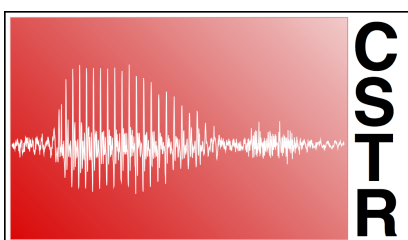




UK Speech Conference

Edinburgh
5–6 September 2022



THE UNIVERSITY *of* EDINBURGH
informatics

Sponsors

The UK Speech workshop is made possible by the kind support of our sponsors who are a very active part of the UK speech community. If you are interested in supporting UK Speech please get in touch with the UK speech committee

Gold Sponsors



Silver Sponsors



Bronze Sponsors



And with support from



UK Speech 2022 Organization

Local Organizers

Peter Bell, *Co-chair*
Catherine Lai, *Co-chair*
Emily Martin, *Event admin*
Sarenne Wallbridge, *Volunteer Coordinator*
Seona Wharrie
Yuanchao Li
Atli Sigurgeirsson
Dan Wells
Emelie Van Der Vreken
Jie Chi
Andrea Carmantini
Christoph Minixhofer
Johannah O'Mahony
Electra Wallington
Sung-Lin Yeh

UK Speech Committee

Thomas Merritt, Amazon
Sébastien Le Maguer, Trinity College Dublin
Simone Graetzer, University of Salford
Catherine Lai, The University of Edinburgh
João Cabral, Trinity College Dublin

Short Program

Monday, 5th September 2022

Time Slot	Session Name
12h00-13h00	Lunch & Registration
13h00-13h15	Welcome message
13h15-15h15	Keynote 1 (Dr. Joanne Cleland) - Chair: Korin Richmond Using Ultrasound to Image the Articulators in the Speech Therapy Clinic
14h15-15h30	break
14h30-15h30	Poster Session A
15h30-16h00	Coffee Break
16h00-17h00	Keynote 2 (Dr. Jennifer Williams) - Chair: Simon King Speech Privacy: Where Are We Going and How to Get There?
18h30	Conference Drinks Reception, Dinner and Ceilidh at the Scottish National Gallery

Tuesday, 6th September 2022

Time Slot	Session Name
09h30-10h30	Oral Session A - Chair: Jon Barker
10h30-11h30	Keynote 3 (Prof. Naomi Harte) - Chair: Kate Knill Multimodal Speech Embracing the Iceberg!
11h30-12h00	Coffee Break
12h00-13h00	Poster Session B
13h00-14h00	Lunch
14h00-15h00	Poster Session C
15h00-16h00	Oral Session B - Chair: Julie Berndsen
16h00-16h15	Closing Ceremony

Long Program

Monday, 5th September 2022

12h00-13h00 - Lunch & Registration

13h00-13h15 - Welcome message

13h15-15h15 - Keynote 1 (chair: Korin Richmond)

Using Ultrasound to Image the Articulators in the Speech Therapy Clinic

Dr. Joanne Cleland

14h15-15h30 - break

14h30-15h30 - Poster Session A

Text-free non-parallel many-to-many voice conversion using normalising flows

Thomas Merritt, Abdelhamid Ezzer, Piotr Biliski, Magdalena Proszewska, Kamil Pokora, Roberto Barra-Chicote and Daniel Korzekwa

Leveraging Explicit Acoustic Features for Controllable TTS

Tian Huey Teh, Devang S Ram Mohan, Vivian Hu, Alexandra Torresquintero, Zack Hodari, Tomás Gómez Ibarrondo, Christopher G. R. Wallis and Simon King

Treating the noisy phase issue in speech enhancement using complex ratio masks

Georgiana-Elena Sfecelis

Comparing human emotion perception and automatic emotion recognition of user turns in human-machine dialogues

Norbert Braunschweiler, Rama Doddipatla, Simon Keizer and Svetlana Stoyanchev

Language Modelling with Recurrent Neural Networks for Code-Switching

Olga Iakovenko and Thomas Hain

Speaker Diarization: Importance of the Modulation Spectrum and Incorporating Uncertainty Modelling

Simon McKnight

Modelling trajectories of human speech articulators using general Tau theory

Benjamin Elie, David Lee and Alice Turk

Multi-sentence TTS with Expressive and Coherent Prosody

Marcel Granero-Moya, Amith Nagaraj, Peter Makarov, Ammar Abbas, Mateusz Lajszczak, Arnaud Joly, Sri Karlapati, Alexis Moinet, Thomas Drugman and Penny Karanasou

Investigating perception of spoken dialogue acceptability through surprisal

Sarenne Wallbridge, Peter Bell and Catherine Lai

Peter 2.0: Building a Cyborg

Matthew Aylett, Ari Shapiro, Sai Prasad, Lama Nachman, Stacy Marsella and Peter Scott-Morgan

Monitoring sleep disordered breathing of long-Covid patients at home using acoustic AI technology

