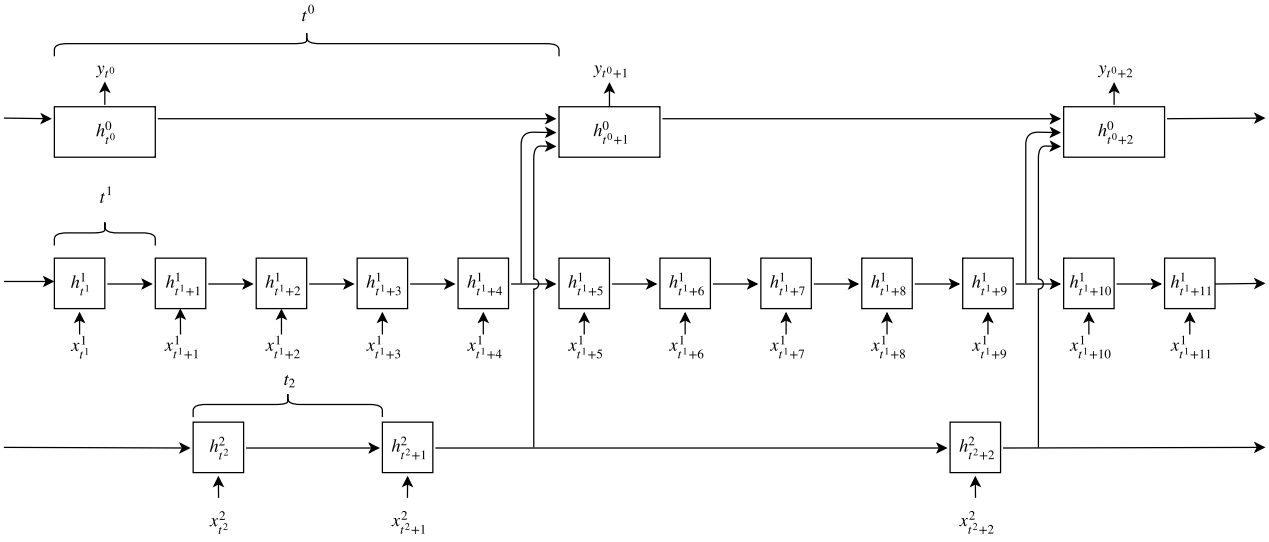


Figure 1: *Multiscale RNN architecture*

2. References

[1] G. Skantze, "Towards a General, Continuous Model of Turn-taking in Spoken Dialogue using LSTM Recurrent Neural Networks," in *Proceedings of SigDial*, Saarbrucken, Germany, 2017.

Expanding Alexa's Knowledge Base: Relation extraction from unstructured text

Christos Christodoulopoulos, Amazon UK

Abstract

These days, most general knowledge question-answering systems rely on large-scale knowledge bases comprising billions of facts about millions of entities. Having a structured source of semantic knowledge means that we can answer questions involving single static facts (e.g. "Who was the 8th president of the US?") or dynamically generated ones (e.g. "How old is Donald Trump?). More importantly, we can answer questions involving multiple inference steps ("Is the queen older than the president of the US?").

In this talk, I'm going to be discussing some of the unique challenges that are involved with building and maintaining a consistent knowledge base for Alexa, extending it with new facts and using it to serve answers in multiple languages. I will focus on an investigation into fact extraction from unstructured text. I will present a method for creating distant (weak) supervision labels for training a large-scale relation extraction system. I will also discuss the effectiveness of neural network approaches by decoupling the model architecture from the feature design of a state-of-the-art neural network system. Surprisingly, a much simpler classifier trained on similar features performs on par with the highly complex neural network system (at 75x reduction to the training time), suggesting that the features are a bigger contributor to the final performance.

Oral Session B - Monday

Benjamin R. Cowan, University College Dublin **11.30-11.50**

"What can I help you with?": Infrequent users' experiences of Intelligent Personal Assistants

Danny Websdale, University of East Anglia **11.50-12.10**

The Effect of Real-Time Constraints on Automatic Speech Animation

Konstantinos Kyriakopoulos, University of Cambridge **12.10-12.30**

Deep learning for assessing non-native pronunciation of English using phone distances

Title: “What can I help you with?”: Infrequent users’ experiences of Intelligent Personal Assistants

Names of authors:
Benjamin R. Cowan (University College Dublin)
Nadia Pantidi (University College Cork)
David Coyle (University College Dublin)
Kellie Morrissey (Newcastle University)
Peter Clarke (University College Dublin)
Sara Al-Shehri (University College Dublin)
David Earley (University College Dublin)
Natasha Bandeira (University College Dublin)

Abstract:
Intelligent Personal Assistants (IPAs) are widely available on devices such as smartphones. However, most people do not use them regularly. Previous research has studied the experiences of frequent IPA users. Using qualitative methods we explore the experience of infrequent users: people who have tried IPAs, but choose not to use them regularly. Unsurprisingly infrequent users share some of the experiences of frequent users, e.g. frustration at limitations on fully hands-free interaction. Significant points of contrast and previously unidentified concerns also emerge. Cultural norms and social embarrassment take on added significance for infrequent users. Humanness of IPAs sparked comparisons with human assistants, juxtaposing their limitations. Most importantly, significant concerns emerged around privacy, monetization, data permanency and transparency. Drawing on these findings we discuss key challenges, including: designing for interruptability; reconsideration of the human metaphor; issues of trust and data ownership. Addressing these challenges may lead to more widespread IPA use.

Full paper presented at ACM SIGCHI Mobile HCI 2017 and published in ACM Digital Library. Paper available at https://dl.acm.org/citation.cfm?id=3098539

The Effect of Real-Time Constraints on Automatic Speech Animation

Danny Websdale, Sarah Taylor and Ben Milner

University of East Anglia, United Kingdom

d.websdale@uea.ac.uk, s.l.taylor@uea.ac.uk, b.milner@uea.ac.uk

Machine learning has previously been applied successfully to speech-driven facial animation. To account for carry-over and anticipatory coarticulation a common approach is to predict the facial pose using a symmetric window of acoustic speech that includes both past and future context. Using future context limits this approach for animating the faces of characters in real-time and networked applications, such as online gaming. An acceptable latency for conversational speech is 200ms and typically network transmission times will consume a significant part of this. Consequently, we consider asymmetric windows by investigating the extent to which decreasing the future context effects the quality of predicted animation using both deep neural networks (DNNs) and bi-directional LSTM recurrent neural networks (BiLSTMs). Specifically we investigate future contexts from 170ms (fully-symmetric) to 0ms (fully-asymmetric), with a fixed past context of 170ms.

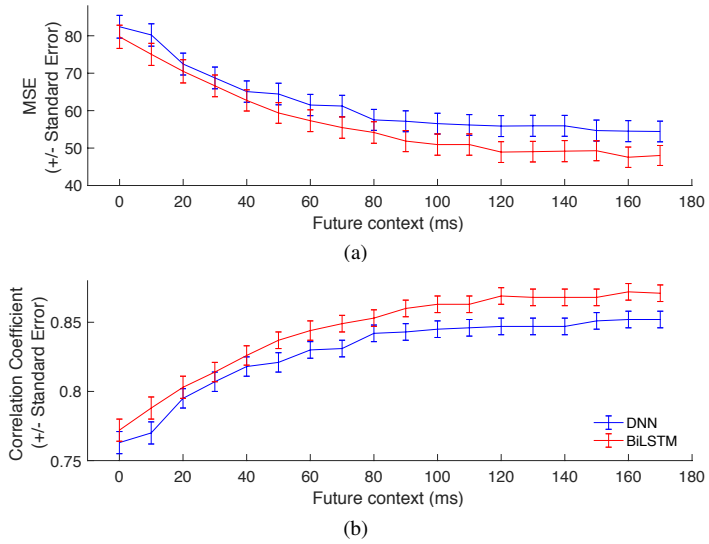


Figure 1: *a) Mean squared error and b) correlation of predicted and ground truth visual features for increasing future contexts.*

Figure 1 shows the MSE (Figure 1(a)) and correlation coefficient (Figure 1(b)) between predicted and ground truth visual features for a DNN (blue) and BiLSTM (red) architecture when increasing the amount of future acoustic context (with standard error bars). We find that a BiLSTM trained using 70ms of future context is able to predict facial motion of equivalent quality as a DNN trained with 170ms, while introducing increased processing time of only 5ms resulting in a 95ms reduction. The results also show that, for a given future context, the BiLSTM is substantially more effective in predicting visual features than the DNN. Subjective tests using the BiLSTM show that reducing the future context from 170ms to 50ms does not significantly decrease perceived realism. Below 50ms, the perceived realism begins to deteriorate, generating a trade-off between realism and latency.

Deep learning for assessing non-native pronunciation of English using phone distances

Konstantinos Kyriakopoulos, Kate M. Knill, Mark J.F. Gales

{kk492, kate.knill, mjfg}@eng.cam.ac.uk

1. Abstract

The way a non-native speaker pronounces the phones of a language is an important predictor of their proficiency. In grading spontaneous speech, the pairwise distances between generative statistical models trained on each phone have been shown to be powerful features (Figure 1). In the system used here as baseline, generative Gaussian models are trained for the pronunciation of each phone and the K-L divergences between them are used to determine the speaker's accent quality.

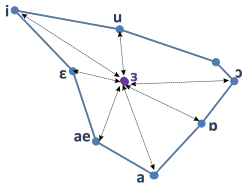


Figure 1: *Illustration of the phone distance concept*

This paper presents a deep learning alternative to model-based phone distances in the form of a tunable Siamese network feature extractor to extract distance metrics directly from the audio frame sequence. Features are extracted at the phone instance level and combined to phone-level representations using an attention mechanism (Figure 2).

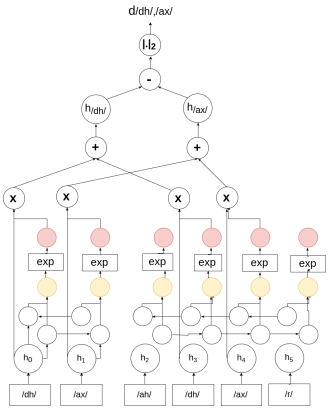


Figure 2: *Illustration of the phone distance concept*

Pair-wise distances between phone features are then projected through a feed-forward layer to predict score. The extraction stage is initialised on either a binary phone instance-pair classification task, or to mimic the model-based features, then the whole system is fine-tuned end-to-end, optimising the learning of the distance metric to the score prediction task. This method is therefore more adaptable and more sensitive to phone instance level phenomena. Its performance is compared against a DNN trained on Gaussian phone model distance features (Table 1). The system outperforms the baseline in terms of mean squared error (MSE) but is comparable for Pearson's correlation coefficient (PCC). This can be explained given that the new system is optimised end-to-end for minimum MSE, whereas for the baseline only the grader is optimised for minimum MSE.

Initialisation	MSE	PCC
Baseline	14.8	0.785
End-to-end system	14.2	0.780

Table 1: *Performance (MSE and PCC of predicted to human-assigned scores) of baseline and Siamese systems, trained using binary and K-L divergence criteria, each trained on TRN and evaluated on EVL*

Index Terms: pronunciation assessment, phone distances, CALL, CAPT, Siamese Networks, attention mechanism

Poster Session A

1	Salil Deena, Raymond W. M. Ng, Pranava Madhyastha, Lucia Specia and Thomas Hain	Exploring the use of Acoustic Embeddings in Neural Machine Translation	University of Sheffield
2	Zack Hodari, Oliver Watts, Srikanth Ronanki, Simon King	Learning interpretable control dimensions for speech synthesis by using external data	University of Edinburgh
3	Christopher G. Buchanan, Matthew P. Aylett, David A. Braude	Adding Personality to Neutral Speech Synthesis Voices	CereProc Ltd., Edinburgh
4	Avashna Govender and Simon King	Using pupillometry to measure the listening effort of synthetic speech	University of Edinburgh
5	Carolina De Pasquale, Charlie Cullen, Brian Vaughan	Towards a protocol for the analysis of interpersonal rapport in clinical interviews through speech prosody	Dublin Institute of Technology
6	Yasufumi Moriya, Gareth. J. F. Jones	Investigating the use of a Multimodal Language Model for Re-Ranking ASR N-best Hypotheses	Dublin City University
7	Jennifer Williams and Simon King	Low-Level Prosody Control From Lossy F0 Quantization	University of Edinburgh
8	Andy Murphy, Irena Yanushevskaya, Christer Gobl, Ailbhe Ní Chasaide	Effects of voice source manipulation on prominence perception	Trinity College Dublin
9	Benjamin R Cowan, Holly P. Branigan, Habiba Begum, Lucy McKenna, Eva Szekely	They Know as Much as We Do: Knowledge Estimation and Partner Modelling of Artificial Partners	University College Dublin
10	Chao Zhang & Phil Woodland	HIGH ORDER RECURRENT NEURAL NETWORKS FOR ACOUSTIC MODELLING	University of Cambridge
11	Xizi Wei, Peter Jančovič, Martin Russell, Khalida Ismail, Tom Marshall	Automatic Assessment of Motivational Interviews with Diabetes Patients	University of Birmingham
12	K.M. Knill, M.J.F. Gales, K. Kyriakopoulos, A. Malinin, A. Ragni, Y. Wang, A.P.Caines	Impact of ASR Performance on Free Speaking Language Assessment	University of Cambridge
13	Carol Chermaz , Cassia Valentini-Botinhao , Henning Schepker and Simon King	Speech pre-enhancement in realistic environments	University of Edinburgh
14	Wissam Jassim and Naomi Harte	Voice Activity Detection Using Neurograms	Trinity College Dublin

Exploring the use of Acoustic Embeddings in Neural Machine Translation

Salil Deena[1]*, Raymond W. M. Ng*[1]*, Pranava Madhyastha*[2]*, Lucia Specia*[2] *and Thomas Hain*[1]

[1]Speech and Hearing Research Group, The University of Sheffield, UK
[2]Natural Language Processing Research Group, The University of Sheffield, UK
{s.deena, wm.ng, p.madhyastha, l.specia, t.hain}@sheffield.ac.uk

1. Abstract

In Neural Machine Translation (NMT) [1], text from a source language is first encoded using a recurrent neural network (RNN), resulting in compressed context vector, which is then passed to the decoder, also a RNN, and takes the encoded context vector and the previously translated word as input and produces the target translated word at the current time step. The compressed context vector is derived by applying an attention mechanism, which is a measure of alignment between the source and target text, to the RNN hidden state vectors of the encoder up to the current time-step.

Auxiliary features can be integrated at the encoder by concatenating the word vectors with features. Linguistic input features such as lemmas were found to improve NMT results when they are appended to the word vector at the encoder [2] or even when added as an extra output at the decoder. In [3], latent Dirichlet allocation (LDA) [4] topic vectors were appended to the hidden state vector for each word and subsequently used to obtain a topic-informed encoder context vector, which is then passed to the decoder.

This work focuses on the integration of auxiliary features extracted from audio accompanying the text. Whilst features extracted from text and images have been explored, the use of audio information for NMT remains an open question. Audio features in the form of show-level i-vectors [5] and Latent Dirichlet Allocation (LDA) topic vectors extracted from audio (acoustic LDA) [6] are explored for machine translation (MT) of source text. These auxiliary features are compared and combined with show-level LDA topic vectors derived from text [4] as well as word embeddings that preserve distance of similar words in vector space [7]. The composition of features at different levels of granularity (show-level and word-level) is also investigated. The features used are listed below:

Feature	Type	Domain	Granularity
Word2Vec	Text	Google pre-trained (out-of-domain)	Word (token) level
Text LDA	Text	In-domain	Document(show) level
Acoustic LDA	Acoustic	In-domain	Document (show) level
i-Vector	Acoustic	In-domain	Document (show) level

The features are investigated on the translation of TED talks transcripts from English to French [8] using NMT augmented with the acoustic information, derived from corresponding audio, and the results are given below:

Model	**TEDdev**		**TEDeval**	
	BLEU	METEOR	BLEU	METEOR
Baseline	30.38	**0.6158**	36.02	0.6485
Word2Vec (300d)	**30.44**	0.6116	35.89	0.6424
i-vector (50d)	29.97	0.6118	35.87	0.6455
i-vector (100d)	29.77	0.6065	36.14	0.6428
tLDA (50d)	30.12	0.6092	36.09	0.6432
tLDA (100d)	30.12	0.6126	36.14	0.6449
aLDA (50d)	30.32	0.6118	36.11	**0.6506**
aLDA (100d)	29.93	0.6125	**36.51**	0.6474

(a) Results on TED data in NMT setting

Model	**TEDdev**		**TEDeval**	
	BLEU	METEOR	BLEU	METEOR
Baseline	30.38	**0.6158**	36.02	0.6485
Word2Vec (300d)	30.44	0.6116	35.89	0.6424
Word2Vec+i-vector (50d)	30.09	0.6146	36.15	0.6499
Word2Vec+i-vector (100d)	30.20	0.6105	36.73	0.6524
Word2Vec+tLDA (50d)	30.38	0.6128	36.23	0.6482
Word2Vec+tLDA (100d)	**30.57**	0.6123	36.27	0.6479
Word2Vec+aLDA (50d)	30.50	0.6087	36.04	0.6463
Word2Vec+aLDA (100d)	30.16	0.6140	**37.21**	**0.6525**

(b) Results on TED data in compositional NMT setting

2. References

[1] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *ICLR'15: Proc. of the International Conference on Learning Representations*, 2015.

[2] R. Sennrich and B. Haddow, "Linguistic input features improve neural machine translation," in *WMT'16: Proceedings of the First Conference on Machine Translation*, 2016, pp. 83–91.

[3] J. Zhang, L. Li, A. Way, and Q. Liu, "Topic-informed neural machine translation," in *COLING'16: Proc. of the 26th International Conference on Computational Linguistics*, 2016, pp. 1807–1817.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[5] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[6] M. Doulaty, O. Saz, and T. Hain, "Unsupervised domain discovery using latent dirichlet allocation for acoustic modelling in speech recognition," in *INTERSPEECH'15: Proc. of the 15th Annual Conference of the International Speech Communication Association*, 2015.

[7] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *ICLR Workshop*, 2013.

[8] I. 2015, "Web inventory of transcribed and translated talks," Jan 2015. [Online]. Available: https://wit3.fbk.eu/mt.php?release=2015-01

Learning interpretable control dimensions for speech synthesis by using external data

Zack Hodari, Oliver Watts, Srikanth Ronanki, Simon King

The Centre for Speech Technology Research, University of Edinburgh, United Kingdom
{zack.hodari, oliver.watts, srikanth.ronanki, simon.king}@ed.ac.uk

1 Abstract

There are many aspects of speech that we might want to control when creating text-to-speech (TTS) systems. We present a general method that enables control of arbitrary aspects of speech, which we demonstrate on the task of emotion control. Current TTS systems use supervised machine learning and are therefore heavily reliant on labelled data. If no labels are available for a desired control dimension, then creating interpretable control becomes challenging. We introduce a method that uses external, labelled data (i.e. not the original data used to train the acoustic model) to enable the control of dimensions that are not labelled in the original data. Adding interpretable control allows the voice to be manually controlled to produce more engaging speech, for applications such as audiobooks. We evaluate our method using a listening test.