Gerardo Roa Dabike, Ning Ma and Guy Brown

Incremental Disfluency Detection for Spoken Learner English

Lucy Skidmore and Roger K. Moore

Audio-Based Computational Analysis of Podcast Expressivity

Shahar Elisha, Emmanouil Benetos, Jussi Karlgren and Mariano Beguerisse-Diaz

Tandem Multitask Training of Speaker Diarisation and Speech Recognition for Meeting Transcription

Xianrui Zheng, Chao Zhang and Phil Woodland

Comparing Human and Machine Perceptions of Voice Anonymisation

Farida Yusuf, Dan Kumpik, Matt Clifford, Jonathan Erskine and Jennifer Williams

ABAIR-ÉIST: recent progress in Irish language low-resource ASR development

Liam Lonergan, Christian Saam, Mengjie Qian, Neasa Ní Chiardáin, Christer Gobl and Ailbhe Ní Chasaide

A summary of the GENEA Challenge 2022 on co-speech gesture generation

Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, Carla Viegas, Teodor Nikolov, Mihail Tsakov and Gustav Eje Henter

Neural formant synthesis a proving ground for speech-synthesis control

Gustavo Teodoro Döhler Beck, Ulme Wennberg, Zofia Malisz and Gustav Eje Henter

Empowering neural TTS with HMMs to get the best of both worlds

Shivam Mehta, Harm Lameris, Éva Székely, Jonas Beskow and Gustav Eje Henter

Unsupervised data selection for Speech Recognition with contrastive loss ratios

Chanh Park, Rehan Ahmad and Thomas Hain

Domain-Informed Probing of wav2vec 2.0 Embeddings for Phonetic Features

Patrick Cormac English, Julie Carson-Berndsen and John Kelleher

Self-supervised Graphs for Audio Representation Learning with Limited Labeled Data

Amir Shirian, Krishna Somandepalli and Tanaya Guha

15h30-16h00 - Coffee Break

16h00-17h00 - Keynote 2 (chair: Simon King)

Speech Privacy: Where Are We Going and How to Get There?

Dr. Jennifer Williams

18h30 - Conference Dinner - Scottish National Gallery

Tuesday, 6th September 2022

09h30-10h30 - Oral Session A (chair: Jon Barker)

Evaluating watchability for video localisation

Zack Hodari, Tian Huey Teh, Vivian Hu, Tomás Gómez Ibarrondo, Devang S Ram Mohan, Alexandra Torresquintero, Chris Wallis, James Leoni and Simon King

Transforming adult to child speech for dubbing

Protima Nomo Sudro, Anton Ragni and Thomas Hain

Deciphering Speech: a Zero-Resource Approach to Cross-Lingual Transfer in ASR

Ondrej Klejch, Electra Wallington and Peter Bell

10h30-11h30 - Keynote 3 (chair: Kate Knill)

Multimodal Speech Embracing the Iceberg!

Prof. Naomi Harte

11h30-12h00 - Coffee Break

12h00-13h00 - Poster Session B

Person-specific automatic speaker recognition: understanding the behaviour of individuals for applications of ASR

Vincent Hughes, Paul Foulkes, Philip Harrison, Jessica Wormald, Chenzi Xu, David van der Vloed and Finnian Kelly

Speaker identification in courtroom contexts: performance of human listeners compared to a state-of-the-art forensic voice comparison system

Philip Weber, Nabanita Basu, Agnes S. Bali, Claudia Rosas-Aguilar, Gary Edmond, Kristy A. Martire and Geoffrey Stewart Morrison

Automatic generation of accented speech using phonetic features

Margot Masson, Anthony Ventresque and Julie Carson-Berndsen

Exploring hidden speech representations of self-supervised automatic speech recognition models

Tamara Soloveva, Ramon Sanabria and Peter Bell

AVSE Challenge: Audio-visual Speech Enhancement Challenge

Lorena Aldana, Cassia Valentini-Botinhao, Ondrej Klejch, Mandar Gogate, Kia Dashtipour, Amir Hussain and Peter Bell

Leveraging linguistic knowledge for accent robustness of end-to-end models

Andrea Carmantini and Peter Bell

A Biological Understanding of Dramatic Speech through Synthesis

Emily Lau, Brechtje Post and Kate Knill

Modelling Pronunciation Variation in Different Spoken Englishes

Emma O'Neill and Julie Berndsen

Using Utterance-Specific Dirichlet Priors to Model Uncertainty in Emotion Class Labels

Wen Wu, Chao Zhang, Xixin Wu and Philip C. Woodland

PSE-Net: Real-time Personalized Sound Enhancement

Abhinav Mehrotra, Alberto Gil C. P. Ramos, Nic Lane and Sourav Bhattacharya

Conversational Speech vs. Sustained Phonation for Diagnosis of Parkinsons Disease

Steve Beet, Phill Restall and Ladan Baghai-Ravary

Tree-Constrained Pointer Generator for End-to-end Contextual ASR

Guangzhi Sun, Chao Zhang and Phil Woodland

Canonical-Correlated Graph Neural Network for Multimodal Energy-Efficient Speech Enhancement

Leandro Aparecido Passos Junior, Ahmed Khubaib, Mohsin Raza, Amir Hussain and Ahsan Adeel

CognoSpeak: a Cognitive Health Assessment Tool (CcHAT)

Nathan Pevy, Heidi Christensen and Daniel Blackburn

Attention Forcing for Speech Synthesis

Qingyun Dou and Mark Gales

Multimodal Emotion Recognition in Conversations

Jiachen Luo, Joshua Reiss and Huy Phan

Addressing user concerns about multi-modal hearing technology

Dorothy Hardy, Michael Akeroyd, Adeel Hussain, Peter Bell and Amir Hussain

Model for Assessor Bias in Automatic Pronunciation Assessment

Jose Antonio Lopez Saenz and Thomas Hain

A siamese RNN architecture to detect deliberate imitation and phonetic convergence in L2-speech

Byron Z. Yuan, Aldo Pastore, Dorina De Jong, Hao Xu, Luciano Fadiga and Alessandro D'Ausilio

Using conversational data to improve prosody in Text-to-Speech synthesis

Johannah O'Mahony, Catherine Lai and Simon King

Non-Linear Pairwise Language Mappings for Low-Resource Multilingual Acoustic Model Fusion

Muhammad Umar Farooq, Darshan Adiga Haniya Narayana and Thomas Hain

RoomReader: A Multimodal Corpus of Online Multiparty Conversational Interactions

Justine Reverdy, Sam O'Connor Russell, Louise Duquenne, Diego Garaialde, Benjamin Cowan and Naomi Harte

Alternative Evaluation Methods of Latent Representations of Speech Audio

Eimear Stanley, Yumnah Mohamied, Peter Bell

13h00-14h00 - Lunch

14h00-15h00 - Poster Session C

Autovocoder: Vocoding Without Spectrograms

Jacob Webber and Simon King

Exploring Prosody Transfer in Speech Synthesis

Atli Sigurgeirsson and Simon King

Code-switched Text Generation on Parallel Data

Jie Chi and Brian Lu

Voice Puppetry for the People: Harnessing Dramatic Performance for Speech Synthesis

Matthew Aylett, Skaiste Butkute and Christopher Pidcock

Improving diagnostic procedures for epilepsy through automated recording and analysis of patients history

Nathan Pevy, Heidi Christensen, Traci Walker and Markus Reuber

Deliberation Based Multi-Pass Speech Synthesis

Qingyun Dou and Mark Gales

Exploring Novel Methods for Automatic Speech Recogniser Based Intelligibility Prediction

Zehai Tu, Ning Ma and Jon Barker

View-Specific Assessment of L2 Spoken English

Stefano Banno, Bhanu Balasu, Mark Gales, Kate Knill and Konstantinos Kyriakopolous

Why is My Social Robot so Slow? How a Conversational Listener can Revolutionize Turn-Taking

Matthew Aylett, Andrea Carmantini and David Braude

Creating New Voices using Normalizing Flows

Piotr Biliski, Thomas Merritt, Abdelhamid Ezzer, Kamil Pokora, Sebastian Cygert, Kayoko Yanagisawa, Roberto Barra-Chicote and Daniel Korzekwa

Phonetic Analysis of Self-supervised Representations of English Speech

Dan Wells, Hao Tang and Korin Richmond

Comparison of Audio-Visual Speech Enhancement Models with Hearing Aid Key Performance Indicators

Jasper Kirton-Wingate, Mandar Gogate, Amir Hussain and Tassadaq Hussain

Simulation of Teacher-Learner Interaction in English Language Pronunciation Learning

Elaf Islam and Thomas Hain

A New Benchmark Multi-modal Speech Corpus With Two Target Speakers

Jasper Kirton-Wingate, Adeel Hussain, Amir Hussain, Kia Dashtipour, Mandar Gogate and Peter Derleth

Gender Bias and Universal Substitution Adversarial Attacks on Grammatical Error Correction Systems for Auto-mated Assessment

Vyas Raina and Mark Gales

Is there an auditory uncanny valley for synthesised speech?

Alice Ross, Catherine Lai and Martin Corley

Exploration of A Self-Supervised Speech Model: A Study on Emotional Corpora

Yuanhao Li, Yumnah Mohamied, Peter Bell and Catherine Lai

Cross Éingual Éav2vec Énetuning Én Étually Éntelligible Éanguage Éairs

Jeffrey Josanne Michael, Toby Godwin and Oscar Saz

Phonetically Guided Transfer Learning for Low-Resource Accented English

Edward Storey and Naomi Harte

Dysarthric Speech Recognition From Raw Waveform with Parametric CNNs

Zhengjun Yue, Erfan Loweimi, Heidi Christensen, Jon Barker and Zoran Cvetkovic

Joint Modelling of Automatic Speaker Verification and Spoofing Countermeasure Systems

Poppy Welch and Jennifer Williams

15h00-16h00 - Oral Session B (chair: Julie Berndsen)