Adding Personality to Neutral Speech Synthesis Voices

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

CereProc Ltd., Edinburgh, UK

{chrisb,matthewa,dave}@cereproc.com

Abstract

A synthetic voice personifies the system using it. Previous work has shown that using sub-corpora with different voice qualities (e.g. tense and lax) can be used to modify the perceived personality of a voice as well as adding expressive and emotional functionality. In this work we explore the use of LPC source/filter decomposition together with modification of the residual, to artificially add voice quality sub-corpora to a voice without recording bespoke data. We evaluate this artificially enhanced voice against a baseline unit selection voice with pre-recorded sub-corpora. Although artificial modification impacts naturalness, it has the advantage of adding emotional range to voices where none was recorded in the source data, deals with data sparsity issues caused by sub-corpora, and results in significant effects in terms of perceived emotion.

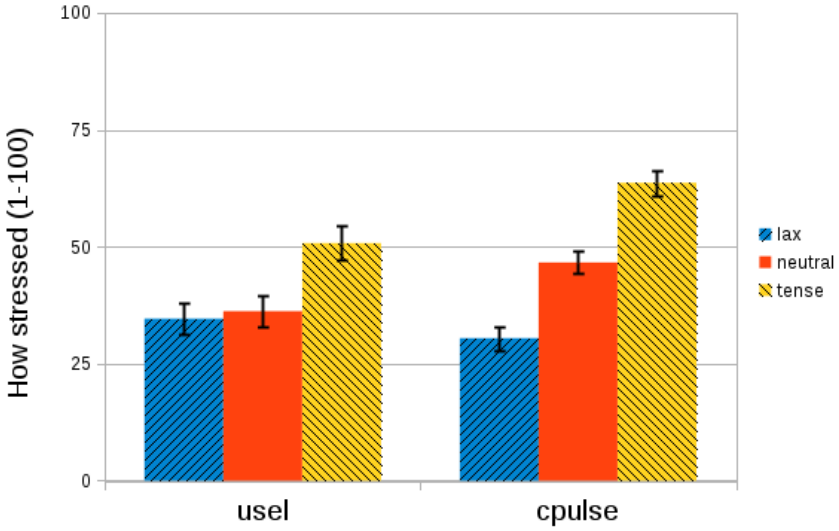


Figure 1: *Perceived emotional stress of synthesised tense, neutral and lax audio stimuli created using voices with natural (*usel*) and artificially created (*cpulse*) tense and lax sub-corpora. Error bars show standard error.*

We invite the reader to listen to two versions of the Arctic STL voice, the first as a standard unit selection voice `https://tinyurl.com/y8p35b5a`, and the second with artificial voice quality sub-corpora added `https://tinyurl.com/y9u9bhf8`. This voice is released free to use as part of ARIA-VALUSPA AVP package[1]. Furthermore the algorithmic approach resulted in a better discrimination between tense, neutral and lax stimuli in terms of perceived emotional stress.

[1] `https://github.com/ARIA-VALUSPA/AVP`

Using pupillometry to measure the listening effort of synthetic speech

Avashna Govender and Simon King

The Centre for Speech Technology Research, University of Edinburgh, United Kingdom

a.govender@sms.ed.ac.uk, Simon.King@ed.ac.uk

1. Abstract

Currently synthetic speech is evaluated through subjective listening tests such as the Blizzard challenge [1]. In these tests, intelligibility is measured by asking a listener to transcribe the words heard or by rating the naturalness on a 5 point scale. Whilst this type of measurement is useful for understanding the systems performance in these areas, what they don't provide is insight on where improvements need to be made within the system. Therefore, there is a need for better evaluation methods. With increased use of Text-to-speech (TTS) in real-world applications, the cognitive load required to understand synthetic speech may be a more appropriate measure. This is motivated by the use of TTS applications in situations where the cognitive load is sufficiently high, which could lead to implications that may be harmful to the user. For example, if listening to TTS compared to natural speech is more distracting, this could be dangerous for a vehicle driver listening to a navigation system or if the effort required to listen to a TTS generated audio-book is high, this could quickly lead to fatigue. Cognitive load has been investigated in the past when rule-based speech synthesizers were popular, but with the recent advancements made in TTS with Deep Neural Networks [2], the speech quality of these systems have drastically improved. To our knowledge, there is little or no recent work evaluating the cognitive load of state-of-the-art text-to-speech systems.

Pupillometry has become popular across many fields of research which include language processing, speech production and visual perception. Studies have consistently shown that there is a correlation between pupil dilation and the mental effort required to carry out a specific task [3, 4]. More recently, the pupil response is used as an index of listening effort which is defined as the amount of mental effort exerted to perform a listening task [5, 6].

An initial study was conducted to determine the feasibility of using pupillometry to measure the listening effort of synthetic speech. Stimuli generated by four speech synthesizers and the human voice used to train them (taken from the Blizzard 2011 Challenge) were evaluated. Our results in Figure 1 show that the pupil dilation is sensitive to the quality of synthetic speech. As expected, results also reflect that synthetic speech imposes a higher cognitive load than natural speech. Whilst these results show differences between various speech synthesizers and natural speech, what may be more interesting is taking this one step further in determining whether differences can be observed within a single system. In this way, more meaningful insight could be gained in pinpointing where the system suffers.

Currently we are in the process of evaluating stimuli generated by a DNN TTS synthesizer. The idea is to generate stimuli that creates a continuum of stepping from natural speech to synthesized speech by changing only a single feature at a time as shown in Table 1.

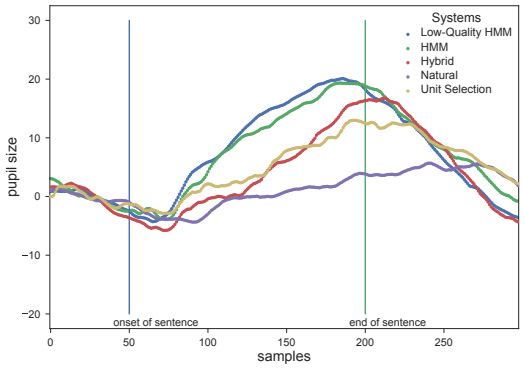


Figure 1: *Pupil response for 4 speech synthesizers from Blizzard Challenge 2011 and the corresponding natural speech for news sentences (50 samples/sec)*

Table 1: *Stimuli generation*

System Name	Mel-cepstral coefficients	F0	Duration
A	Vocoded	Vocoded	Natural
B	Vocoded	Predicted	Natural
C	Predicted	Vocoded	Natural
D	Predicted	Predicted	Natural
E	Predicted	Predicted	Predicted

2. References

[1] S. King, L. Wihlborg, and W. Guo, "The Blizzard Challenge 2017," Stockholm,Sweden, Sept. 2017.

[2] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference*. IEEE, 2013, pp. 7962–7966.

[3] J. Beatty and D. Kahneman, "Pupillary changes in two memory tasks," *Psychonomic Science*, vol. 5, no. 10, pp. 371–372, 1966.

[4] D. Kahnemann and J. Beatty, "Pupillary responses in a pitch-discrimination task," *Attention, Perception, & Psychophysics*, vol. 2, no. 3, pp. 101–105, 1967.

[5] R. McGarrigle, K. J. Munro, P. Dawes, A. J. Stewart, D. R. Moore, J. G. Barry, and S. Amitay, "Listening effort and fatigue: What exactly are we measuring? a British Society of Audiology Cognition in Hearing Special Interest Group 'white paper'," *International journal of audiology*, vol. 53, pp. 443–440, 2014.

[6] A. A. Zekveld, S. E. Kramer, and J. M. Festen, "Cognitive load during speech perception in noise: the influence of age, hearing loss, and cognition on the pupil response," *Ear and hearing*, vol. 32, no. 4, pp. 498–510, 2011.

Towards a protocol for the analysis of interpersonal rapport in clinical interviews through speech prosody.

Carolina De Pasquale[1]*, Charlie Cullen*[2]*, Brian Vaughan*[1]

[1]Dublin Institute of Technology, Ireland [2]University of the West of Scotland, United Kingdom

carolina.depasquale@dit.ie

Abstract

Finding objective measures to assess the quality of a clinical interaction could greatly improve therapy training and evaluation of the interactions. Socio-behavioural signals can provide objective measures to corroborate or even replace expert appraisal: several studies have explored the validity of speech behavioural signals as measures to evaluate the quality of interactions, finding that higher degrees of coordination (also referred to as entrainment, synchrony, and so on) often correlate with higher perceived empathy, higher therapeutic alliance ratings, and better outcomes. However, there is no standard protocol for the acquisition and analysis of acoustic prosodic measures in clinical interactions.This contribution suggests a methodological approach to the analysis of interpersonal prosody in therapeutic interactions as a measure of therapeutic alliance.

Equipment choice and recording set up can affect the audio quality, and choices should be made so as not to compromise the authenticity of the interaction. Professional recording equipment yields better quality audio, but it is expensive and can be obtrusive; sufficient audio quality can be obtained with lesser quality equipment such as smartphones [1]. Naturalistic settings yield audio that needs more preparation before analysis, but have the advantage of not colouring the interaction, while a professional recording studio might affect interpersonal behaviour, especially in a clinical interaction; with proper arrangements, a quiet room in the clinician's place of practice is an ideal location for the recordings.

Naturalistic audio needs to be pre-processed to be separated: this contribution presents an annotation schema that allows to adequately prepare audio for pre-processing by labelling cross-talk, overlaps, noise, and other acoustic events. The annotation schema presented in this contribution was successfully used in an empirical study and is proposed as an extension to existing schemes such as the IMDI. This process can be automated further by using speaker separation/diarization tools, or can be performed manually by researchers with no coding experience, though that is not advisable due to the long processing time that it requires and propensity for human error.

Acoustic features can be extracted with a variety of software (such as Praat or Matlab), each requiring different levels of proficiency: the process can be almost entirely automated, which requires considerable domain knowledge or access to expensive tools, or can be performed in a manual/supervised way, which is time consuming but requires less technical expertise. A time aligned moving average window can be used to time align prosodic features in conversations, which by nature are turn based instead of synchronous; a moving correlation window shows accommodation dynamics throughout the conversation.

The guidelines suggested were followed during a pilot study in a psychiatric setting, with the aim of investigating prosodic accommodation as a way to help with depression diagnosis [2]. Recordings were performed in the psychiatrists office with a Zoom H4N external recorder and Lavalier lapel microphones. Each recording was annotated using Textgrids in PRAAT following the schema proposed and subsequently exported as separate audio files, which were then used for feature extraction.

A number of prosodic features were extracted and analysed through Praat, with an automated script. The features extracted were: pitch (mean and median f_0), speech rate (syllable nuclei per second), intensity, vowel space (calculated by tracking the first and second formants and drawing the frequency region), spectral energy and slope. These features can be considered a good starting point for the investigation of acoustic prosody, and they are easily extracted through speech processing tools.

The analysis focused on f_0 as a measure of prosodic accommodation and vowel space ratio, which Scherer et al.[3] indicate as a robust measure of psychological distress. All the conversations were positively correlated; more pertinent, however, is that the dynamic nature of prosodic accommodation was evident from the results (in agreement with the literature [4]). In agreement with [3], vowel space ratio was reduced in the patients' speech; however, analysis also showed significant reduction in the vowel space ratio of the psychiatrist involved in the interactions, which represents a potential indication of adaptive behaviour in response to that of the patient.

This contribution suggests a methodological approach to the analysis of interpersonal prosody in therapeutic interactions as a measure of therapeutic alliance. It suggests guidelines that can be followed with ease even without speech analysis expertise, and presents preliminary results from a pilot study that was performed according to the guidelines, as validation for the robustness of the guidelines.

Index Terms: annotation schema, audio segmentation, speech analysis, clinical interviews, methodology, prosody

1. References

[1] E. U. Grillo, J. N. Brosious, S. L. Sorrell, and S. Anand, "Influence of Smartphones and Software on Acoustic Voice Measures." *International Journal of Telerehabilitation*, vol. 8, no. 2, pp. 9–14, dec 2016.

[2] B. Vaughan, C. De Pasquale, L. Wilson, C. Cullen, and B. Lawlor, "Investigating Prosodic Accommodation in Clinical Interviews with Depressed Patients," in *Proceedings of 7th International MindCare Workshop on Pervasive Computing Paradigms for Mental Health*, Boston, USA.

[3] S. Scherer, L.-P. Morency, J. Gratch, and J. Pestian, "Reduced vowel space is a robust indicator of psychological distress: A cross-corpus analysis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, apr 2015, pp. 4789–4793.

[4] C. De Looze, S. Scherer, B. Vaughan, and N. Campbell, "Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction," *Speech Communication*, vol. 58, pp. 11–34, mar 2014.

Investigating the use of a Multimodal Language Model for Re-Ranking ASR N-best Hypotheses

Yasufumi Moriya, Gareth. J. F. Jones

ADAPT Centre, School of Computing, Dublin City University, Ireland

yasufumi.moriya@adaptcentre.ie, gareth.jones@dcu.ie

Abstract

The primary basis of indexing spoken content retrieval (SCR) systems for multimedia content is often transcripts of the spoken content created using automatic speech recognition (ASR) systems. Due to this dependency, ASR transcription errors can severely degrade the effectiveness of SCR systems. In [1, pp. 1393], it is observed that recognition of content words, especially named-entities, has a significant impact on reliability of SCR systems, whereas mis-recognition of function words often has no impact on the behaviour of these systems. Further, mis-recognition of words that result in semantically incorrect transcripts can lead to worse performance for search tasks.

Integration of visual information into ASR is a potential mechanism to improve the semantic accuracy of ASR transcripts, as information related to semantic context absent from or unclear in the audio stream, can be contained in the associated visual stream. Visual features have already been demonstrated to be useful for various natural language processing tasks, such as image caption generation, parsing, and machine translation. In ASR, recent research has applied a language model, that exploits visual context, to re-ranking of N-best hypotheses generated from an ASR systems [2], [3]. However, in this work the potentially contribution of visually adapted language models (referred to as "multimodal language models") over non-adapted models ("unimodal language models") is not clear, since direct comparison of these models is not provided.

Our work presents analysis of experiments on re-ranking of ASR 30-best hypotheses using a multimodal language model with comparison to a unimodal language model. Our experiments were conducted on two different datasets: (1) online lecture videos of Udacity, and (2) instruction videos of the CMU How-to corpus [2]. Both multimodal and unimodal language models were based on a recurrent neural network (RNN) with long-short term memory (LSTM) units. On training of a multimodal language model, a 4096 dimensional visual embedding was extracted with the VGG19 *fc7* layer from an image corresponding to each utterance. The embedding was fed to a language model before the first word of each utterance. Perplexity and word error rates (WER) are shown in Table 1. As seen from Table 1, the multimodal language model outperformed a unimodal counterpart with a very small margin both in perplexity and WER. Qualitative analysis, however, did not show proof that the multimodal language model was superior to the unimodal one in terms of semantics of output transcripts. Phonetically similar or identical, but semantically different word pairs, were often mis-recognised in the experiments, and the use of the multimodal language model did not help to reduce this type of errors using this method. For the Udacity data, recognition of "ad" and "ads" (often substituted by "add" and "adds") was correct 64 and 58 times using the multimodal, and 66 and 58 times by the unimodal. On the CMU How-to data, similar behaviour was observed for the word pair "hair" and "here". Substitution of "hair" by "here" occurred 7 times using a multimodal language model, and 6 times by a unimodal language model. This is perhaps, because feeding image embeddings is not sufficient for the models to learn semantic relationships between audio and visual streams. In our future work, we plan to explore alternative approaches to the integration of visual information into language models.

Table 1: *Perplexity and WER results on the test set of Udacity and CMU How-to data. "no-rerank" is an ASR system without N-best hypotheses re-ranking. "oracle" is the lowest possible WER achievable from 30-best hypotheses.*

	Udacity		**CMU How-to**	
	Perplexity	WER (%)	Perplexity	WER (%)
no-rerank	N/A	14.43	N/A	20.25
oracle	N/A	7.92	N/A	15.13
unimodal	110.69	12.48	70.29	18.59
multimodal	108.82	12.40	67.81	18.42

References

[1] L. S. Lee, J. Glass, H. Lee, and C. Chan, "Spoken content retrieval – beyond cascading speech recognition with text retrieval," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 9, pp. 1389-1420, 2015.

[2] A. Gupta, Y. Miao, L. Neves, and F. Metze, "Visual features for context-aware speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5020-5024, 2017.

[3] F. Sun, D. Harwath, and J. Glass, "Look, listen, and decode: Multimodal speech recognition with images," in *IEEE Workshop on Spoken Language Technology (SLT)*, pp. 573-578, 2016.

Low-Level Prosody Control From Lossy F0 Quantization

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

Abstract

A single utterance has potentially many different plausible surface realizations of prosody. This is a problem for speech synthesis because prosody can significantly affect intelligibility as well as the perception of naturalness. It is therefore necessary to represent and control prosody in a way that facilitates this one-to-many mapping. Prosody control is a vibrant area of research in speech synthesis (Skerry-Ryan et al., 2018). One persisting challenge is to find a low-level representation of speech pitch ($F0$) which meets the following criteria:

- aids user-oriented control in text-to-speech (TTS)
- enables meaningful automatic $F0$ prediction
- preserves human perceptual salience
- flexible towards synthesis method (hybrid, unit-selection, parametric, end-to-end, etc)

In this work, we adopt a templatic approach to $F0$ representation. It has been shown that $F0$ templates at the syllable-level can be predicted directly from source text (Ronanki et al., 2016). We also build on the idea that $F0$ can be stylized so that it is objectively lossy while at the same time does not compromise human perception (d'Alessandro and Mertens, 1995). We explore the effects of lossy $F0$ parameterization by generating sets of pitch templates and we compare these across natural and synthesized speech. If an effective representation of $F0$ can be identified then it will aide efforts to build expressive TTS. It will also provide a scaffolding for future work that disentangles speaking style from underlying emotion.

References

Christophe d'Alessandro and Piet Mertens. 1995. Automatic Pitch Contour Stylization Using a Model of Tonal Perception. *Computer Speech and Language*, 9(3):257–288.

Srikanth Ronanki, Gustav Eje Henter, Zhizheng Wu, and Simon King. 2016. A Template-Based Approach for Speech Synthesis Intonation Generation Using LSTMs. In *INTERSPEECH*, pages 2463–2467.

RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J Weiss, Rob Clark, and Rif A Saurous. 2018. Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron. *arXiv preprint arXiv:1803.09047*.

Effects of voice source manipulation on prominence perception

Andy Murphy, Irena Yanushevskaya, Christer Gobl, Ailbhe Ní Chasaide

Trinity College Dublin, Ireland

murpha61@tcd.ie, yanushei@tcd.ie, cegobl@tcd.ie, anichsid@tcd.ie

Abstract

This paper explores the effect of voice source modulation on the perception of prominence, continuing from previous analyses of accentuation [1], focus/deaccentuation [2] and declination [3]. Elaborating the role of the voice in prosody is desirable for a more adequate account of the functions of prosody in speech communication. This has many implications in speech technology. For example, our current parallel research developing synthetic voices for Irish dialects (www.abair.ie and [4]) and deploying these in interactive games for language learning requires a more sophisticated modelling of prosody in synthesis than is currently available.

Synthetic stimuli were generated based on a sentence with three accented, prominent syllables (P1, P2, P3, indicated in bold type below) produced by a male speaker of Munster (Kerry) Irish:

Bhí **CÁIT** cupla **LÁ** ar an **TRÁ**laer

[vʲi kɑtʲ kʊplʲə lʲɑ ərˠ ənˠ trˠɑlʲərˠ]

'was Kate couple days on the trawler' (word gloss), 'Kate was a couple of days on the trawler' (translation)

Using inverse filtering, source parameterisation and resynthesis, a 'flattened' baseline version of the utterance was generated, with only slight declination of f_0 and other voice parameters. The global waveshape parameter R_d [5, 6] was modulated to provide (i) source boosting (tenser phonation) on either P1 or P2, and/or (ii) source attenuation (laxer phonation) following (Post-attenuation) or preceding (Pre-attenuation) P1 or P2.

The R_d parameter is derived from f_0, E_e and U_p as follows: $(1/0.11)\times(f_0 \cdot U_p/E_e)$, where E_e is the excitation strength (measured as the negative amplitude of the differentiated glottal flow at the time point of maximum waveform discontinuity) and U_p is the peak flow of the glottal pulse. R_d requires the parameters E_e, U_p and f_0 to be known. Given that for the individual stimuli in this test f_0 was kept constant (other than for the sentence declination), changes to R_d could be implemented with covariation of E_e (with U_p kept constant) or covariation of U_p (with E_e constant). In this experiment both options were tested, and two sets of stimuli (and two baseline sentences) were thus generated.

Twenty nine listeners rated the prominence level of all syllables in the utterance. Results show the E_e-varying stimuli to be more effective in signaling prominence (Figure 1). Phrasal position (P1 vs. P2) makes a large difference to prominence judgements. P1 emerged as overall more prominent and more readily 'enhanced' by the source modifications. Post-attenuation was particularly important for P1, with effects equal to or greater than local P-boosting. In the case of P2, Pre-attenuation was much more important than Post-attenuation.

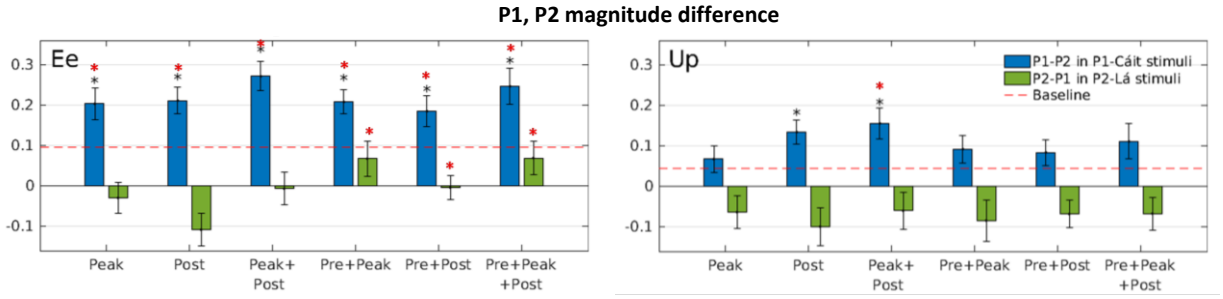


Figure 1: *P1-P2 differences for P1-Cáit target stimuli (blue), and P2-P1 differences for P2- Lá target stimuli (green). Red line = P1-P2 difference in baseline.* = significant difference (* = difference re baseline).*

References

[1] A. Ní Chasaide, I. Yanushevskaya, J. Kane, and C. Gobl, "The Voice Prominence Hypothesis: the interplay of F0 and voice source features in accentuation," in *Interspeech 2013*, Lyon, France, 2013, pp. 3527-3531.

[2] I. Yanushevskaya, C. Gobl, J. Kane, and A. Ní Chasaide, "An exploration of voice source correlates of focus," in *Interspeech 2010*, Makuhari, Japan, 2010, pp. 462-465.

[3] A. Ní Chasaide, I. Yanushevskaya, and C. Gobl, "Prosody of voice: declination, sentence mode and interaction with prominence," in *XVIIIth International Congress of Phonetic Sciences*, Glasgow, UK, 2015, pp. 1-5.

[4] A. Ní Chasaide, N. Ní Chiaráin, C. Wendler, H. Berthelsen, A. Murphy, and C. Gobl, "The ABAIR initiative: bringing spoken Irish into the digital space," in *Interspeech 2017*, Stockholm, Sweden, 2017, pp. 2113-2117.

[5] G. Fant, "The LF-model revisited: transformations and frequency domain analysis," *STL-QPSR,* vol. 2-3, pp. 119-156, 1995.

[6] G. Fant, "The voice source in connected speech," *Speech Communication,* vol. 22, pp. 125-139, 1997.

Title of Poster:
They Know as Much as We Do: Knowledge Estimation and Partner Modelling of Artificial Partners

Names of Authors:
Benjamin R Cowan (University College Dublin)
Holly P. Branigan (University of Edinburgh)
Habiba Begum (University of Birmingham)
Lucy McKenna (ADAPT Centre, Trinity College Dublin)
Eva Szekely (KTH Royal Institute of Technology, Stockholm)

Abstract:
Conversation partners' assumptions about each other's knowledge (their *partner models*) on a subject are important in spoken interaction. However, little is known about what influences our partner models in spoken interactions with artificial partners. In our experiment we asked people to name 15 British landmarks, and estimate their identifiability to a person as well as an automated conversational agent of either British or American origin. Our results show that people's assumptions about what an artificial partner knows are related to their estimates of what other people are likely to know - but they generally estimate artificial partners to have more knowledge in the task than human partners. These findings shed light on the way in which people build partner models of artificial partners. Importantly, they suggest that people use assumptions about what other humans know as a heuristic when assessing an artificial partner's knowledge.

Paper presented at CogSci 2017- paper available at https://mindmodeling.org/cogsci2017/papers/0355/index.html

HIGH ORDER RECURRENT NEURAL NETWORKS FOR ACOUSTIC MODELLING

Chao Zhang & Phil Woodland

Cambridge University Engineering Dept., Trumpington St., Cambridge, CB2 1PZ U.K.
{cz277,pcw}@eng.cam.ac.uk

Vanishing long-term gradients are a major issue in training standard recurrent neural networks (RNNs), which can be alleviated by long short-term memory (LSTM) models with memory cells. However, the extra parameters associated with the memory cells mean an LSTM layer has four times as many parameters as an RNN with the same hidden vector size. This paper addresses the vanishing gradient problem using a high order RNN (HORNN) which has additional connections from multiple previous time steps. Speech recognition experiments using British English multi-genre broadcast (MGB3) data showed that the proposed HORNN architectures for rectified linear unit and sigmoid activation functions reduced word error rates (WER) by 4.2% and 6.3% over the corresponding RNNs, and gave similar WERs to a (projected) LSTM while using only 20%–50% of the recurrent layer parameters and computation. When the savings in parameter number and computation are used to implement wider or deeper recurrent layers, (projected) HORNNs gave a 4% relative reduction in WER over the comparable (projected) LSTMs .

ID	System	D_h	D_p	tg	cn
L_1^{275h}	1L LSTMP	1000	500	26.5	26.0
S_1^{275h}	1L sigmoid HORNNP	1000	500	26.4	25.8
R_1^{275h}	1L ReLU HORNNP	1000	500	26.4	25.9
L_3^{275h}	2L LSTMP	1000	500	25.7	25.2
S_4^{275h}	2L sigmoid HORNNP	1000	500	25.6	25.2
R_4^{275h}	2L ReLU HORNNP	1000	500	25.3	25.0
D_1^{275h}	7L sigmoid DNN	1000		28.4	27.5

Table 1. *3-gram LM %WERs on dev17b for 275h MGB3 systems. LSTMP and HORNNP refer to the projected LSTM and HORNN.*

Automatic Assessment of Motivational Interviews with Diabetes Patients

Xizi Wei[1], *Peter Jančovič*[1], *Martin Russell*[1], *Khalida Ismail*[2], *Tom Marshall*[3]

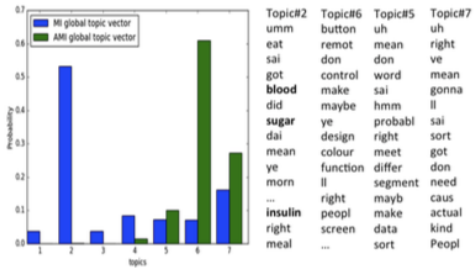
[1]School of Engineering, The University of Birmingham, UK

[2]Institute of Psychiatry, Psychology and Neuroscience, King's College London, UK

[3]Institute of Applied Health Research, The University of Birmingham, UK

{xxw395,p.jancovic,m.j.russell,t.p.marshall}@bham.ac.uk; khalida.2.ismail@kcl.ac.uk

Abstract : The cost of diabetes to the UK in 2016 is estimated to be £14 billion. The impact on the health service can be reduced if patients take ownership of day-to-day monitoring and medication. Motivational Interviewing (MI) is a kind of goal-driven clinical con- version that seeks to achieve this objective. This paper presents initial results on automatic quality assessment of MIs. This is challenging because the speech is conversational, medical terminology is used and recordings are made with a distant micro- phone in natural environments. We describe the development of automatic speech recognition (ASR), speaker diarisation and topic analysis for MIs. We explore approaches to DNN-HMM training and adaptation for ASR using out-of-domain AMI data plus a limited amount of in-domain MI speech. Diarisation, using DNN i-vectors and GMM-based clustering, is used to separate clinician and patient speech, and enables speaker-specific fMLLR training. On a test set of over 45 minutes of MI data, our best ASR and diarisation systems achieve 43.59% word error rate and an F-measure of 0.765, respectively. Finally, we explore the impact of ASR error rate on LDA topic modeling.



Topic#2	Topic#6	Topic#5	Topic#7
umm	button	uh	uh
eat	remot	mean	right
sai	don	don	ve
got	control	word	mean
blood	make	sai	gonna
did	maybe	hmm	ll
sugar	ye	probabl	sai
dai	design	right	sort
mean	colour	meet	got
ye	function	differ	don
morn	ll	segment	need
...	right	mayb	caus
insulin	peopl	make	actual
right	screen	data	kind
meal	...	sort	Peopl

Figure 1: *Results of LDA-based topic modelling. The global distribution of topics in MI and AMI documents (left) and the most frequent words in the four most frequent topics (right).*

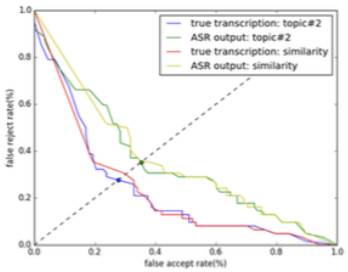


Figure 2: *DET curve for automatic detection of 'diabetes' topic using the true transcription of MI_test data and ASR output.*

Impact of ASR Performance on Free Speaking Language Assessment

K.M.Knill[1], *M.J.F.Gales*[1], *K. Kyriakopoulos*[1], *A. Malinin*[1], *A. Ragni*[1], *Y. Wang*[1], *A.P.Caines*[2]

[1]ALTA Institute / Engineering Department
[2]ALTA Institute / Computer Lab
Cambridge University, UK

{kate.knill,mjfg,kk492,am969,ar527,yw396}@eng.cam.ac.uk,apc38@cam.ac.uk

Abstract

In free speaking tests candidates respond in spontaneous speech to prompts. This form of test allows the spoken language proficiency of a non-native speaker of English to be assessed more fully than read aloud tests. As the candidate's responses are unscripted, transcription by automatic speech recognition (ASR) is essential for automated assessment, as shown in Fig. 1. ASR will never be 100% accurate so any assessment system must seek to minimise and mitigate ASR errors.

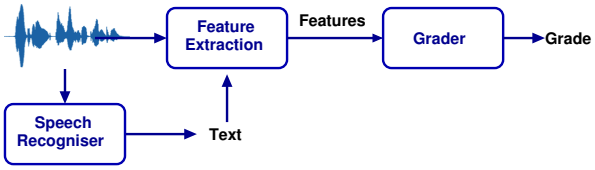
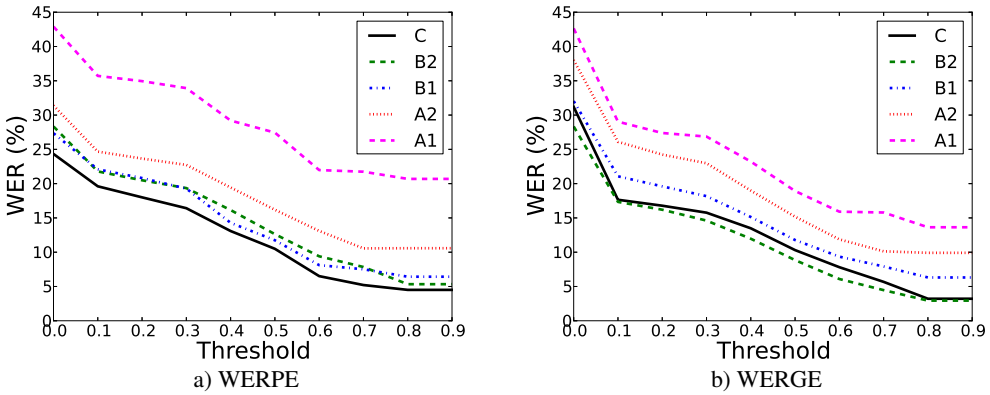


Figure 1: *Free speaking language assessment auto-marker framework.*

This paper considers the impact of ASR errors on the performance of free speaking test auto-marking systems. Firstly rich linguistically related features, based on part-of-speech tags from statistical parse trees, are investigated for assessment. Then, the impact of ASR errors on how well the system can detect whether a learner's answer is relevant to the question asked is evaluated. Finally, the impact that these errors may have on the ability of the system to provide detailed feedback to the learner is analysed. Feedback on what a candidate needs to do to improve their proficiency is highly desirable, but precision is important as the learner will get confused if the feedback is based on an incorrectly recognised word. When a speaker mis-pronounces a word or speaks ungrammatically the ASR task is made harder and the WER for words with pronunciation (WERPE) and grammatical (WERGE) errors is much higher than for all words. Fig. 2 illustrates the WERPE and WERGE at different CEFR grade levels. The lower the grade, and least proficient the speaker, the higher the number of recognition errors. As the recognition confidence threshold is increased the WERPE and WERGE significantly reduce. This can be used to mitigate the effect of WER on feedback.



a) WERPE b) WERGE

Figure 2: *WER against confidence threshold of all words with pronunciation (WERPE) and grammatical (WERGE) errors for System 2 for CEFR grades A1 (lowest) to C (highest).*

This research was funded under the ALTA Institute, Cambridge University. Thanks to Cambridge English Language Assessment for supporting this research and providing access to the BULATS data.

Speech pre-enhancement in realistic environments

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, United Kingdom
[2]Dept. Medical Physics and Acoustics and Cluster of Excellence Hearing4all, University of Oldenburg, Germany
c.chermaz@sms.ed.ac.uk

In recent years there has been growing interest - both from industry and academia - in improving the intelligibility of speech playback, a field known also as NELE (Near End Listening Enhancement). Across different applications (eg. television, telephony, public address systems), there is one shared problem: speech is harder to understand in the presence of noise and reverberation.
Speech pre-enhancement techniques are typically designed to deal with either additive noise or reverberation, but few attempts have been made at dealing with both - as each of them poses different challenges and requires a different approach. Moreover, algorithms are typically tested in lab conditions (eg. with speech shaped noise), which are easy to control and describe, but do not correspond to realistic situations. Finally, these methods are often evaluated with objective measures only, as tests involving human listeners - the golden standard in this field - are time consuming and expensive.
The present study aims at testing NELE techniques with human listeners - with particular interest on adaptive methods - in simulated real life scenarios, where fluctuating noise and reverberation pose an obstacle to communication.

This project has received funding from the EUs H2020 research and innovation programme under the MSCA GA 67532 (the ENRICH network: www.enrich-etn.eu)

Voice Activity Detection Using Neurograms

Wissam A. Jassim and Naomi Harte

Sigmedia, ADAPT Centre, School of Engineering, Trinity College Dublin, Ireland
wissam.jassim@tcd.ie , nharte@tcd.ie

Abstract:

Existing acoustic-signal-based algorithms for Voice Activity Detection (VAD) do not perform well in the presence of noise. In this study, we propose a method to improve VAD accuracies by employing another type of speech representation which is derived from the responses of the human Auditory-Nerve (AN) system. The neural responses referred to as a neurogram are simulated for each input speech signal using a computational model of the AN system with a range of Characteristic Frequencies (CFs). Features are extracted from neurograms using Discrete Cosine Transform (DCT), and a Multilayer Perceptron (MLP) classifier is trained based on the extracted features to predict the true VAD intervals. The proposed method was evaluated using QUT-NOISE-TIMIT corpus [1], and the Detection Cost Function (DCF) scoring algorithm [2] was employed as an accuracy measure. The proposed neural-feature-based method outperformed most of the baseline methods such as G.729 algorithm [3], statistical-model-based method by Sohn *et al.* [4], low complexity method by Tan and Lindberg [5], and the feature-combination-based method by Segbroeck *et al.* [6]. Figure 1 shows the block diagram of the proposed method.