The 2nd Clarity Enhancement Challenge: A machine learning challenge for hearing aid speech intelligibility enhancement

Will Bailey, Michael Akeroyd, Jon Barker, Trevor Cox, John Culling, Simone Graetzer, Graham Naylor, Zuzanna Podwiska and Zehai Tu

Back to the Future: Extending the Blizzard Challenge 2013

Sébastien Le Maguer, Simon King and Naomi Harte

Fine Grained Spoken Document Summarization Through Text Segmentation

Samantha Kotey, Rozenn Dahyot and Naomi Harte

16h00-16h15 - Closing Ceremony

Keynotes

Keynote 1 - Dr. Joanne Cleland

Using Ultrasound to Image the Articulators in the Speech Therapy Clinic



Abstract

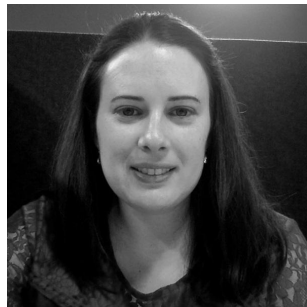
Speech sound disorders are common in childhood and can affect the education and wellbeing of children. This talk will first provide an overview of using medical ultrasound to image the articulators for assessment and treatment of speech sound disorders in children. Using this technique, we are able to see the tongue moving in real-time and use this for both the assessment of speech sound disorders and as a biofeedback tool for intervention. In this talk I will make a case for how imaging the articulators leads to more precise diagnosis of speech sound disorders and provides insight into the underlying nature of such disorders. I will also explore how in the future we might automate the classification of speech sound disorders using dynamic ultrasound, leading not only to quicker but also more precise diagnosis.

Biography

Joanne Cleland is a Reader in Speech and Language Therapy at the University of Strathclyde in Glasgow. Her research focuses on using instrumental techniques, particularly ultrasound tongue imaging, for the assessment of speech sound disorders in children. You can find out more about her work by following her on twitter: [@DrJoanneCleland](https://twitter.com/DrJoanneCleland)

Keynote 2 - Dr. Jennifer Williams

Speech Privacy: Where Are We Going and How to Get There?



Abstract

Audio recording devices and speech technologies are becoming increasingly commonplace in everyday life. At the same time, commercialised speech and audio technologies do not provide consumers with a range of privacy choices. Even where privacy is regulated or protected by law, technical solutions to privacy assurance and enforcement fall short. Within the speech research community there are no standard technical definitions of privacy and security. However, privacy is usually taken to refer to "controlling access to information" whereas security is often taken to refer to "how information can be used (or misused)". This talk highlights several critical challenges to developing trustworthy speech and audio technologies in the context of privacy and security. Particularly, a new type of speech privacy will be introduced as an emerging research area: content-based privacy. True progress toward trustworthy speech and audio technology will require an interdisciplinary approach that combines perspectives from multiple science, legal, artistic, and social domains. Interdisciplinary approaches are especially important because sometimes issues of privacy and security are well-known among technical researchers (who create the technologies). Issues may become known among other scholars only once technology has been commercialised. In fact, with speech technology moving "at the speed of light" lately, timing is critical. Such a gap must be closed for progress on speech technology that moves us toward a freer and more just world.

Biography

Dr Jennifer Williams is a postdoctoral Research Fellow at the University of Southampton. She currently works in two main areas: citizen-centric AI systems and trustworthy autonomous systems. One aspect of her research explores the creation of trustworthy, private, and secure speech/audio solutions for smart buildings that can contribute to accessibility as well as resource management and "low-carbon comfort". Dr. Williams is also the PI of a small interdisciplinary project between University of Southampton and University of Nottingham which explores regulation and policy development in the context of speech applications for the creative industries. She completed her PhD at the University of Edinburgh (2022) in the area of representation learning and speech signal disentanglement. She applied that work to a variety of speech technology applications (voice conversion, speech synthesis, anti-spoofing, naturalness assessment, and privacy). She also holds a position in industry as a Senior Speech Scientist at MyVoice AI where she develops ultra-low power speech technology to run on edge devices. Dr Williams was previously a staff member at MIT Lincoln Laboratory for five years developing prototype speech and text technology for the US Government. She is a member of IEEE and ISCA, serves as a committee member of the ISCA-PECRAC group, and co-organizes ISCA SPSC-SIG events. She is a reviewer for multiple conferences involving AI, text, speech, and multimedia. She holds an MScR in Data Science from University of Edinburgh (2018), an MS in Computational Linguistics from Georgetown University (2012) and a BA in Applied Linguistics, magna cum laude, from Portland State University (2009).

Keynote 3 - Prof. Naomi Harte

Multimodal Speech Embracing the Iceberg!



Abstract

This talk will consider the multimodal nature of speech and speech technology. Human speech communication is extremely rich. We use many elements to communicate, from words to gestures and eye gaze, and seamlessly interpret these many cues in our conversations. How can we exploit this in technology? In my talk, I'll look at how visual and linguistic information can be integrated into deep learning frameworks for audio-visual speech recognition and turn taking prediction. I'll also explore how conversational interaction online can be challenging due to disruptions to the cues we usually rely on, and consider whether multimodal approaches can help.