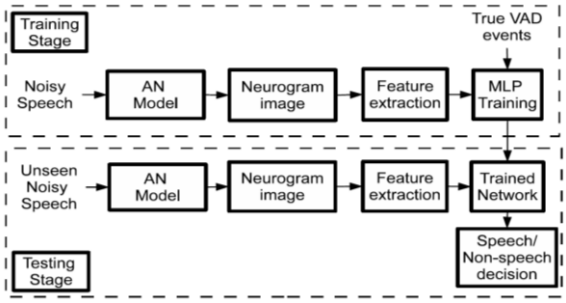


Fig. 1. Block diagram of the proposed VAD algorithm

Results:

Table 1. DCF (%) errors for enrolment set

Method \ SNR, dB	15	10	5	0	-5	-10
G.729	19.30	21.20	22.21	23.90	28.33	30.34
Sohn *et al.*	11.60	13.69	19.94	25.59	30.31	36.27
Tan & Lindberg	8.18	9.53	11.20	14.13	19.18	26.33
Segbroeck *et al.*	7.15	7.39	10.85	10.60	17.75	24.10
HSR	8.80	8.90	10.95	14.57	20.66	25.45
MSR	6.12	6.93	9.29	10.83	19.02	22.63
LSR	7.15	7.18	9.97	12.57	19.77	24.74

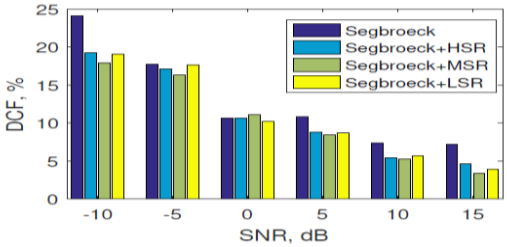


Fig. 2. DCF (%) errors of combining features for enrolment set

Table 2. DCF (%) errors for verification set

Method \ SNR, dB	15	10	5	0	-5	-10
G.729	14.69	20.17	21.52	24.65	29.94	32.05
Sohn *et al.*	10.57	13.04	19.48	26.23	32.38	38.98
Tan & Lindberg	8.80	8.91	10.98	15.79	17.33	22.28
Segbroeck *et al.*	4.99	4.59	9.98	11.04	18.90	22.27
HSR	7.23	6.38	10.48	11.82	18.75	26.21
MSR	4.82	4.96	10.68	9.62	15.98	24.86
LSR	5.52	5.38	10.95	10.72	15.35	26.66

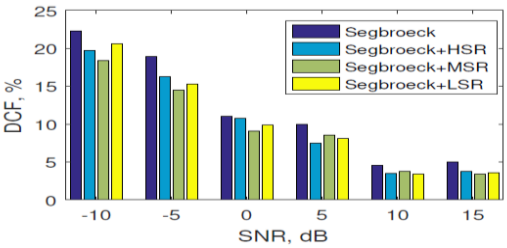


Fig. 3. DCF (%) errors of combining features for verification set

Reference:

[1] David B. Dean, Sridha Sridharan, Robert J. Vogt, and Michael W. Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," in Interspeech 2010, September 2010.

[2] National Institute of Standards and Technology (NIST), "NIST open speech-activity-detection evaluation," https://www.nist.gov/itl/iad/mig/nist-open-speech-activity-detection-evaluation , May 2016.

[3] A. Benyassine, E. Shlomot, H. Y. Su, D. Massaloux, C. Lamblin, and J. P. Petit, "ITU-T recommendation G.729 Annex B: A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," Comm. Mag., vol. 35, no. 9, pp. 64–73, Sept. 1997.

[4] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung, "A statistical model-based voice activity detection," IEEE Signal Processing Letters, vol. 6, no. 1, pp. 1–3, Jan 1999.

[5] Z. H. Tan and B. Lindberg, "Low-complexity variable frame rate analysis for speech recognition and voice activity detection," IEEE Journal of Selected Topics in Signal Processing, vol. 4, no. 5, pp. 798–807, 2010.

[6] Maarten Van Segbroeck, Andreas Tsiartas, and Shrikanth Narayanan, "A robust frontend for vad: exploiting contextual, discriminative and spectral cues of human voice.," in Interspeech, 2013, pp. 704–708.

Poster Session B

1	Matthew P. Aylett, David A. Braude	Grassroots: Using Speech Sythesis to Curate Audio Content for Low Power Community FM Radio	CereProc Ltd., Edinburgh
2	Joanna Rownicka, Steve Renals, Peter Bell	Understanding deep speech representations	University of Edinburgh
3	Feifei Xiong, Jon Barker, Heidi Christensen	Deep Learning of Articulatory-Based Representations for Dysarthric Speech Recognition	University of Sheffield
4	Manuel Sam Ribeiro, Aciel Eshky, Korin Richmond, Steve Renals	Towards Robust Word Alignment of Child Speech Therapy Sessions	University of Edinburgh
5	Oliver Watts, Cassia Valentini-Botinhao, Felipe Espic, Simon King	Exemplar-based speech waveform generation	University of Edinburgh
6	Leigh Clark, Phillip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Benjamin Cowan	The State of Speech in HCI: Trends, Themes and Challenges	University College Dublin
7	Danny Websdale and Ben Milner	Using Visual Speech Information for Noise and Signal-to-Noise Ratio Independent Speech Enhancement	University of East Anglia
8	Eva Fringi and Martin Russell	Analysis of phone errors attributable to phonological effects associated with language acquisition through bottleneck feature visualisations	University of Birmingham
9	Maria O'Reilly, Amelie Dorn, Ailbhe Ní Chasaide	Intonation of declaratives and questions in South Connaught and Ulster Irish	Trinity College Dublin
10	Ilaria Torre, Emma Carrigan, Killian McCabe, Rachel McDonnell, Naomi Harte	Mismatched audio-video smiling in an avatar and its effect on trust	Trinity College Dublin
11	Jeremy H. M. Wong and Mark J. F. Gales	Teacher-student learning and ensemble diversity	University of Cambridge
12	Emer Gilmartin, Brendan Spillane,Maria O'Reilly, Ketong Su, Christian Saam, Benjamin R. Cowan,Carl Vogel, Nick Campbell, VincentWade	Dialog Acts in Greeting and Leavetaking in Social Talk	Trinity College Dublin
13	Brendan Spillane, Emer Gilmartin, Christian Saam, Leigh Clark, Benjamin R. Cowan, Vincent Wade	Introducing ADELE: A Personalized Intelligent Companion	Trinity College Dublin
14	George Sterpu, Christian Saam, Naomi Harte	Progress on Lip-Reading Sentences	Trinity College Dublin

Grassroots: Using Speech Sythesis to Curate Audio Content for Low Power Community FM Radio

Matthew P. Aylett, David A. Braude

CereProc Ltd., Edinburgh, UK

{matthewa,dave}@cereproc.com

Abstract

The Grassroot Wavelengths project will create a game changing network of inclusive digital platforms for citizen engagement and community deliberation. Our approach includes features of the Living Lab and Participatory Design methods for setting up stations and services and understanding the processes in which they will be used and appropriated, along with an emphasis on synthetic speech to support the curation of audio content, thus turning data into media. Building on the success of the existing RootIO platform with its proven commons-oriented technology and catalytic capacities for promoting/enabling collective awareness and action, participatory innovation, community resilience, and media pluralism we will enhance use and accessibility of networked community radio through text-to-speech (TTS), community oriented programming applications, generating automated audio advertising, dedications, government information and trusted news partners. We will compare commercial TTS (CereVoice) and open source TTS (Kaldi-Idlak) in four target languages - Madeiran Portuguese, Irish Gaelic and Romanian. We will present an overview of the project and our planned research work together with current achievements to date.

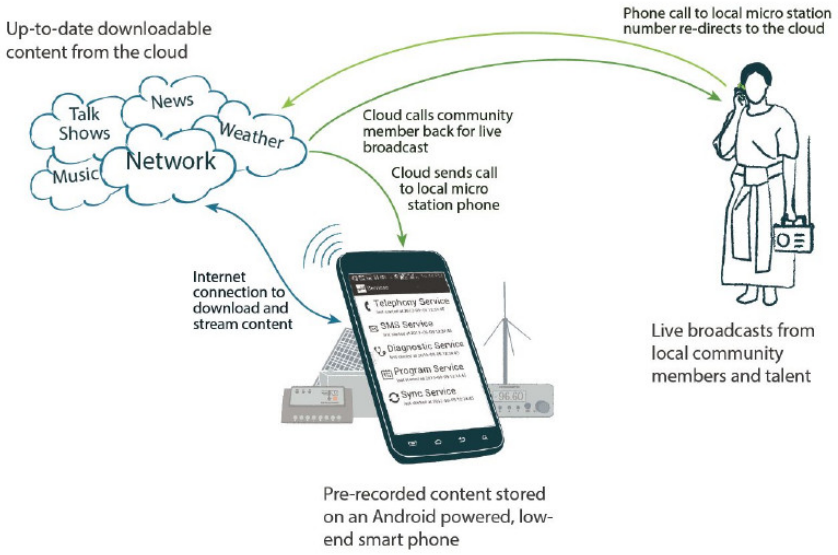


Figure 1: *RootIO's technical stack. TTS will be integrated into the RootIO system to generate automatic radio program content.*

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

- ASOCIATIA ACTIVEWATCH, Romania
- ADENORMA - ASSOCIACAO DESENVOLVIMENTO COSTA NORTE DA MADEIRA, Portugal
- ASSOCIATION MONDIALE DES RADIODIFFUSEURS COMMUNAUTAIRES - EUROPE, Belgium
- THE BERE ISLAND PROJECTS GROUP, Ireland
- CEREPROC LTD, United Kingdom
- CENTRUL ROMAN PENTRU JURNALISM DE INVESTIGATIE, Romania
- ROOTIO LTD, Portugal
- UNIVERSITY COLLEGE CORK - NATIONAL UNIVERSITY OF IRELAND, CORK, Ireland

Understanding deep speech representations

Joanna Rownicka, Steve Renals, Peter Bell

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

j.m.rownicka@sms.ed.ac.uk, {s.renals, peter.bell}@ed.ac.uk

Methods for interpreting and understanding neural network behavior have become an important aspect of robust neural network validation. They help to verify that the reported accuracy stems from a proper problem representation rather than from the exploitation of artifacts in the data. Learning about the network's learning process can also be useful in identifying the model's bias or in predicting the model's behavior when applied to new conditions.

Deep convolutional neural network architectures with small kernels have previously been shown to perform well in modeling speech, resulting in high speech recognition accuracy. In this work, we try to extend the analysis in order to better understand why deep CNNs, primarily used for modeling spatial relations in images, are also effective in speech recognition. The aim of this work is to gain more insight on the type of representations learned by deep CNNs and to compare them with DNN representations.

In the qualitative evaluation part of our work, we visualize deep representations learned at the utterance level, using the Aurora-4 and MGB English datasets. We use gender, speaker, acoustic condition and channel labels to show how the utterance-level representations cluster in the activation space. We also quantitatively evaluate discriminative power of the learned representations based on speaker and acoustic conditions. The learned representations can be regarded as vectors of fixed length that characterize utterances, hence we experiment with the use of learned representations as the additional input for acoustic models adaptation task. We further compare deep CNN and DNN learned embeddings with an i-vector representation for the adaptation of TDNN models.

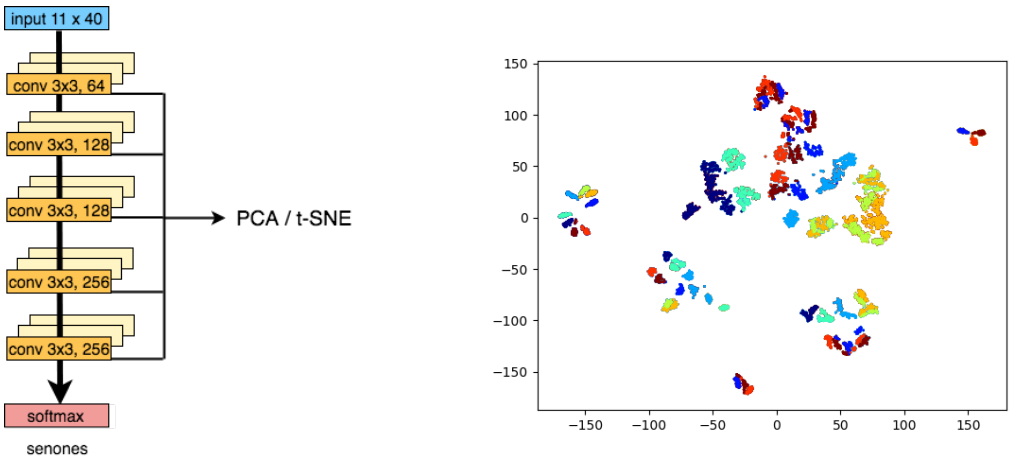


Figure 1: (*left*) Deep CNN embedding extraction framework. We take the frames activations averaged over utterance of all of the channels of the last layer of each convolutional block of a fully-convolutional network, we vectorize the output of each block, we concatenate resulting vectors, and we use the PCA transform to reduce the dimensionality of an embedding. Those embeddings are then used for TDNN adaptation. (*right*) Example of t-SNE visualization of deep CNN embeddings. Different colors represent distinct speakers of the Aurora4 dataset.

Deep Learning of Articulatory-Based Representations for Dysarthric Speech Recognition

Feifei Xiong[1], *Jon Barker*[1], *Heidi Christensen*[1,2]

[1]Speech and Hearing Group (SPandH), Dept. of Computer Science, University of Sheffield, Sheffield, UK, [2]Centre for Assistive Technology and Connected Healthcare (CATCH), University of Sheffield, Sheffield, UK
{*f.xiong, j.p.barker, heidi.christensen*}*@sheffield.ac.uk*

Improving the robustness of dysarthric speech recognition is an active and challenging research field despite the fact that state-of-the-art automatic speech recognition (ASR) technology has obtained great progress. The large inter-speaker variability observed in disordered speech significantly degrades the performance of off-the-shelf ASR systems. In this work, we investigate articulatory-based representations to augment the conventional acoustic features to better model dysarthric variability, motivated by the observation that such low-level articulatory information behaves less variant to the atypical speech variability caused by dysarthria.

Specifically, the articulatory-based representation is estimated by an acoustic-articulatory inverse mapping learned using deep neural networks (DNNs). GnuSpeech [1] is employed to generate normal speech data as well as the corresponding aligned articulatory features for training, and long short-term memory (LSTM) units are used in DNNs to temporally smooth the estimated articulatory-based output. The estimated articulatory features are then applied to augment current hybrid DNN/HMM ASR system, and evaluation is based on single-word UASPEECH task [2], for which we have released the baseline Kaldi script in [3]. Experimental results show that a relative recognition accuracy improvement of 12.6% can be achieved by introducing articulatory-based representations in the feature domain.

Acknowledgements

The research is supported by the DeepArt project sponsored by Google Inc. The authors would like to thank Siddharth Sehgal for valuable discussions.

References

[1] GnuSpeech website: `https://www.gnu.org/software/gnuspeech/`

[2] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. Huang, K. Watkin, and S. Frame, *Dysarthric Speech Database for Universal Access Research*, in Proceedings of Interspeech, Brisbane, Australia, Sep. 2008.

[3] Baseline Kaldi script for UA-SPEECH corpus: `https://github.com/ffxiong/uaspeech`

Towards Robust Word Alignment of Child Speech Therapy Sessions

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

Centre for Speech Technology Research, University of Edinburgh, UK

{sam.ribeiro, aeshky, s.renals, korin}@ed.ac.uk

Abstract

Developmental Speech Sound Disorders (SSDs) are a common communication impairment in childhood. These describe cases where children consistently exhibit difficulties in the production of specific speech sounds in their native language. SSDs have the potential to negatively affect the lives and the development of children. For example, self-awareness of disordered speech may lead to low-confidence in social situations or introduce communication barriers that lead to increased difficulty in learning and decreased literacy levels [1].

Clinical intervention is typically available for children with SSDs. However, current clinical methods for speech therapy are subjective and inaccurate [2]. Instrumented methods, such as spectrogram analysis or articulatory imaging, are useful, but require a large amount of manual effort from speech pathologists. In the Ultrax Speech Project (`www.ultrax-speech.org`), we explore objective methods that could alleviate manual processes undertaken by Speech and Language Therapists (SLTs) using audio and ultrasound. This paper lays out some of the major challenges for processing this type of data and presents initial results for the tasks of speaker labeling and word alignment.

We use the `UltraSuite Repository` [3], a collection of datasets of ultrasound and acoustic data collected from recordings of child speech therapy sessions. The repository contains one dataset of Typically Developing children and two datasets of children with SSDs. There are various challenges associated with this type of data. For example, the interaction between speech therapist and child, insertions and deletions with respect to the given prompt, mispronounciations, and the various challenges associated with child speech processing and disordered speech processing. These challenges are noticeable when force-aligning the audio with the expected prompt. Using baseline standard methods, we observe an f1-score of word recovery of 69% in Typically Developing children and 30% in diagnosis sessions of Speech Disordered children.

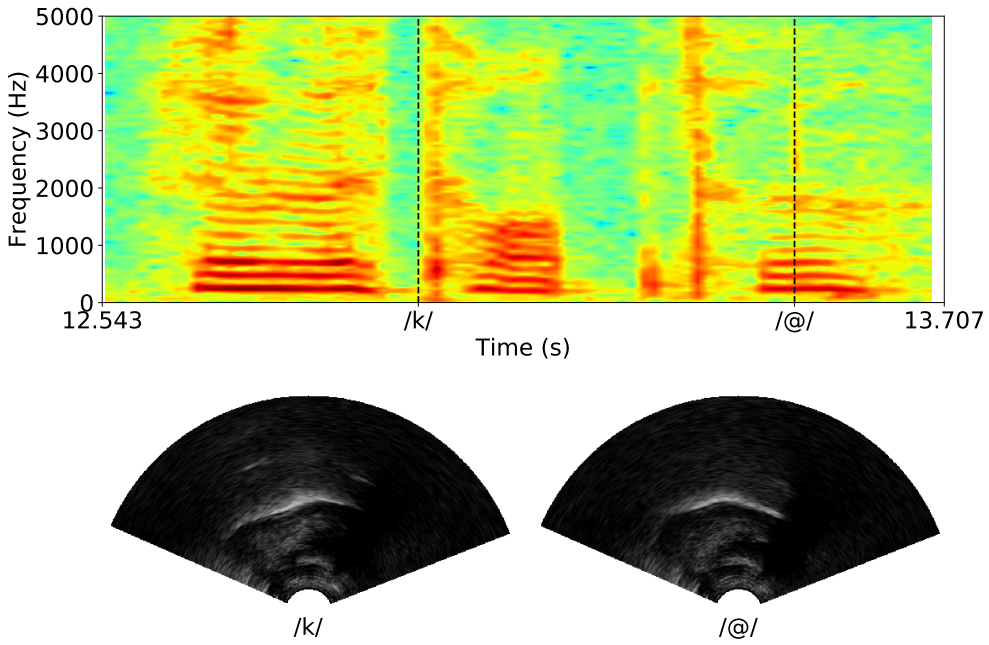


Figure 1: *Spectrogram for the word* helicopter *with two corresponding ultrasound frames, ellicited during a session with a six-year-old child diagnosed with velar fronting. Ultrasound frames show a mid-saggital view of the oral cavity with the tip of the tongue facing right.*

1. References

[1] S. McLeod and E. Baker, *Children's speech: An evidence-based approach to assessment and intervention*. Pearson, 2016.

[2] S. Howard and A. Lohmander, *Cleft palate speech: assessment and intervention*. John Wiley & Sons, 2011.

[3] A. Eshky, M. S. Ribeiro, J. Cleland, K. Richmond, Z. Roxburgh, J. Scobbie, and A. Wrench, *UltraSuite: A Repository of Ultrasound and Acoustic Data from Child Speech Therapy Sessions*. Manuscript submitted for publication, 2018.