Biography

Naomi is Professor in Speech Technology in the School of Engineering in Trinity College Dublin. She is Co-PI and a founding member of the ADAPT SFI Centre in Ireland. In ADAPT, she has led a major Research Theme centered on Multimodal Interaction involving researchers from Universities across Ireland and was instrumental in developing the future vision for the Centre for 2021-2026. She is also a lead academic in the Sigmedia Research Group in the School of Engineering. Prior to starting her lectureship in TCD in 2008, Naomi worked in high-tech start-ups in the field of DSP Systems Development, including her own company. She also previously worked in McMaster University in Canada. She was a Visiting Professor at ICSI in 2015, and became a Fellow of TCD in 2017. She earned a Google Faculty Award in 2018 and was shortlisted for the AI Ireland Awards in 2019. She currently serves on the Editorial Board of Computer Speech and Language, and will Chair Interspeech 2023 in Dublin.

Poster Session A

Text-free non-parallel many-to-many voice conversion using normalising flows

*Thomas Merritt¹, Abdelhamid Ezzer¹, Piotr Biliński¹, Magdalena Proszewska^{2*},
Kamil Pokora¹, Roberto Barra-Chicote¹, Daniel Korzekwa¹*

¹ Amazon Alexa, ² Jagiellonian University, Poland

{thommer, ezzer, bilipiot, kamipoko, rchicote, korzekwa}@amazon.com

Abstract

Non-parallel voice conversion (VC) is typically achieved using lossy representations of the source speech. However, ensuring only speaker identity information is dropped whilst all other information from the source speech is retained is a large challenge. This is particularly challenging in the scenario where at inference-time we have no knowledge of the text being read, i.e., text-free VC. To mitigate this, we investigate information-preserving VC approaches.

Normalising flows have gained attention for text-to-speech synthesis, however have been under-explored for VC. Flows utilize invertible functions to learn the likelihood of the data, thus provide a lossless encoding of speech. We investigate normalising flows for VC in both text-conditioned and text-free scenarios. Furthermore, for text-free VC we compare pre-trained and jointly-learned priors. Flow-based VC evaluations show no degradation between text-free and text-conditioned VC, resulting in improvements over the state-of-the-art. Also, joint-training of the prior is found to negatively impact text-free VC quality.

Index Terms: voice conversion, Flow-TTS, Glow-TTS, Copy-Cat, AutoVC

Work performed during an internship at Amazon.

Leveraging Explicit Acoustic Features for Controllable TTS

T. H. Teh¹, D. S. R. Mohan¹, V. Hu¹, A. Torresquintero¹, Z. Hodari¹, T. G. Ibarrondo¹, C. G. R. Wallis¹ and S. King^{1, 2}

¹Papercup Technologies Ltd., ²University of Edinburgh

Email: tian@papercup.com

Text does not fully specify the spoken form, so text-to-speech models must be able to learn from speech data that vary in ways not explained by the corresponding text. One way to reduce unexplained variation in the training data is to provide acoustic information as an additional learning signal. Since much of the unexplained variation is in the prosody, we adopt a model that generates speech explicitly conditioned on the three primary acoustic correlates of prosody: F0, energy and duration.

During inference, these explicit acoustic features can either be predicted from text, predicted then subsequently modified, or externally provided. The latter two inference modes enable multiple distinct renditions of a text to be produced, and offer interpretable, temporally-precise, and disentangled control over the prosody of the synthesised speech. These characteristics open up myriad ways for a human-in-the-loop to manipulate the prosody of the synthesised speech.

We demonstrate two modes of control, that trade off between the efficiency and specificity of control by a human-in-the-loop. Both modes of control interact directly with the sequence of acoustic features via different user interface modalities. Subjective evaluations show that both modes of control are able to improve listener preference when applied appropriately.

References

D. S. R. Mohan, V. Hu, T. H. Teh, A. Torresquintero, C. G. R. Wallis, M. Staib, L. Foglianti, J. Gao, and S. King, “Ctrl-P: Temporal control of prosodic variation for speech synthesis,” INTERSPEECH, 2021.

Treating the noisy phase issue in speech enhancement using complex ratio masks

Georgiana-Elena Sfecelis

University of East Anglia
Norwich, UK

`g.sfecelis@uea.ac.uk`

Index Terms: speech enhancement, deep learning, complex ratio masks

Abstract

The presence of noise and distortion is the main cause of information loss and decreased quality in speech-based communication channels. One of the most challenging tasks within the speech processing study-area is speech enhancement, whose aim is to extract an estimate of clean speech from a noisy utterance. The state-of-the-art is to tackle this problem using deep learning algorithms.

The time-domain representation of a signal provides little usable information with regards to the contents of an audio signal. A common analysis process involves converting the speech utterance to a time-frequency representation using the short-time Fourier transform (STFT), which outputs a complex spectrogram of the input.

The direct prediction of an enhanced signal from a noisy input is often difficult, and thus several masking approaches have been proposed to simplify this task. Most masking-based algorithms only make use of the magnitude of the time-frequency signal representation and reconstruct the speech using the noisy phase, as the phase component is usually neglected across literature due to the human inability to perceive it. Such examples are the ideal binary mask (IBM) and the ideal ratio mask (IRM), or the phase-sensitive mask (PSM).

Nonetheless, none of these magnitude masks provide a perfect reconstruction of the clean speech due to the noisy phase issue. Although methods for enhancing the phase exist (e.g. PSM), none of them can derive an ideal prediction of the oracle phase due to the unstructured nature of this component. The concept of complex ideal ratio mask (CIRM) has been introduced to tackle this issue, where Cartesian coordinates are used to represent the complex spectrograms, from which a real and imaginary mask is estimated for each speech signal.

This study explores the use of CIRMs as targets, rather than magnitude and phase based masks, and compares their implicit way of enhancing the phase with a magnitude-based mask such as IRM. We use a Y-shaped CNN model to extract features from the complex spectrograms and output two concatenated sets for the complex components, namely the real and imaginary masks. The results are compared with a Y-shaped DNN model for complex mask estimation. These results are further measured against the magnitude-based IRM mask, which uses the noisy phase upon signal reconstruction. Our experiments are performed on utterances contaminated with babble noise with SNRs between -5 and 5 dB, as the distortion caused by phase is higher at low SNRs. The enhanced signals are evaluated in terms of quality (PESQ) and enhanced intelligibility (ESTOI) but also assessed using an objective framework to measure speech quality (SIG), background noise quality (BAK), and overall audio quality (OVRL). Preliminary results show intelligibility improvements when speech is enhanced using CIRM as compared to IRM.

Submitted to:
UK Speech 2022

Comparing human emotion perception and automatic emotion recognition of user turns in human-machine dialogues

Norbert Braunschweiler, Rama Doddipatla, Simon Keizer, Svetlana Stoyanchev

Toshiba Europe Limited, Cambridge Research Laboratory, Cambridge, CB4 0GZ, UK

{firstname.lastname}@crl.toshiba.co.uk

Abstract

The success of automatic emotion recognition depends on its ability to resemble human-level performance in human-machine interaction such as in emotionally-aware spoken dialogue systems (SDS). Incorrect recognition could trigger an inadequate response and invalidate its original purpose. In light of this, the current paper presents a comparison of human perception vs. automatic recognition of spoken user turns from human-machine dialogues. In a series of listening tests subjects are asked to rate which emotion is expressed in a user turn presented as 1) speech-only, 2) speech with text of prior dialogue, and 3) speech with both text & speech of prior dialogue. Results show that human ratings are affected by prior dialogue context most noticeably when both speech and text is available. However, human inter-rater agreements show only fair to moderate agreement. Agreement between human ratings and two model estimates (speech-only vs. speech&text) is also moderate, varies by emotion-class and differs noticeably between models. However, adding speech-only model estimates to human ratings only marginally lowers agreement rates indicating that model estimates are comparable to human agreement on this task.

Listening tests

The 3 listening tests presented the same 40 user turns selected from the Toshiba Laptop corpus with no previous context (LT1), text-only (LT2) or text & speech (LT3) context. Subjects were asked to rate which emotion is expressed in a user turn by selecting from the 4 emotion classes: *angry*, *happy*, *sad*, *neutral* or to select *other*. Figure 1 shows an example of the interface seen by subjects during listening test LT3. In read-only test RT subjects were asked to read a transcript of a dialogue snippet and rate which emotion they expected to be expressed by the user at the end.

SYS: Hello, welcome to the laptop info system, what laptop are you after?
 USR: I WANT AN EXPENSIVE LAPTOP IN THE MID WEIGHT RANGE
 SYS: There are 27 laptops in the expensive price range. What product family would you like: satellite - satellite pro - tecra - or portege?
 USR: I DON'T CARE
 SYS: Would you like something in the expensive or moderate price range?
 USR: EXPENSIVE PRICE RANGE

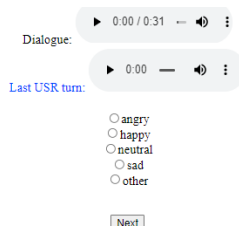


Figure 1: Interface seen by subjects in listening test LT3.

Automatic emotion estimation

All user turns in the Toshiba laptop corpus were automatically labeled with one of the emotions *angry*, *happy*, *neutral*, *sad* by state-of-the-art speech emotion recognition models using either speech-only (Sp) or speech&text (SpTxt).

Results

Table 1 provides an overview of human ratings and model estimates, showing the percentage of assignments for each of the 4 emotion classes *angry* (=ang), *happy* (=hap), *sad* (=sad), *neutral* (=neu) or the *other* (=oth) class and by models speech-only (Sp) and speech&text (SpTxt). In addition, the lower half of Table 1 shows inter-rater agreement values, calculated with Krippendorff's alpha (α)-metric to show agreement amongst human raters (α humans) and human plus model predictions for speech-only model (α +Sp) and for speech&text-model (α +SpTxt).