Exemplar-based speech waveform generation

Oliver Watts, Cássia Valentini-Botinhão, Felipe Espic, Simon King

The Centre for Speech Technology Research, Edinburgh University, UK

owatts@inf.ed.ac.uk, cvbotinh@inf.ed.ac.uk, felipe.espic@ed.ac.uk, Simon.King@ed.ac.uk

Abstract

This work presents a simple but effective method for generating speech waveforms by selecting small units of stored speech to match a low-dimensional target representation. The method is designed as a drop-in replacement for the vocoder in a deep neural network-based text-to-speech system. Most previous work on hybrid unit selection waveform generation relies on phonetic annotation for determining unit boundaries, or for specifying target cost, or for candidate preselection. In contrast, our waveform generator requires no phonetic information, annotation, or alignment. Unit boundaries are determined by epochs, and spectral analysis provides representations which are compared directly with target features at runtime. As in unit selection, we minimise a combination of target cost and join cost, but find that greedy left-to-right nearest-neighbour search gives similar results to dynamic programming. The method is fast and can generate the waveform incrementally. We use publicly available data and provide a permissively-licensed open source toolkit for reproducing our results.

1. Introduction

We present a waveform generation module which can be dropped in to a statistical parametric text-to-speech (TTS) synthesis system to turn it into a 'hybrid' synthesiser. By *hybrid*, we mean that waveforms are produced by waveform unit selection and concatenation, but that the selection is guided by the output of a high quality acoustic model. Typically, the acoustic features used to guide selection could themselves be passed through a vocoder to produce a stable, intelligible and reasonably natural-sounding waveform. Until recent developments in the direct time-domain prediction of waveforms, such hybrid systems were the state of the art in natural-sounding speech synthesis, and they are still a dominant form of synthesiser in commercial applications.

In the majority of hybrid synthesis work, the speech units selected are relatively large, phonetically determined units, such as diphones and halfphones. The current work aims to use smaller units which can be determined without phonetic annotation. There are several possible benefits to this: it means the unit selection module can be agnostic about the symbolic content of speech to be synthesised in the same way as a vocoder, it opens up the possibility of simply sharing unit databases across dialects and languages, and systems selecting smaller units are conceivably less susceptible to degradation due to inadequate amounts of data and poor annotation.

Our work is different from most previous work in that we make no reliance on phonetic labels. It differs from all previous work in that no use is made of dynamic programming for unit selection: we find greedy search to be effective. Furthermore, we select units whose temporal bounds are defined by knowledge of speech structure: we select units pitch synchronously rather than using an arbitrary frame size in voiced regions.

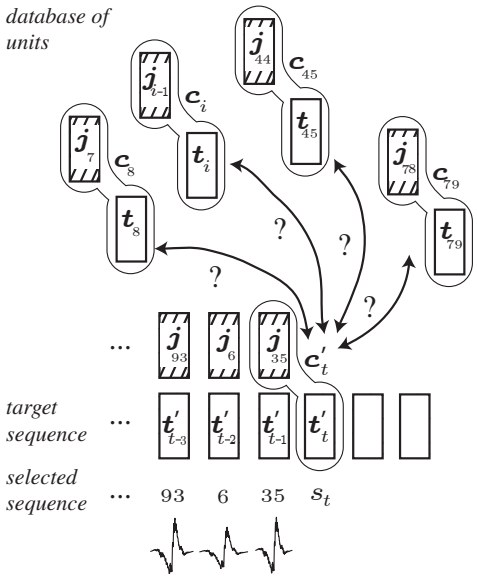


Figure 1: *Exemplar-based synthesis.*

2. Proposed system

The task performed by the module at synthesis time is to generate a waveform consistent with a given sequence of feature vectors. As with the inputs to the synthesis module of a standard vocoder, this sequence might be predicted by a statistical model (as part of e.g. a TTS or voice conversion system), or could simply be extracted from natural speech. In all cases, we term this the *target sequence*. For each vector $\boldsymbol{t}'_t$ in the target sequence, an exemplar is chosen by searching a database of units, and finally the exemplars chosen to cover each vector in the target sequence are joined to produce a speech waveform. This process is illustrated in Fig.1, in the case that a unit is made of one epoch. As with other unit selection approaches, the goal of database search is to select a sequence of units to minimise a cost which incorporates two types of constraint. Firstly, each unit should be similar to the acoustics encoded by the target vector (divergence is penalised by the *target* component of the cost); secondly, neighbouring selected units should be acoustically compatible in order to minimise audible artifacts when they are joined (incompatibility is penalised by the *join* component of the search cost). The target and join components can also be thought of as *fidelity* and *fluency* measures: the first scores how faithfully the message encoded by the target sequence is rendered, and the second, how fluently this is done.

In our poster we will present more details on the system as well as results of listening tests.

Acknowledgements: This research was supported by EPSRC Standard Research Grant EP/P011586/1, *Speech Synthesis for Spoken Content Production (SCRIPT)*.

Title: The State of Speech in HCI: Trends, Themes and Challenges

Authors and Affiliations: Leigh Clark[1], Phillip Doyle[1], Diego Garaialde[1], Emer Gilmartin[2], Stephan Schlögl[3], Jens Edlund[4], Matthew Aylett[5], João Cabral[6], Cosmin Munteanu[7], Benjamin Cowan[1]

[1]School of Information and Communication Studies, University College Dublin, Ireland

[2]Speech Communication Laboratory, Trinity College Dublin, Ireland

[3]MCI Management Center, Innsbruck, Austria

[4]Department of Speech, Music and Hearing, KTH Royal Institute of Technology, Sweden

[5]School of Informatics, University of Edinburgh, United Kingdom

[6]School of Computer Science and Statistics, Trinity College Dublin, Ireland

[7]Institute of Communication, Culture, Information and Technology, University of Toronto Mississauga, Canada

Abstract:

Speech interfaces are growing in popularity. Through a review of 68 research papers this work maps the trends, themes, findings and methods of empirical research on speech interfaces in HCI. We find that most studies are usability/theory-focused or explore wider system experiences, evaluating Wizard of Oz, prototypes, or developed systems by using self-report questionnaires to measure concepts like usability and user attitudes. A thematic analysis of the research found that speech HCI work focuses on nine key topics: system speech production, modality comparison, user speech production, assistive technology & accessibility, design insight, experiences with interactive voice response (IVR) systems, using speech technology for development, people's experiences with intelligent personal assistants (IPAs) and how user memory affects speech interface interaction. From these insights we identify gaps and challenges in speech research, notably the need to develop theories of speech interface interaction, grow critical mass in this domain, increase design work, and expand research from single to multiple user interaction contexts so as to reflect current use contexts. We also highlight the need to improve measure reliability, validity and consistency, in the wild deployment and reduce barriers to building fully functional speech interfaces for research.

Using Visual Speech Information for Noise and Signal-to-Noise Ratio Independent Speech Enhancement

Danny Websdale and Ben Milner

University of East Anglia, United Kingdom

d.websdale@uea.ac.uk, b.milner@uea.ac.uk

This work is concerned with using neural networks for estimating ideal ratio masks within a speech enhancement framework. Our previous work focused on supplementing traditional audio-only applications with visual information within noise type and signal-to-noise ratio (SNR) dependant systems. Results found that combining both modalities gave large gains in intelligibility over audio-only at low SNRs, and equivalent performance at high SNRs. We also found that extracting visual information using convolutional neural networks (CNN) provided further performance gains when compared against traditional active appearance models (AAM) features.

In this work we expand on this by comparing audio-only, visual-only and audio-visual models for noise type and SNR independent systems. Figure 1 shows our fully end-to-end audio-visual convolutional recurrent feed-forward hybrid speech enhancement model, for audio-only, the visual convolutional network is removed, and for visual-only the audio channel is removed. Visual features are extracted by first fitting a 90×110 pixel box centred around the mouth before downsampling to 64×64 for input to the network. The acoustic feature selected is the multi-resolution cochleagram, which combines 4 different cochleagrams into a single feature, specifically designed for mask estimation within a cochleagram framework.

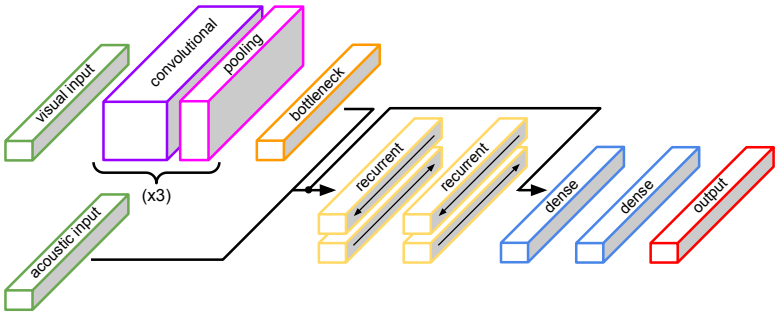


Figure 1: *Audio-visual convolutional recurrent feed-forward hybrid speech enhancement architecture.*

All models are trained with babble, factory and speech-shape noise at SNRs of -10 dB, 5 dB, 0 dB and +5 dB. We then evaluate the generalisation of these models with unseen noise types, cafeteria babble and street noise. Evaluations of the noise-independent models show that in seen noise conditions, audio-visual provides large gains over audio-only at low SNRs with equivalent performance at high SNRs, and large gains over visual-only across all SNRs. In unseen noise conditions, audio-visual and visual-only still provide large improvements in intelligibility over unprocessed, however, audio-only performs poorly, particularly at low SNRs. This degradation in performance for audio-only reveals the strength of the noise-independent visual channel, which is unaffected by acoustic noise. This allows both audio-visual and visual-only to generalise well to unseen noise conditions, while combining both audio and visual information still provides best performance.

Analysis of phone errors attributable to phonological effects associated with language acquisition through bottleneck feature visualisations

Eva Fringi, Martin Russell

Department of Electronic Electrical and Systems Engineering,
University of Birmingham B15 2TT, UK
exf111@bham.ac.uk, M.J.RUSSELL@bham.ac.uk

Previous work aimed to investigate the extent to which errors attributable to phonological effects associated with language acquisition (PEALA) contribute to the output of children's ASR. Opposite to what was intuitively expected, the proportion of errors predictable from PEALA was positively correlated with recognition accuracy, therefore increased across ages. In order to interpret this finding, the present paper employs a DNN-HMM automatic speech recognition system, built on the CSLU children's speech corpus, to produce bottleneck feature (BNF) visualisations of phones and examine how these relate with respect to PEALA. The focus is drawn particularly on ASR errors caused by phone confusions, which are compared against phone substitution pairs indicated by PEALA. The ASR results confirm the previously observed interaction between errors predictable from PEALA and rising accuracy, but also suggest that these errors only account for a small percentage of the total phone substitution error. The BNF visualisations for the most part outline the age progression smoothly and demonstrate clear clusters of neighbouring phones consistently. The distance between PEALA related phones can be partitioned in four sets; two that increase with age (at a higher or lower rate), one that roughly remains constant and one that decreases with age.

Intonation of declaratives and questions in South Connaught and Ulster Irish

Maria O'Reilly[1], *Amelie Dorn*[2], *Ailbhe Ní Chasaide*[1]

[1]Trinity College Dublin, Ireland; [2]Austrian Academy of Sciences, Austria

moreil12@tcd.ie amelie.dorn@oeaw.ac.at, anichsid@tcd.ie

Abstract

This paper provides an overview of the ongoing work on the intonation in two dialects of Irish, South Connaught Irish and Ulster (Donegal) Irish. The melody of these dialects is of particular interest to us for a number of reasons. First, this work contributes to our growing understanding of the prosody of Irish, of Celtic languages, and of intonational typology in general. Second, we aim to incorporate it into our multi-dialect text-to-speech for Irish dialects [1].

South Connaught Irish (SCI) and Ulster Irish (UI) are interesting in that they use diverse intonation patterns. Specifically, in declaratives SCI typically uses a sequence of falling (H*+L) accents, while UI typically employs a sequence of rising (L*+H) accents [2] (Figure 1). UI is particularly unique, as across languages a rising intonation is typically associated with questions [3].

Figure 2 compares the intonation of declaratives (DEC) and questions (WHQ and YNQ) in the two dialects (A = accented syllable, shaded grey). It appears that SCI and UI make little use of contour type to differentiate questions from declaratives, but chiefly use phonetic f0 markers [4-5]. Questions are signalled by raising the pitch of the first accent, A1 in the phrase (WHQ and YNQ in both dialects), and that of the phrase-final accent, AN (WHQ and YNQ in SCI; only YNQ in UI). With respect to tonal patterns, SCI favours H*+L (H*+L) H*+L %, while UI – the L*+H L*+H L*+H % sequence in DEC, WHQ and YNQ alike. Alternatives include a phrase-final rise (H%) in YNQ (both dialects, not shown), and a phrase-initial H* in WHQ (UI, Figure 2).

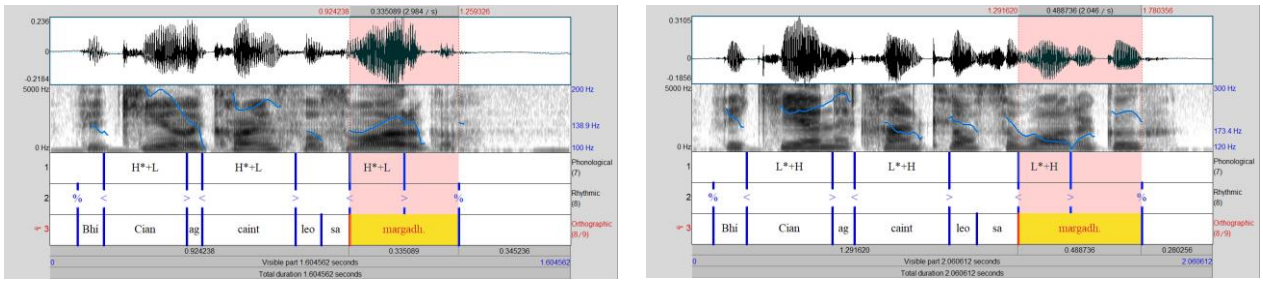


Figure 1: *Examples of f0 contours of a declarative in SCI (left) and UI (right).*

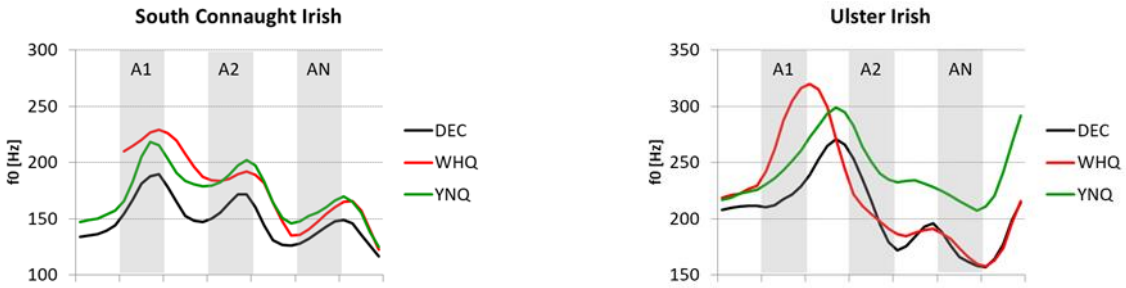


Figure 2: *Examples of time-normalised f0 contours in declaratives and wh- and yes/no questions in SCI and UI.*

References

[1] "abair.ie – An Sintéiseoir Gaeilge," Available at www.abair.tcd.ie.

[2] M. Dalton, and A. Ní Chasaide, "Modelling intonation in three Irish dialects," in 15th International Congress of Phonetic Sciences, Barcelona, Spain, 2003, pp. 1073-1076.

[3] D. Hirst, and A. Di Cristo, "A survey of intonation systems," in *Intonation systems: A survey of twenty languages*, D. Hirst and A. Di Cristo, Eds. Cambridge: Cambridge University Press, 1998, pp. 1-44.

[4] A. Dorn, M. O'Reilly, and A. Ní Chasaide, "Prosodic signalling of sentence mode in two varieties of Irish (Gaelic)," in 17th International Congress of Phonetic Sciences, Hong Kong, China, 2011, pp. 611-614.

[5] M. O'Reilly, and A. Ní Chasaide, "Declination, peak height and pitch level in declaratives and questions of South Connaught Irish," in INTERSPEECH 2015, Dresden, Germany, 2015, pp. 978-982.

Mismatched audio-video smiling in an avatar and its effect on trust

Ilaria Torre, Emma Carrigan, Killian McCabe, Rachel McDonnell, Naomi Harte

Abstract

Correctly interpreting an interlocutor's emotional expression is paramount to a successful interaction, and the vocal channel contributes to expressing it. But what happens when the interlocutor expressing an emotion is a machine? The facilitation of human-machine communication and cooperation is of growing importance as smartphones, autonomous cars, or social robots increasingly pervade human social spaces. Previous research has shown that emotionally expressive avatars generally elicit higher cooperation and trust than "neutral" ones (e.g. Elkins & Derrick, 2013; Krumhuber et al., 2007). Since emotional expressions are multi-modal, the question of which of the available channels should be most carefully considered in design arises. Would a mismatch in the emotion expressed in the face and voice influence people's cooperation with an avatar? To answer this question, we developed a simulated survival game where people had to cooperate with a computer-generated avatar in order to survive a crash landing on the moon. The avatar's face and voice were designed to either smile or not, in 2 matched and 2 mismatched conditions: smiling voice and face, neutral voice and face, smiling voice only (neutral face), smiling face only (neutral voice). The experiment was set up in a museum over the course of several weeks, where visitors were invited to interact with it. We report preliminary results from hundreds of visitors, showing that people tend to trust the avatar in the mismatched condition with the smiling face and neutral voice more.

References

Elkins, A. C., & Derrick, D. C. (2013). The sound of trust: Voice as a measurement of trust during interactions with embodied conversational agents. *Group Decision and Negotiation*, *22*(5), 897–913.

Krumhuber, E., Manstead, A. S. R., Cosker, D., Marshall, D., Rosin, P. L., & Kappas, A. (2007). Facial dynamics as indicators of trustworthiness and cooperative behavior. *Emotion*, *7*(4), 730–735.

Teacher-student learning and ensemble diversity

Jeremy H. M. Wong and Mark J. F. Gales

University of Cambridge, United Kingdom

jhmw2@cam.ac.uk, mjfg@eng.cam.ac.uk

Abstract

Ensemble methods have often been shown to produce significant performance gains in Automatic Speech Recognition (ASR). These can be viewed as a Monte Carlo approximation to Bayesian inference of the word sequence hypothesis. The ensemble may capture uncertainty about the model parameters and design, represented within the diversity of the models used. Often for simplicity, ASR ensembles are generated while constraining some of the model design aspects to be the same across the ensemble. However, this restricts the diversity of models within the ensemble, and may result in a biased Bayesian inference estimate. Allowing more design aspects to vary between models may allow the ensemble to capture a greater model diversity, and therefore more holistically represent model uncertainty.