Table 1: Percentage of assignments for one of 4 emotion classes or the 'other' class by human raters and models.

	Human ratings [%]				Models [%]	
	LT1	LT2	LT3	RT	Sp	Sp-Txt
ang	19	28	18	32	25	7.5
hap	3	3	2	0	25	7.5
neu	58	52	63	62	25	75
sad	14	11	15	6	25	10
oth	5	6	2	0	0	0
α humans	0.378	0.388	0.236	0.359		
α +Sp	0.372	0.362	0.265	0.208		
α +SpTxt	0.330	0.314	0.171	0.221		

Numbers are rounded to the nearest integer.

→ Agreement of emotion estimation:

- for human subjects is only fair to moderate
- between human ratings and two model estimates (speech-only vs. speech&text) is also only fair to moderate and varies by emotion-class
- however, adding estimates from the speech-only model to human ratings marginally lowered overall agreement ratings indicating that model estimates are comparable to human agreement on this task

→ Impact of context:

- human emotion ratings showed lower agreement when both text and speech were presented

Language Modelling with Recurrent Neural Networks for Code-Switching

Olga Iakovenko¹, Thomas Hain¹

¹The University of Sheffield, Sheffield, UK

oiakovenko1@sheffield.ac.uk, t.hain@sheffield.ac.uk

Abstract

Code-switching (CS) is a process of speaker changing between languages in the context of a language production act. This phenomenon can appear in communication between speakers when they are fluent in two or more languages. In Automatic Speech Recognition (ASR) this process can be modelled by using monolingual data in a single ASR system. Although being feasible in theory, in practice models trained on highly resourced monolingual data perform worse than models trained on smaller amount of CS data. This happens because the current ASR models are not trained to capture the patterns of how exactly speech units are being interchanged within an utterance in respect to their languages. To solve the issue is to model code-switching production in spoken languages by using monolingual data and incorporating linguistically informed Matrix Language Frame and 4-M models. In this work analysis and modelling of sequences of morphemes was carried out for a SEAME dataset of Singaporean colloquial language which contains English/Mandarin code-switching. Recurrent Neural Network Language Models (RNNLM) were trained on monolingual transcriptions and applied to CS transcriptions to test the generalised 1st Matrix Language Frame model principle. Results show that clauses tend to follow Matrix Language (ML) word order if the clause contains singleton lexemes from the Embedded Language (EL) while also applicable to cases with Embedded Language Islands.

Index terms should be included as shown below.

Index Terms: language model, recurrent neural network, code-switching, morpheme segmentation

This is an ongoing work.

Speaker Diarization: Importance of the Modulation Spectrum and Incorporating Uncertainty Modelling

Simon W. McKnight*

Department of Electrical and Electronic Engineering, Imperial College London, UK

s.mcknight18@imperial.ac.uk

Abstract

This Thesis covers research into speaker diarization, an important area of research that has several practical applications. It focuses on three main areas: (a) analyses of the evaluation methodology based on diarization error rates (DER) and use of forgiveness collars; (b) the use of modulation spectrum features to distinguish speakers, including overlapping speakers, and investigation into salient parts; and (c) uncertainty quantification of machine learning models and using those uncertainties to improve performance.

The analyses of evaluation methodology highlight shortcomings in the use of DERs and sensitivity to the ground truth (GT) used. Initial research showed that using simple DERs without forgiveness collars can unfairly penalise diarization systems and proposed using phoneme-dependent collars. However, subsequent human subject-based experiments on a 5-minute AMI meeting extract were conducted and compared to state-of-the-art systems, showing that forgiveness collars generally are not a satisfactory way of dealing with insignificant GT deviations. Figure 1 shows human reviewers score well if the predicted number of segments matches that used by the GT, which is not accounted for by forgiveness collars. Furthermore, using forgiveness collars improves results but increases the standard deviation of those results. Consequently, the use of forgiveness collars is not recommended and is not used in subsequent results.

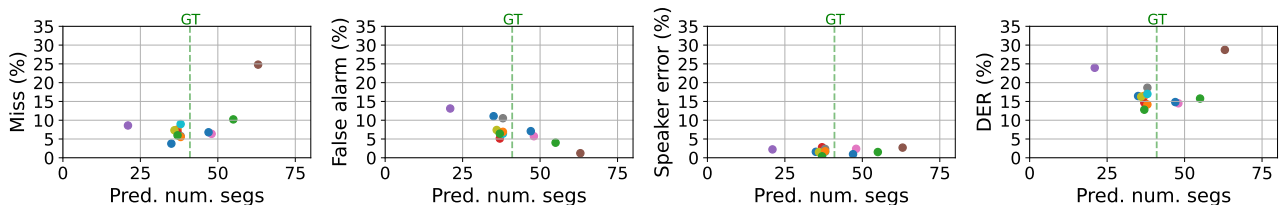


Figure 1: Human reviews errors against predicted number of speech segments.