Although an ensemble may perform well, it can be computationally expensive to use for recognition, with the computational cost generally scaling linearly with the ensemble size. Teacher-student learning is one possible method to mitigate this cost, by training a single student model, $\boldsymbol{\Theta}$, to emulate the combined behaviour of the ensemble, $\widehat{\boldsymbol{\Phi}}$. Only this single student needs to be used for recognition. A commonly used criterion to train the student is to minimise the KL-divergence between per-frame state cluster, s, posteriors, shown in the first row of Table 1. Here, t is the time step and $\mathbf{O}_t$ are the observations. However, this criterion does not take into account the sequential nature of the data, and the interactions between the acoustic, alignment, and language models. Furthermore, this criterion requires that all models within the ensemble must use the same set of state clusters as the student. This limits the diversity that the ensemble is allowed to capture.

The sequential nature of speech data can be taken into account by using a KL-divergence criterion over sequence posteriors. One possible criterion is the KL-divergence between word sequence, $\boldsymbol{\omega}$, posteriors. This criterion is general, in that it does not place any constraints on the forms of diversity that are allowed within the ensemble. Although the gradient of this criterion can be computed over n-best lists, the computational cost can become impractical as the number of competing hypotheses increases. Approximations can be taken to allow more efficient gradient computation over lattices or lattice-free graphs. One possible approximation is to instead consider the KL-divergence between arc sequences, $\mathbf{a}$. The difference between words and arcs is that arcs have defined start and end times. This results in a gradient that does not have a sum over sequences, and can be computed using two levels of forward-backward operations over lattices, or lattice-free graphs.

The arcs can be marked with words, phones, or state clusters. When marked with state clusters, the gradient can be computed even more efficiently, using only a single level of forward-backward operations. However, this marking again forces all models in the ensemble to use the same set of state clusters as the student, thereby limiting the diversity that the ensemble can capture. Instead, the arcs can be marked with logical Context Dependent (CD) states, c. This generalises the teacher-student criterion to allow for a diversity of state cluster sets, while still preserving the computational efficiency of a single level of forward-backward operations in the gradient computation. The gradient of this criterion can be computed over intersect states, $\widehat{s}$. Both the KL-divergences over state cluster sequences and logical CD state sequences can be realised using existing lattice-free implementations.

This work presents experimental results on the AMI-IHM dataset, for a preliminary investigation into training the student by minimising the KL-divergence between logical CD state sequence posteriors, implemented within a lattice-free framework. Lattice-free MMI training gives a single model WER of 26.8 %. An ensemble of models, each with a different set of state clusters, are combined to give a WER of 20.9 %. The student model, trained using the proposed logical CD state sequence criterion, has a WER of 22.2 %.

Table 1: Teacher-student learning criteria and gradients

Macrostate	Criterion	Student gradient
per-frame state cluster	$-\sum_{t}\sum_{s_t} P\left(s_t\middle\vert\boldsymbol{o}_t,\widehat{\boldsymbol{\Phi}}\right)\log P\left(s_t\middle\vert\boldsymbol{o}_t,\boldsymbol{\Theta}\right)$	$P\left(s_t\middle\vert\boldsymbol{o}_t,\boldsymbol{\Theta}\right)-P\left(s_t\middle\vert\boldsymbol{o}_t,\widehat{\boldsymbol{\Phi}}\right)$
word sequence	$-\sum_{\boldsymbol{\omega}} P\left(\boldsymbol{\omega}\middle\vert\mathbf{O},\widehat{\boldsymbol{\Phi}}\right)\log P\left(\boldsymbol{\omega}\middle\vert\mathbf{O},\boldsymbol{\Theta}\right)$	$P\left(s_t\middle\vert\mathbf{O},\boldsymbol{\Theta}\right)-\sum_{\boldsymbol{\omega}} P\left(s_t^{\boldsymbol{\Theta}}\middle\vert\boldsymbol{\omega},\mathbf{O},\boldsymbol{\Theta}\right)P\left(\boldsymbol{\omega}\middle\vert\mathbf{O},\widehat{\boldsymbol{\Phi}}\right)$
arc sequeuce	$-\sum_{\boldsymbol{\omega}}\sum_{\mathbf{a}\in\mathcal{G}_{\boldsymbol{\omega}}} P\left(\mathbf{a},\boldsymbol{\omega}\middle\vert\mathbf{O},\widehat{\boldsymbol{\Phi}}\right)\log P\left(\mathbf{a},\boldsymbol{\omega}\middle\vert\mathbf{O},\boldsymbol{\Theta}\right)$	$P\left(s_t\middle\vert\mathbf{O},\boldsymbol{\Theta}\right)-\sum_{a_t} P\left(s_t^{\boldsymbol{\Theta}}\middle\vert a_t,\mathbf{O},\boldsymbol{\Theta}\right)P\left(a_t\middle\vert\mathbf{O},\widehat{\boldsymbol{\Phi}}\right)$
state cluster sequence	$-\sum_{\boldsymbol{\omega}}\sum_{\mathbf{s}\in\mathcal{G}_{\boldsymbol{\omega}}} P\left(\mathbf{s},\boldsymbol{\omega}\middle\vert\mathbf{O},\widehat{\boldsymbol{\Phi}}\right)\log P\left(\mathbf{s},\boldsymbol{\omega}\middle\vert\mathbf{O},\boldsymbol{\Theta}\right)$	$P\left(s_t\middle\vert\mathbf{O},\boldsymbol{\Theta}\right)-P\left(s_t\middle\vert\mathbf{O},\widehat{\boldsymbol{\Phi}}\right)$
CD state sequence	$-\sum_{\boldsymbol{\omega}}\sum_{\mathbf{c}\in\mathcal{G}_{\boldsymbol{\omega}}} P\left(\mathbf{c},\boldsymbol{\omega}\middle\vert\mathbf{O},\widehat{\boldsymbol{\Phi}}\right)\log P\left(\mathbf{c},\boldsymbol{\omega}\middle\vert\mathbf{O},\boldsymbol{\Theta}\right)$	$P\left(s_t\middle\vert\mathbf{O},\boldsymbol{\Theta}\right)-\sum_{\widehat{s}_t\in\mathcal{G}_{s_t^{\boldsymbol{\Theta}}}} P\left(\widehat{s}_t\middle\vert\mathbf{O},\widehat{\boldsymbol{\Phi}}\right)$

Dialog Acts in Greeting and Leavetaking in Social Talk

Emer Gilmartin[1], *Brendan Spillane*[1], *Maria O'Reilly*[1], *Ketong Su*[1], *Christian Saam*[1], *Benjamin R. Cowan*[3],*Carl Vogel*[2], *Nick Campbell*[1], *Vincent Wade*[1]

[1]ADAPT Centre, School of Computer Science and Statistics, Trinity College Dublin
[2]Computational Linguistics Group, School of Computer Science and Statistics, Trinity College Dublin
[3]University College Dublin

gilmare, moreil, vogel, nick@tcd.ie, brendan.spillane, kesu, sammc, vwade@adaptcentre.ie

Text has become a medium for practically synchronous interaction. Historically, written messages were asynchronous and did not approach the fine-grained interaction and collaboration of talk. Dialog systems model spoken or written synchronous/near-synchronous interactions, often to fulfill a task but increasingly to create the illusion of social interaction. With live text exchange a part of everyday life we have seen an explosion of casual writing, performed not for a formal purpose but to fulfill social goals.

Existing dialog act annotation schemes are often quite task-based. The ISO standard is very useful as it (i) amalgamates contributions from pre-existing schemes, and (ii) is multifunctional and multidimensional - several acts can apply to stretches within the same contribution. Most schemes include a number of social obligation management functions. In a survey of 14 schemes, Petukova found that 10 included greeting functions, 4 included introductions, 6 had goodbyes, 5 included apology type functions, and 5 contained thanking. The Social Obligations Management dimension of the ISO standard contains nine communicative functions.

A corpus of 187 dialogs was collected and annotated with the ISO standard to provide training data for the ADELE project, a personalized intelligent companion. The dialogs were text-based and dyadic via a web-based interface. Each participant was given a persona with information on home, relationships, nationality, job, hobbies and interests, and instructed to discover this information about the interlocutor and also to discover any facts or interests in common. The corpus contained examples of greeting and leave-taking and casual talk for practically all of the conversations gathered. There were 37 participants (26M/11F, age 18-43), either native English speakers or meeting IELTS level 6.5. During a pilot annotation, annotators noted components in extended greeting/introductions and leavetaking (henceforth GIL) sequences which could not easily be annotated. Additional acts were created to more easily mark such sequences and similarly problematic sequences in leave-taking. The GIL sequences in 187 conversations were annotated and then analysed.

Greeting sections were marked as beginning with the first utterance of the conversation, and ending with the last production of a formulaic greeting/introduction or greeting/introduction response. Leave-taking sequences were marked from the first attempt to close the conversation to the final utterance of the conversation. The data contained 9231 turns or 'utterances' where a turn was defined as the text entered before a user pressed return. The vast bulk of utterances were tagged with a single label (7811, 84.7%), 1209 (13%) had two tags, 181 (2%) had three tags, while 26 (0.3%) and 3 utterances had four and five tags.

Of 10889 dialog act tags, 2336 or 21.5% were included in GIL sequences. 1329 tags related to greeting and 1007 to leave-taking. GIL sequences sometimes contained other acts unrelated to greeting, introduction, or leave-taking. The number of dialog acts directly involved in GIL sequences was calculated by disregarding such 'interloping' acts. Greeting/introduction alone accounted for 1034 labels, while leave-taking alone accounted for 786 labels, making a total of 1820 acts of greeting/introduction and leave-taking, or 16.7% of all dialog acts tagged in the corpus. The leave-taking totals include 194 Leave-taking Introductions – utterances which introduce the closure of the dialog. These utterances could be included in the Discourse Structuring dimension, in which case the total for GIL drops to 1626 or 15% of all dialog act labels, which is the most conservative estimate of the proportion of GIL tags in the corpus. The total SOM acts in the corpus including SOM categories outside GIL from the ISO standard amounts to 1824 or 17%. In terms of the prevalence of the new greeting tags, in 187 conversations the hay (How are you?) tag appeared 68 times, the ntmy (Nice to meet you) tag appeared 101 times, and the extra greet tag appeared 66 times (each conversation contained two initialGreets). The response tags repHay and repNtmy appeared less frequently, with 49 instances of repHay and 25 of repNtmy. For the leavetaking tags, there were 139 wntmy (It was nice to meet you) tags and 47 repWntmy tags.

There is a high proportion of SOM acts in the ADELE corpus, and GIL acts contribute greatly to this total. Petukova reports SOM acts in task-based corpora as ranging from 0.5 to 7.8% of total dialog acts, compared with 17% in the ADELE corpus. Most SOM in ADELE are greetings/introductions and leavetaking. Increasing interest in friendly interfaces strengthens the need for greater understanding and more accurate modelling of social dialogue. There are large areas of such dialogue which are not well represented in dialog annotation schemes, ranging from simple politeness formulae, such as the greeting and leavetaking acts treated here, to larger concerns of how to represent the relationship building and maintenance functions integral to casual social talk.

Introducing ADELE: A Personalized Intelligent Companion

Brendan Spillane
ADAPT Centre,
Trinity College Dublin,
Ireland
brendan.spillane@adaptcentre.ie

Emer Gilmartin
ADAPT Centre,
Trinity College Dublin,
Ireland
gilmare@tcd.ie

Christian Saam
ADAPT Centre,
Trinity College Dublin,
Ireland
christian.saam@adaptcentre.ie

Leigh Clark
ADAPT Centre,
University College Dublin,
Ireland
leigh.clark@ucd.ie

Benjamin R. Cowan
ADAPT Centre,
University College Dublin,
Ireland
benjamin.cowan@ucd.ie

Vincent Wade
ADAPT Centre,
Trinity College Dublin,
Ireland
vincent.wade@adaptcentre.ie

ABSTRACT

ADELE is a Personalized Intelligent Companion designed to engage with users such as the elderly through spoken social dialog to help them explore topics of interest, update them about news and events, and provide exercise coaching, companionable chat, and healthcare monitoring. Personalised conversational dialogue agents create an opportunity for both deeper user engagement and companionship for the elderly or those in need of care. ADELE is being designed to provide entertainment and social companionship through naturalistic conversation, while recommending activities such as memory games or physical exercise. It will also be capable of monitoring medication and wellbeing. The system will maintain a user model of information consumption habits and preferences in order to (1) personalize the user's experience for ongoing interactions, and (2) build the user-machine relationship to model that of a friendly companion.

The ADELE Personalized Intelligent Companion is a virtual agent which will be capable of engaged, yet natural and informed, casual conversation. It is being designed to assist a user, not only in small tasks that the user initiates, but in more prolonged dialogue to inform the user of news events and other information of interest through conversation, using a mix of user and agent-initiated interaction.
To achieve these aims the companion will engage the user in conversations that employ a natural mix of linguistic and paralinguistic devices to give and seek information but also to entertain. These may carry elements of task execution but will to a large extent comprise social talk. Accordingly, topic, style and register must be varied at levels ranging from lexical and syntactic to socio- and dialectal pronunciation, to tone of voice, speed and rhythm.

ADELE is intended to converse with the user on topics relating to their personal interests, current events and social feeds, while in turn learning about the user. With repeated interaction, the system will build a profile of the user's interests, and how the user wants to receive information in order to provide editorialized summaries of news articles personalized to the user's consumption preferences in terms of source, granularity, depth, and detail.

Progress on Lip-Reading Sentences

George Sterpu, Christian Saam, Naomi Harte

Sigmedia, ADAPT Centre, School of Engineering, Trinity College Dublin, Ireland

sterpug@tcd.ie, saamc@scss.tcd.ie, nharte@tcd.ie

Abstract

In spite of the grown interest for lip-reading, recent efforts have still largely focused on isolated words or trivial phrases, and, with the exception of [1], not on full complex sentences which may offer sufficient temporal context to accurately decode visual speech. We explore recent advancements in the area of Machine Learning for image processing and sequence-to-sequence mapping, evaluating them on a sentence-level lip-reading task. This includes visual front-ends consisting of Convolutional Neural Networks without and with residual connections, Sequence to Sequence models, and several structural and optimisation improvements. We report results on the TCD-TIMIT dataset [2], consisting of a main set of 59 volunteers saying 98 sentences from the scripts of the TIMIT corpus and amounting to approximately 7 hours of high definition audio-visual recordings. We compare our results against traditional approaches based on hand-crafted visual features and Hidden Markov Models, seeing an improvement of more than 100% overall. An analysis done at the viseme level reveals improvements for most of the visual units, demonstrating the efficiency of neural networks even when the training data is not abundant.

Index Terms: Lip-reading, Sequence to Sequence Recurrent Neural Networks, TCD-TIMIT

1. Results

Table 1: *Lip-reading accuracy on TCD-TIMIT. The right column shows the number of epochs needed to reach convergence (or* nc *for* no convergence*).*

Feature	Accuracy	Epochs
A. DCT + HMM baseline [3]	31.59 %	-
B. AAM + HMM baseline [3]	25.28 %	-
C. Eigenlips + DNN-HMM [4]	46.61 %	-
D. zeros + LSTMs	45.87 %	160
E. DCT + LSTMs	61.52 %	250
F. DCT + BiLSTMs	60.72 %	180
G. **E** w/o attention	48.29 %	270
H. **E** w/ monotonic attention	61.58 %	170
I. DCT + joint CTC-Seq2seq	61.18 %	180
J. 2D CNN + LSTMs		*nc*
K. 2D CNN + BiLSTMs	66.27 %	400
L. **J** on RGB + joint CTC-Seq2seq	66.20%	150
M. **J** on 64x64 + joint CTC-Seq2seq		*nc*
N. Gray 3D CNN + LSTMs		*nc*
O. 2D CNN + joint CTC-Seq2seq	64.61%	260
P. 2D CNN ResNet + LSTMs	**71.21**%	240

Table 2: *Viseme accuracy of the best DNN system (**P**) and relative change from HMM baseline (**A**). Visemes sorted by decreasing visibility.*

Viseme	TIMIT Phonemes	Accuracy P [%]	Δ Accuracy P - A [%]
Lips to teeth	/f/ /v/	94.4	33.71
Lips puckered	/er/ /ow/ /r/ /q/ /w/ /uh/ /uw/ /axr/ /ux/	89.4	61.66
Lips together	/b/ /p/ /m/ /em/	97.1	33.56
Lips relaxed moderate opening to lips narrow-puckered	/aw/	55.1	51.79
Tongue between teeth	/dh/ /th/	71.6	56.67
Lips forward	/ch/ /jh/ /sh/ /zh/	78.5	41.95
Lips rounded	/oy/ /ao/	31.5	-8.70
Teeth approximated	/s/ /z/	90.7	69.22
Lips relaxed narrow opening	/aa/ /ae/ /ah/ /ay/ /ey/ /ih/ /iy/ /y/ /eh/ /ax-h/ /ax/ /ix/	96.1	74.41
Tongue up or down	/d/ /l/ /n/ /t/ /el/ /nx/ /en/ /dx/	89.8	65.38
Tongue back	/g/ /k/ /ng/ /eng/	62.5	23.03
Silence	/sil/ /pcl/ /tcl/ /kcl/ /bcl/ /dcl/ /gcl/ /h#/ /#h/ /pau/ /epi/	94.0	0.64

2. References

[1] J. Son Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[2] N. Harte and E. Gillen, "TCD-TIMIT: An audio-visual corpus of continuous speech," *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 603–615, May 2015.

[3] G. Sterpu and N. Harte, "Towards lipreading sentences using active appearance models," in *AVSP*, Stockholm, Sweden, August 2017.

[4] K. Thangthai, H. L. Bear, and R. Harvey, "Comparing phonemes and visemes with dnn-based lipreading," in *Workshop on Lip-Reading using deep learning methods*, ser. BMVC 2017, 2017.

Supported by the ADAPT Centre for Digital Content Technology, which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

Poster Session C

1	Christopher J. Pidcock, Blaise Potard, Matthew P. Aylett	Creating a New JFK Speech 55 Years Later	CereProc Ltd., Edinburgh
2	Mark Huckvale, András Beke & Iya Whiteley	Longitudinal study of voice reveals mood changes of cosmonauts on a 500 day simulated mission to Mars	University College London
3	Gerardo Roa and Jon Barker	"Automatic Speech Recognition in Music using ACOMUS Musical Corpus"	University of Sheffield
4	Catherine Lai and Gabriel Murray	Predicting Group Satisfaction in Meeting Discussions	University of Edinburgh
5	Mengjie Qian, Xizi Wei, Peter Jancovic, Martin Russell	The University of Birmingham 2018 Spoken CALL Shared Task Systems	University of Birmingham
6	"Ailbhe Ní Chasaide, Neasa Ní Chiaráin, Harald Berthelsen, Christoph Wendler, Andrew Murphy, Emily Barnes, Irena Yanushevskaya, Christer Gobl"	Speech technology and resources for Irish: the ABAIR initiative	Trinity College Dublin
7	Emer Gilmartin, Carl Vogel, Nick Campbell, VincentWade	Chats and Chunks: Annotation and Analysis ofMultiparty Long Casual Conversations	Trinity College Dublin
8	Emma O'Neill, Mark Kane, and Julie Carson-Berndsen	Two Data-Driven Perspectives on Phonetic Similarity	University College Dublin
9	Brendan Spillane, Emer Gilmartin, Christian Saam, Leigh Clark, Benjamin R. Cowan, Vincent Wade	Identifying Topic Shift and Topic Shading in Switchboard	Trinity College Dublin
10	Andrea Carmantini, Simon Vandieken, Alberto Abad, Julie-Anne Meaney, Peter Bell, Steve Renals	Automatic speech recognition for cross-lingual information retrieval in the IARPA MATERIAL programme	University of Edinburgh
11	Felipe Espic and Simon King	The Softmax Postfilter for Statistical Parametric Speech Synthesis	University of Edinburgh
12	Joao P. Cabral	Estimation of the assymetry parameter of the glottal flow waveform using the Electroglottographic signal	Trinity College Dublin
13	Jason Taylor, Korin Richmond	Combilex G2P with OpenNMT	University of Edinburgh

Creating a New JFK Speech 55 Years Later

Christopher J. Pidcock, Blaise Potard, Matthew P. Aylett

CereProc Ltd., Edinburgh, UK

{chris,blaise,matthewa}@cereproc.com

Abstract

Text-to-speech voice creation projects typically use data specifically recorded for the purpose. Recording scripts are crafted to cover multiple phonetic contexts, and data is collected in high quality, low-reverberant environments. Audio recordings from the 1950s and 1960s do not fit this model, with varying quality based on underlying analogue hardware. Additionally Presidential speeches are often recorded in reverberant and noisy environments. To create a new speech from this data requires additional processing for quality analysis, noise reduction, and audio environment smoothing. Additionally a DNN-based prosodic model was implemented to model JFKs unique speaking style. Post processing to partially restore the original environment was found to improve the perceived quality.

Figure 1: *Photo portrait of John F. Kennedy, President of the United States. This image is a work of an employee of the Executive Office of the President of the United States, taken or made as part of that person's official duties. As a work of the U.S. federal government, the image is in the public domain.*

The speech audio can be found at https://s3-eu-west-1.amazonaws.com/cereproc/jfk_speech_180326.wav, with the text at https://www.jfklibrary.org/Research/Research-Aids/JFK-Speeches/Dallas-TX-Trade-Mart-Undelivered_19631122.aspx. This project was commissioned by The Times newspaper and Rothco, part of Accenture Interactive.

Longitudinal study of voice reveals mood changes of cosmonauts on a 500 day simulated mission to Mars

Mark Huckvale[1], András Beke[1] & Iya Whiteley[2]
[1]Speech, Hearing and Phonetic Sciences, University College London
[2]Centre for Space Medicine, University College London

Long duration space missions beyond Earth orbit will create considerable psychological stresses on future space flight crews and explorers, and it will be essential to find ways to monitor their mental health even when they are beyond real-time communication with mission control. In this paper we propose that changes to the voice might be used as an index into the effects of the mission on the psychological state of the crew. Speech is a useful indicator in this situation since it is an activity that engages both the physiological and psychological systems of the speaker and is easy to collect as part of everyday working activities.

In this study we analysed recordings made by the crew on the Mars500 simulated mission to Mars that took place between 2010-2011 in Moscow under the supervision of IBMP and ESA [1]. In total over 2400 voice mail messages made by the six members of the crew between mission days 60 and 500 were studied. We showed that significant changes occurred in both mean fundamental frequency (F0) and fundamental frequency range of the speakers over that period. A general additive model (GAM) was used to study the effects of time-of-day and day-of-mission on speaker F0. While differences were observed across speakers, some common patterns emerged. Speakers demonstrated a period of adaptation in the first few months in which their F0 reduced as they became settled to their work; they showed an increase in mean F0 over the period of the simulated Mars landing in the middle of the mission as they achieved their mission objective; and they demonstrated a large fall in mean F0 in the third quarter of the mission, corresponding to a period of low mood where the main aims of the mission had been fulfilled and all that remained was the long journey home. These changes in mean F0 correlate both with our understanding of how psychological arousal and mood affects the voice and with the testimony of the cosmonauts themselves after the mission [2]. Interestingly, voice analysis seems to detect the changes in psychological health with at least as good detail as reported in other Mars500 studies of changing sleep patterns and changing hormone levels of the crew. This was the case even though the recordings were not made with this kind of analysis in mind. We conclude that voice analysis is promising as an additional means for monitoring psychological health in long term space missions.

[1] Grigoriev, I. Ushakov, B. Morukov, B.V., First results of the Mars-500 international megaexperiment. Pilotiruemye Polety Kosmos, 1 (2012) 5-14.
[2] Šolcová, I. Šolcová, I. Stuchlíková, Y. Mazehóová, The story of 520 days on a simulated flight to Mars. Acta Astronautica 126 (2016) 178-189.

Automatic Speech Recognition in Music using ACOMUS Musical Corpus

Gerardo Roa, Jon Barker

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

Abstract

This research considers the task of automatically transcribing song lyrics from an audio recording. This challenging task is made more difficult by the lack of suitable data for training recognition systems. Current systems have been trying to address this problem by using spoken speech resources in conjunction with traditional speaker adaptation techniques [1] or by generating training data by the 'singification' of normal speech (i.e., processing existing speech corpora to simulate the characteristics of signing) [2]. Nevertheless, the results obtained with these approaches have been poor, with accuracies only sufficient to support certain keyword based music information retrieval applications. As a first step toward building a high performance transcription system, we have been constructing a novel audio corpus from acoustic covers versions of popular songs performed by amateur artists. This corpus, called ACOMUS, is built from a collection of performances that have been sourced from YouTube. The corpus currently includes 120 songs with a solo singer and acoustic guitar or piano accompaniment. The audio track of each video has been carefully annotated to provide ground truth start and end times and word sequence transcriptions for each sung phrase. Baseline recognition systems have then been built using state-of-art HMM-DNN ASR systems trained and evaluated using the Kaldi toolkit [3]. Some early training data augmentation experiments have been conducted using pitch and tempo modification [4]. Current works are focused on increasing the size of the ACOMUS corpus, developing representations that are better suited to the singing voice and evaluate blind source separation and speech enhancement techniques. The poster will present some initial results and a presentation of the immediate research plans.

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] Kruspe, A.M. (2016) "Retrieval of Textual Song Lyrics from Sung Inputs". In *Interspeech 2016*, 2140-2144.

[3] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Vesely, K. (2011). "The Kaldi Speech Recognition Toolkit". In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society.

[4] Roa, G. (2016). "Automatic Speech Recognition in Music". Unpublished master's dissertation, University of Sheffield, UK.

Predicting Group Satisfaction in Meeting Discussions

Catherine Lai[1], *Gabriel Murray*[2]

[1]The Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK
[2]University of the Fraser Valley, Canada

clai@inf.ed.ac.uk, gabriel.murray@ufv.ca

Abstract

We address the task of automatically predicting group satisfaction in meetings using acoustic, lexical, and turn-taking features. Participant satisfaction is measured using individual post-meeting questionnaires from the AMI corpus. We focus on predicting two aspects of satisfaction: overall participant satisfaction with the meeting, and more specifically whether participants felt that everyone received sufficient attention during the meeting. All predictions are made at the aggregated group level. In general, we find that combining features across modalities improves prediction performance. In particular, acoustic and lexical models benefit from the addition of turn-taking features. Our experiments also show how data-driven methods can be used to explore how different facets of group satisfaction are expressed through different modalities. For example, inclusion of prosodic features improves prediction of feeling sufficient attention but hinders prediction of overall satisfaction, and vice-versa for lexical features. Moreover, feelings of sufficient attention were better reflected by acoustic features than by speaking time. Feature ablation experiments indicate that more abstract lexical features were helpful for predicting overall satisfaction. Thus, a greater focus on extracting affective lexical content from meeting interactions appears warranted for this task, as does further examination of potential interactions between features from different modalities in expressing participant affect. Overall, this study indicates that group dynamics can be revealed as much by *how* participants speak, as by what they say.

Q7: Overall Satisfaction			
Feature set	RFR	SVR	BRR
turn	6.51	6.54	6.36
acoustic	6.39	7.23	6.70
lex	6.67	7.05	6.76
acoustic+lex	6.19	6.95	6.49
acoustic+turn	**6.16**	6.64	6.34
lex+turn	6.40	**6.53**	***6.10***
acoustic+lex+turn	6.23	6.78	6.23

Q16: Attention satisfaction			
Feature set	RFR	SVR	BRR
turn	8.72	9.03	8.84
acoustic	7.20	7.05	6.85
lex	7.68	7.95	7.74
acoustic+lex	6.94	7.38	6.90
acoustic+turn	**6.93**	**7.01**	***6.65***
lex+turn	7.88	7.82	7.43
acoustic+lex+turn	7.13	7.40	6.76

Table 1: *Mean Squared Error results for Q7: 'All in all, I am very satisfied' (left) and Q16: 'All team members received sufficient attention' (right) for Random Forest Regression (RFR), Support Vector Regression (SVR), and Bayesian Ridge Regression (BRR) models. Mean Score Baseline: (Q7) 7.08, (Q16) 9.06.*

Ablation modality	Q7	Q16
turn	6.23	6.70
acoustic	5.89	6.73
lex	6.05	6.60
all modalities	***5.57***	***6.52***

Table 2: *Results after removing feature sets that caused decreased performance in the ablation studies.*

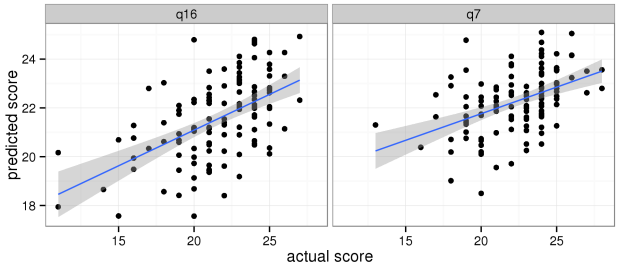


Figure 1: *Scatterplots of group scores vs best model predictions with linear model fit.*

The University of Birmingham 2018 Spoken CALL Shared Task Systems

Mengjie Qian, Xizi Wei, Peter Jančovič, Martin Russell
Schoool of Engineering, University of Birmingham, Birmingham B15 2TT, UK
{mxq486, xxw395, p.jancovic, m.j.russell}@bham.ac.uk

Abstract

Shared tasks have been a major factor in the development of many areas of speech and language technology. Following the success of the first edition of the Spoken CALL Shared Task challenge [1], a second edition was introduced in 2018. In this work we present the systems developed by the University of Birmingham for the 2018 CALL Shared Task (ST) challenge. The task is to perform automatic assessment of grammatical and linguistic aspects of English spoken by German-speaking Swiss teenagers. Our developed systems consist of two components, automatic speech recognition (ASR) and text processing (TP). We explore several ways of building a DNN-HMM ASR system using out-of-domain AMI [2] speech corpus plus a limited amount of ST data. In development experiments on the initial ST data, our final ASR system achieved the word-error-rate (WER) of 12.00%, compared to 14.89% for the official ST baseline DNN-HMM system. The WER of 9.28% was achieved on the test set data. For TP component, we first post-process the ASR output to deal with hesitations and then pass this to a template-based grammar, which we expanded from the provided baseline. We also developed a TP system based on machine learning methods using word2vec embeddings [3, 4], which enables to better accommodate variability of spoken language. In one of our submissions, a linear logistic regression was used to fuse outputs from several systems. The results of our submissions are shown in Table 1 and Table 2 shows the results we achieved after the challenge.

Table 1: Results of submissions for the 2018 Spoken CALL Shared Task challenge with different ASR components.

Submission	Evaluation measure		
	F-measure	D	D_{full}
Submission 1	0.915	10.714	5.778
Submission 2	0.904	8.804	4.958
Submission 3	0.914	10.764	5.691

Table 2: Results of ML-based TP systems employing different classifiers.

Classifier	Evaluation measure		
	F-measure	D	D_{full}
LDA	0.88	9.767	4.136
logReg (PCA)	0.884	10.263	4.281
SVM (PCA)	0.891	10.939	4.616
NN	0.928	12.716	7.101

References

[1] M. Qian, X. Wei, P. Jančovič, and M. Russell, "The University of Birmingham 2017 SLaTE Share CALL Share Task systems," in *7th ISCA Workshop on Speech and Language Technology in Education, Stockholm, Sweden*, pages 91 -96, 2017.

[2] J. Carletta, et al., "The AMI meeting corpus: A pre-announcement," in *Int. Workshop on Machine Learning for Multimodal Interaction.* Springer, 2005, pp. 28-39.

[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781* (2013).

[4] T. Mikolov, I. Sutskever, et al., "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems,* 2013.

Speech technology and resources for Irish: the ABAIR initiative

Ailbhe Ní Chasaide, Neasa Ní Chiaráin, Harald Berthelsen, Christoph Wendler, Andrew Murphy, Emily Barnes, Irena Yanushevskaya, Christer Gobl

Phonetics and Speech Lab, CLCS, Trinity College Dublin, Ireland

anichsid@tcd.ie, nichiarn@tcd.ie, bertheslen@tcd.ie, wendlec@tcd.ie

Abstract

An overview is presented of the ongoing research of the ABAIR initiative, which aims at the provision of speech/linguistic resources and technologies for Irish. A multidisciplinary approach focusses on three parallel research strands, to ensure that the resources and technologies developed are harnessed for the language user in a way that supports the maintenance of Irish. It is hoped that many aspects of the work can serve as a model for other endangered languages [1]. The strands of ABAIR research include:

(1) Provision of speech and linguistic-phonetic resources. This covers different types of activities, including the recording and processing of speech corpora, carrying out linguistic-phonetic analyses, which provide modules for use in the development of speech technology (2 below) and for deployment in targeted applications (3 below). One challenge, common to minority languages is the fact that, although there is a written standard, there no spoken 'prestige variety' to serve as a target for speech technology. A further challenge is that the sound system and writing system differ considerably from English, in ways that have to be taken into account in text processing for technology. Dialect diversity is taken into account, and so, for example pronunciation lexica, letter to sound rules, prosodic models are being developed for the main dialects. Prosody modelling is being carried out with a view to capturing both the linguistic aspects of prosody, e.g. stress rules, question intonation etc., which vary across dialects [2.3], and the paralinguistic prosody which expresses emotion and attitude [4]. These are needed for interactive and dialogue-based applications, see (3) below.

(2) Core Technologies for the dialects of Irish. To date multi-dialect synthesis (three main dialects) is available at ***www.abair.ie*** and further, more endangered dialects and children's voices are targeted. The website is being accessed by large numbers worldwide, and provides linguistic resources such as the automatic phonetic transcription of text. Future developments envisaged include speech recognition, and dialogue systems for specific purposes.

(3) This strand exploits the outputs of both (1) and (2) to develop applications for Irish-language users, and it is through these that the impact of the research is felt. A web-browser plug-in allows easy integration of Irish synthesis (with choice of dialect) into websites. Applications target particularly language learning and disability/access. A screen-reader for the blind, with simultaneous Braill output allows speech output with speed control as required. Synthesis-based interactive applications being piloted for language learners involve multimodal games, virtual worlds and an embryonic dialogue system (the latter, a talking monkey, who for the moment, responds in speech to text input). Evaluation of these games elicit feedback not only on their educational effectiveness, but also on the quality of the synthesis [5]. A current collaboration with IT Carlow, involves development of an interactive game to train awareness of the sound contrasts of the language, a prerequisite to literacy acquisition. It is intended as the first in a series of pronunciation and literacy tools which, although designed for all, will particularly impact dyslexic learners. Linguistic/phonetic resources underpin educational applications, and will facilitate future intelligent ICALL applications, integrating linguistic knowledge to guide the learner's interaction, and allowing adaptation to the learner's level of language [6].

References

[1] Ní Chasaide, A., Ní Chiaráin, N., Berthelsen, H., Wendler, C., Murphy. A. (2015). Speech technology as documentation for endangered language preservation: the case of Irish. *Proceedings of ICPhS*, Glasgow

[2] O'Reilly, M. and Ní Chasaide, A. (2016). Modelling the timing and scaling of nuclear pitch accents of Connaught and Ulster Irish with the Fujisaki model of intonation. *Proceedings Speech Prosody 8*, Boston, pp. 355–359.

[3] Dorn, A. and Ní Chasaide, A. (2016). Donegal Irish rises: Similarities and differences to rises in English varieties. *Proceedings of Speech Prosody 8*, Boston, pp. 163–167.

[4] Ní Chasaide, A., Yanushevskaya, I. and Gobl, C. (2017). Voice-to-affect mapping: inferences on language voice baseline settings. *INTERSPEECH 2017*, Stockholm, Sweden, pp. 1258-1262.

[5] Ní Chiaráin, N. and Ní Chasaide, A. (2016). Chatbot Technology with Synthetic Voices in the Acquisition of an Endangered Language: Motivation, Development and Evaluation of a Platform for Irish. *Proceedings of LREC 2016*, Portorož, Slovenia.

[6] Ní Chiaráin, N. and Ní Chasaide, A. (2016). Faking Intelligent CALL: the Irish context and the road ahead. *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC*, Umeå, Sweden.

Chats and Chunks: Annotation and Analysis of Multiparty Long Casual Conversations

Emer Gilmartin[1], *Carl Vogel*[2], *Nick Campbell*[1], *Vincent Wade*[1]

[1]ADAPT Centre, School of Computer Science and Statistics, Trinity College Dublin
[2]Computational Linguistics Group, School of Computer Science and Statistics, Trinity College Dublin

gilmare@tcd.ie, vogel@tcd.ie, nick@tcd.ie, vwade@adaptcentre.ie

Abstract

Casual talk or social conversation is a fundamental form of spoken interaction. Corpora of casual talk often comprise relatively short dyadic conversations, although research into such talk has found longer multiparty interaction to be very common. This genre of spoken interaction is attracting more interest with attempts to build more friendly and natural spoken dialog systems. To study longer multiparty casual talk, we have assembled a collection of conversations from three existing corpora. Casual conversation is not monolithic but rather comprises phases of interactive 'chat' and more monologic 'chunks' where one speaker dominates the conversation. We describe the annotation of structural chat and chunk phases in these conversations, and statistical analysis of the characteristics of these phases. We review our preliminary results, noting significant differences in the distribution of overlap, laughter and disfluency in chat and chunk, and finding that chunk dominates as conversations get longer. We also outline our continuing work on the structure and prosody of such conversations, investigating the distribution of chat and chunk phases and phrase final pitch movements for between- and within- speaker silence and overlaps. We conclude with discussion of how greater understanding of this genre of spoken interaction could aid the design of spoken dialog systems.

Index Terms: multiparty dialogue, human-computer interaction, turntaking

Figure 1: *Examples of chat (top) and chunk (bottom) phases in two stretches from the 5-party conversation analysed in this work. Each row denotes the activity of one speaker across 120 seconds. Speech is green, and laughter is yellow on a grey background (silence).The chat frame, taken at the beginning of the conversation, can be seen to involve shorter contributions from all participants with frequent laughter. The chunk frame shows longer single speaker stretches.*

1. References

[1] E. Ventola, "The structure of casual conversation in English," *Journal of Pragmatics*, vol. 3, no. 3, pp. 267–298, 1979.

[2] S. Eggins and D. Slade, *Analysing casual conversation*. Equinox Publishing Ltd., 2004.

[3] E. Gilmartin, F. Bonin, L. Cerrato, C. Vogel, and N. Campbell, "What's the game and who's got the ball? Genre in spoken interaction," 2014.

Two Data-Driven Perspectives on Phonetic Similarity

Emma O'Neill[1], Mark Kane[2], and Julie Carson-Berndsen[1]

[1]ADAPT Centre, School of Computer Science, University College Dublin, Ireland
[2]Daon Ireland

Abstract

We present two very different data-driven perspectives on phonetic similarity, first focusing on similarities in speech data generated by substitution and then considering similarities from a distributional standpoint derived from text. In [3], a method to enhance standard phone confusion matrices generated during speech recognition (on the TIMIT training data set [2]) was proposed. This approach involved iteratively removing one acoustic phone model at a time during recognition to identify which phones were recognised in place of the missing phone. The resultant confusions were incorporated into an enhanced matrix thus reducing sparseness and capturing additional degrees of phonetic similarity in a purely data-driven manner. The enhanced matrix was seen to perform better on a spoken term detection task using different corpora and speech recognition methods. The difference between the standard and enhanced matrices are visualised in Figures 1 and 2 respectively using hierarchical clustering[1]. The dendrogram in Figure 2 can be seen to clearly capture phonetic similarities not found in the standard matrix.

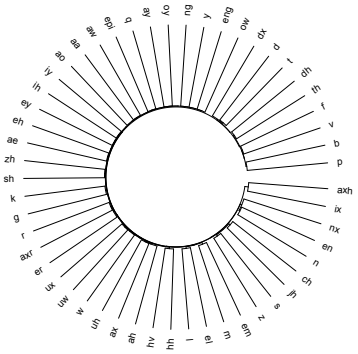


Figure 1: Standard Confusions

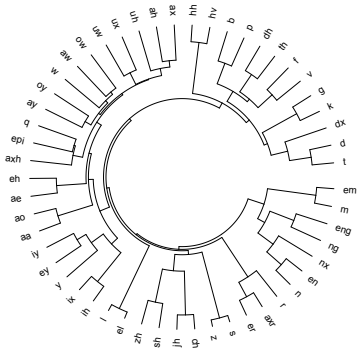


Figure 2: Enhanced Confusions

Phonetic similarities can also be considered from the distributional perspective and using text rather than speech as the starting point. Here we used the CMU Pronouncing Dictionary to obtain phonetic representations of 2.38M syllable tokens in the Brown Corpus [1] and applied the *w*ord2vec embedding technique [4] at the level of phones. Figures 3 and 4 depict the resultant clustering of phones encompassing similarities based on their distribution within the syllables and the syllable onsets. The poster will provide further details on how the similarities are generated, how they may complement each other and how they can be evaluated.

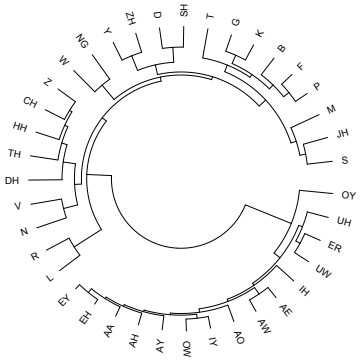


Figure 3: All syllables in Brown Corpus

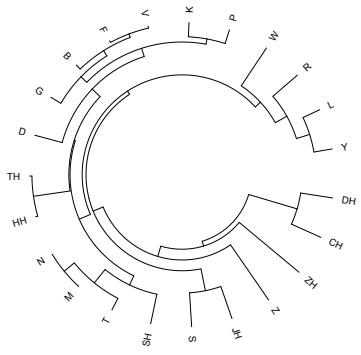


Figure 4: All onsets in Brown Corpus

References

[1] W. N. Francis and H. Kučera. *A Standard Corpus of Present-Day Edited American English, for use with Digital Computers (Brown)*. Brown University, Providence, Rhode Island, 1964.

[2] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. DARPA TIMIT acoustic phonetic continuous speech corpus CDROM, 1993.

[3] M. Kane and J. Carson-Berndsen. Enhancing data-driven phone confusions using restricted recognition. In *Proceedings of Interspeech 2016*, 2016.

[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[1]We use the R dendextend package https://cran.r-project.org/web/packages/dendextend/vignettes/introduction.html

Identifying Topic Shift and Topic Shading in Switchboard

Brendan Spillane
ADAPT Centre,
Trinity College Dublin,
Ireland
brendan.spillane@adaptcentre.ie

Emer Gilmartin
ADAPT Centre,
Trinity College Dublin,
Ireland
gilmare@tcd.ie

Christian Saam
ADAPT Centre,
Trinity College Dublin,
Ireland
christian.saam@adaptcentre.ie

Leigh Clark
ADAPT Centre,
University College Dublin,
Ireland
leigh.clark@ucd.ie

Benjamin R. Cowan
ADAPT Centre,
University College Dublin,
Ireland
benjamin.cowan@ucd.ie

Vincent Wade
ADAPT Centre,
Trinity College Dublin,
Ireland
vincent.wade@adaptcentre.ie

ABSTRACT

This paper highlights some of the ongoing work on the ADELE project, namely the identification and annotation of topic shift and topic shading in the Switchboard-1 Release-2 corpus. The purpose of this is to train an Artificial Neural Network to create a digital companion for the elderly that can communicate through informal, yet informed social dialogue, on a variety of topics of interest to a user over a prolonged time scale. To this end the project is focussing on topic shift and shading, the mechanisms which underpin the development of such conversations [6, 8]. In the past, dialogue systems have predominantly focussed on practical tasks due to the complexity of modelling realistic everyday social talk [1]. With increasing awareness of the need for home robots and virtual home care agents to help assist in the provision of care for a rapidly ageing population, it is necessary to develop a more caring, involved, and personalised virtual care agent capable of such social dialogue.

In any social conversation, the topic being discussed changes based on a host of factors. As such, it is nigh on impossible to predict the path of a normal conversation between two people. It is this random, yet mostly smooth transition, from one topic to another that often differentiates real conversation between two people and that between a person and an artificial agent. In the past, many systems bounded any possible interaction by limiting the range of possible responses that could be recognised by the agent, e.g. limited speech or buttons. Such approaches meant that the agent did not have to deal with the variability of a normal conversation, but this also limited the naturalness of the interaction.

However, to properly interact with a person in a more natural conversational form, ADELE must be able to identify, strategise, render, and initiate topic shift and topic shading in a conversation. There are several reasons for this. Firstly, it will allow ADELE to be able to change the course of a conversation, thus limiting awkward disjunctions. Secondly, it will enable ADELE to more easily and more naturally follow a conversation strategy, such as reminding or prompting the user to take medication or engage in exercise. Thirdly, it will enable ADELE to form more natural dialogue. Fourthly, it will be able to better identify topic shift or shading by the user which is a sign they want to reorient the conversation.

Automatic speech recognition for cross-lingual information retrieval in the IARPA MATERIAL programme

Andrea Carmantini[1], *Simon Vandieken*[1], *Alberto Abad*[1,2], *Julie-Anne Meaney*[1], *Peter Bell*[1], *Steve Renals*[1]

[1] Centre for Speech Technology Research, University of Edinburgh, Edinburgh EH8 9AB, UK
[2] Instituto Superior Tecnico, University of Lisbon, Portugal

The IARPA MATERIAL programme [1] seeks to develop methods for searching speech and text in low-resource languages, using English-language queries. Methods must use limited training data and be rapidly deployable to new languages and domains. This presentation will focus on Edinburgh's work on low-resource speech recognition for MATERIAL.

A core challenge for ASR in MATERIAL is the requirement to train systems to operate on multi-genre data – including telephone conversations, news broadcasts and other topical content – with limited quantities of transcribed source-language training data, drawn only from telephone conversations.

We will present an overview of work from the past six months on:

- acoustic modelling using TDNN-LSTM models with lattice-free MMI training
- language modelling from web-crawled text data
- semi-supervised acoustic model training
- speaker and domain adaptation

Speech recognition work was carried out in collaboration with colleagues in the speech group at Cambridge University, and other members of the MATERIAL SCRIPTS team: the Universities of Columbia, Maryland and Yale.

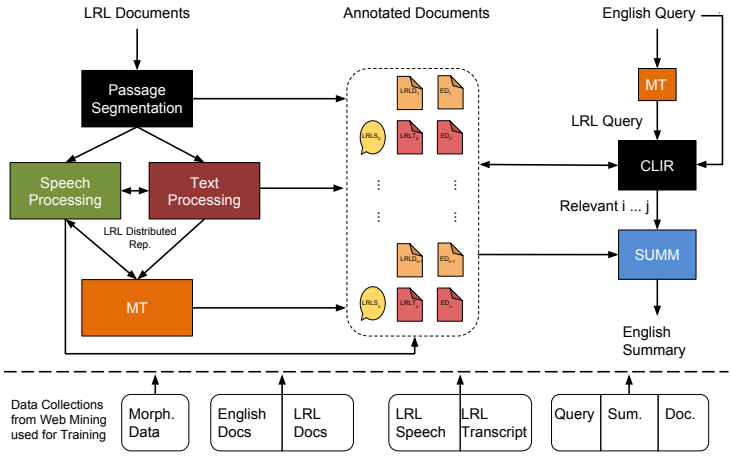


Figure 1: Architecture for the complete SCRIPTS system

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract # FA8650-17-C-9117. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

[1] https://www.iarpa.gov/index.php/research-programs/material

The Softmax Postfilter for Statistical Parametric Speech Synthesis

Felipe Espic and Simon King

The Centre for Speech Technology Research (CSTR), University of Edinburgh, UK

felipe.espic@ed.ac.uk, Simon.King@ed.ac.uk

code and samples:
http://www.felipeespic.com/softmax_pf

Abstract

Recently, new generative models (e.g., *WaveNet*, *Tacotron*, *SampleRNN*) have produced high quality TTS that is comparable to natural speech. However, they still exhibit some drawbacks, especially when implemented for practical use. The large number of required layers involve high computational burden and large footprint, which constrain their applicability especially in embedded systems. Moreover, their controllability is limited as they are almost full end-to-end neural network-based systems, such that it makes very difficult to do fixings that are easily done in more modular architectures. Also, these systems require a large amount of training data to work properly.

On the contrary, Statistical Parametric Speech Synthesis (SPSS) has been proven to be very suitable for constrained applications, such as embedded systems. SPSS architectures are controllable, computationally efficient, whilst generating high quality speech. Furthermore, they can be trained with relatively small databases.

We propose a new postfilter for SPSS based on neural networks, which is inspired by the new generation of generative models, but replacing costly operations by efficient signal processing. For this purpose, *MagPhase* features and vocoder are used as a deterministic representation of speech, and as a waveform generator, respectively. The Figure 1 shows the general diagram of the proposed system, which can be applied to different feature streams, even to speech signals directly.

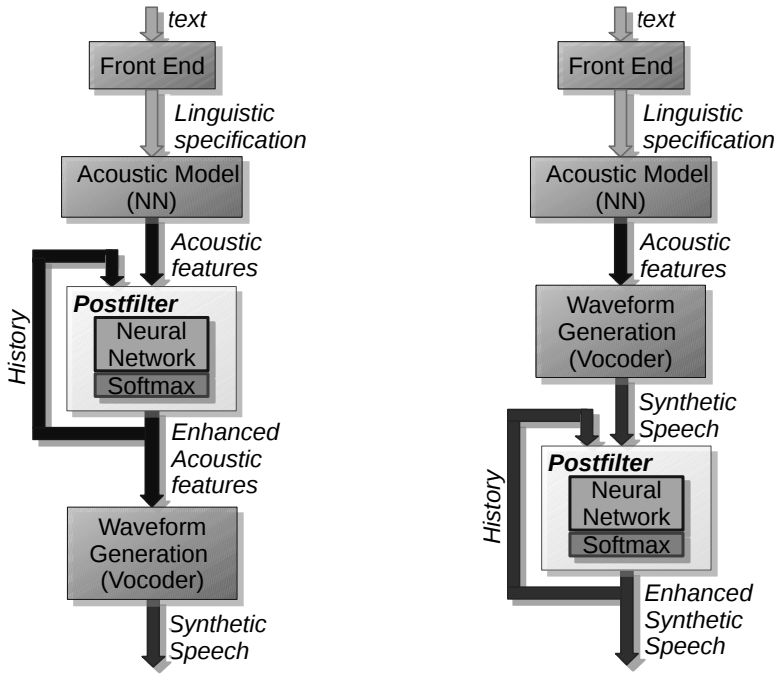


Figure 1. Two alternative configurations for the Softmax Postfilter.

Estimation of the assymetry parameter of the glottal flow waveform using the Electroglottographic signal

João P. Cabral

The ADAPT Research Centre, Trinity College Dublin, Ireland

cabralj@scss.tcd.ie

1. Abstract

Glottal activity information can be very important in several speech processing applications, such as in speech therapy, voice disorder diagnosis, voice transformation and text-to-speech synthesis. However, the use of algorithms for estimating glottal parameters from the speech signal is very limited in those applications because of problems with robustness and accuracy. An alternative way to obtain more accurate and reliable glottal parameter estimates is to use other recording equipment besides the audio microphone. Electroglottography is the most popular non-invasive measurement of vocal fold motion. This paper proposes an automatic method for estimation of the closing quotient parameter of the glottal source from the electroglottographic (EGG) signal that permits to measure an additional parameter related to the assymetry of the glottal flow pulse. This parameter is related to the assymetry of the glottal flow pulse, which is a very important characteristic correlated with voice quality and widely studied in voice source analysis.

The EGG signal permits to accurately estimate two important time instants of the glottal cycle: the glottal opening and closing instants [1, 2, 3, 4, 5, 6, 7, 8, 9]. Although the EGG signal has been widely used to calculate the open phase duration of the glottal pulse (duration between those glottal instants), in general it is not used to obtain the other important time instants of the glottal flow. This work studies the correlation between the EGG signal and an important characteristic of the glottal flow: the assymetry of the glottal pulse shape. For example, the closing quotient (CQ) and speed quotient (SQ) parameters represent the assymetry of the glottal pulse.

By performing measurements of the CQ parameter on the speech and the EGG signals it was found in this work that there is a high correlation of the parameter estimates between the two signals. In addition, this work also proposes a method to robustly estimate the CQ parameter by using measurements performed on the EGG signal.

2. Acknowledgments

This research is supported by the Science Foundation Ireland (Grant 13/RC/2106) as part of the ADAPT centre (www.adaptcentre.ie) at Trinity College Dublin.

3. References

[1] D. G. Childers, "Glottal source modeling for voice conversion," *Speech Communication*, vol. 16, no. 2, pp. 127–138, 1995.

[2] D. Thotappa and S. R. M. Prasanna, "Reference and automatic marking of glottal opening instants using egg signal," in *2014 International Conference on Signal Processing and Communications (SPCOM)*, 2014, pp. 1–5.

[3] M. R. P. Thomas, J. Gudnason, and P. A. Naylor, "Detection of glottal closing and opening instants using an improved dypsa framework," in *2009 17th European Signal Processing Conference*, 2009, pp. 2191–2195.

[4] K. Ramesh and S. R. Prasanna, "Glottal opening instants detection using zero frequency resonator," *Int. J. Speech Technol.*, vol. 20, no. 1, pp. 127–141, 2017.

[5] K. Ramesh, S. R. M. Prasanna, and D. Govind, "Detection of glottal opening instants using hilbert envelope," in *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, 2013, pp. 44–48.

[6] M. R. P. Thomas, J. Gudnason, and P. A. Naylor, "Estimation of glottal closing and opening instants in voiced speech using the yaga algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 82–91, 2012.

[7] A. Bouzid and N. Ellouze, "Voice source parameter measurement based on multi-scale analysis of electroglottographic signal," *Speech Communication*, vol. 51, no. 9, pp. 782 – 792, 2009, special issue on non-linear and conventional speech processing.

[8] N. Sturmel, C. dAlessandro, and B. Doval, "A spectral method for estimation of the voice speed quotient and evaluation using electroglottography," 2006.

[9] R. S. Prasad and B. Yegnanarayana, "Determination of glottal open regions by exploiting changes in the vocal tract system characteristics," *The Journal of the Acoustical Society of America*, vol. 140, no. 1, pp. 666–677, 2016.

Combilex G2P with OpenNMT

Jason Taylor, Korin Richmond

[1]Centre for Speech Technology Research, University of Edinburgh, UK
{jason.taylor, korin.richmond}@ed.ac.uk

Abstract

Predicting the pronunciation of Out-of-Vocabulary (OOV) words using Grapheme-to-Phoneme modelling (G2P) is still commonplace in TTS and LVCSR systems. The current state of the art involves training deep, bi-directional RNN models with LSTM units [1]. Such an architecture allows for the simultaneous prediction of stress markers [2].

We experimented with the training data in OpenNMT G2P models [3]. We used the Received Pronunciation and the General American accents of the Combilex Speech Technology Lexicon [4] with and without stress and prior alignments of graphemes to phonemes. The results showed using OpenNMT with Combilex gave lower Word Error Rates (WER) than with the CMUdict [5] and than a previous Combilex G2P method based on Decision Trees [6]. Future work will study whether accurate pronunciations can be learnt using inverted indices from ASR lattices.

References

[1] K. Rao, F. Peng, H. Sak, and F. Beaufays, "Grapheme-to-phoneme conversion using Long Short-Term Memory recurrent neural networks." [Online]. Available: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/43264.pdf

[2] D. Van Esch, M. Chua, and K. Rao, "Predicting pronunciations with syllabification and stress with recurrent neural networks," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2016.

[3] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush, "OpenNMT: Open-Source Toolkit for Neural Machine Translation." [Online]. Available: https://arxiv.org/pdf/1701.02810.pdf

[4] CSTR, "Combilex," 2018. [Online]. Available: http://www.cstr.ed.ac.uk/research/projects/combilex/

[5] CMU, "The Carnegie Mellon Pronouncing Dictionary," 2018. [Online]. Available: https://github.com/cmusphinx/cmudict

[6] K. Richmond, R. A. J. Clark, and S. Fitt, "Robust LTS rules with the Combilex speech technology lexicon." [Online]. Available: http://www.cstr.ed.ac.uk/downloads/publications/2009/IS090308.pdf