Modulation spectrum features (Φ) are thought to be a promising way to generate features that distinguish speakers well, particularly the joint acoustic and modulation spectrum form. The modulation spectrum features investigated include both temporal envelope and temporal fine structure. It highlights that modulation frequencies in the 0-0.5 Hz range and around the fundamental frequencies of the speakers are most useful, contradicting earlier research preferring the 1-16 Hz range, and shows that although the temporal envelope contains more speaker information than the temporal fine structure, the temporal fine structure contains useful additional speaker information and should not be excluded.

Training machine learning models on both Φ and mel-frequency cepstral coefficients (MFCCs) on the AMI Corpus are shown to consistently give better results than either alone. These features have different durations, so a model using convolutional neural networks (CNNs) on the modulation spectrum features and recurrent neural networks - long short-term memory (LSTMs) on the MFCCs, and those outputs concatenated before feeding into dense layers. One example had 27.78% for LSTM on MFCCs, 29.09% for CNN on Φ and 19.44% for the combined model. However, these models do not distinguish overlapping speakers as well as anticipated.

Machine learning models that indicate their prediction confidence are clearly more helpful than those that simply predict. This Thesis investigates models quantifying aleatoric and epistemic uncertainties, using output probability distributions and Monte Carlo dropout respectively and illustrated in Figure 2. Kalman filters to smooth individual model predictions and/or combine with other model predictions (including deep neural networks (DNNs) based on x-vectors) are tested. The probabilities enable meaningful conclusions to be drawn from entropies, particularly where frame-based error metrics are used rather than typical time-based error metrics. DER performance is improved for certain hyperparameters, both for single models and model ensembles, and the best ensemble DER of 9.29% on the relevant AMI meetings is comparable to benchmark state-of-the-art models (one was 6.23%, another was 11.25%).

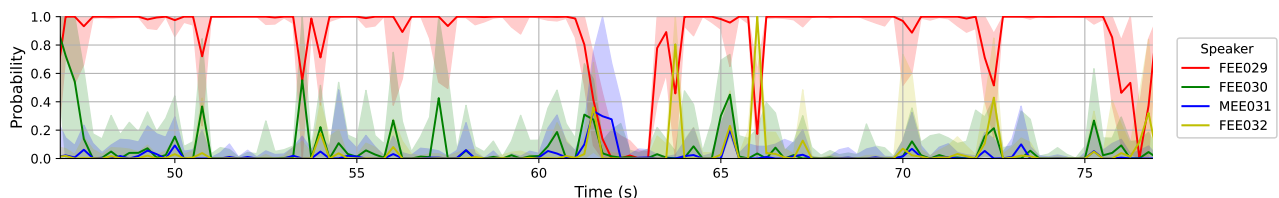


Figure 2: 30 s extract showing aleatoric and epistemic probabilities for joint speaker diarization and identification.

* Now a research assistant in the ALTA team in Cambridge University Engineering Department: swm35@cam.ac.uk.

Modelling trajectories of human speech articulators using general Tau theory

Benjamin Elie¹, David N. Lee², Alice Turk¹

¹Linguistics and English Language; School of Philosophy, Psychology and Language Sciences; the University of Edinburgh; Edinburgh, Scotland, United Kingdom

²Psychology; School of Philosophy, Psychology and Language Sciences; the University of Edinburgh; Edinburgh, Scotland, United Kingdom

benjamin.elie@ed.ac.uk, d.n.lee@ed.ac.uk, a.turk@ed.ac.uk

Abstract

We present an application of general Tau theory for modelling and analysing the trajectories of speech articulators. Tau theory has been applied successfully in the past to several kind of bodily movements [1, 2, 3, 4, 5, 6]. The basic assumption of Tau theory is that purposeful movements aim at closing gaps, e.g. a distance gap, an angle gap, ... The gap-closure function is defined such that the target is reached at the right time, *i.e.* the gap is closed at the movement endpoint. Another interest for trajectory modelling is that Tau-guided movements exhibit single-peaked velocity profiles whose symmetry can be changed by adjusting a unique parameter, the Tau-coupling parameter k . Considering these features, general Tau theory is thus a good candidate to explain the production of articulatory trajectories in speech.

We evaluated its relevance for speech, both for the generation and analysis of articulatory trajectories, by comparing the fit of the Tau theory equation to real articulatory trajectories, extracted from electromagnetic articulometry (EMA) data, with the fit of existing methods [7, 8, 9]. The comparison is carried over the trajectories of all sensor signals from a dataset from the DoubleTalk corpus, corresponding to EMA recordings of 12 native speakers of English reading an English text [10, 11]. Our experiments show that the general Tau theory model gives a better fit than existing methods (critically damped oscillators and sequential target approximation). These findings support the hypothesis of Tau-guided movements of articulators during speech production.

This paper also presents a statistical investigation of Tau equation parameters that provide the best fit to observed movement trajectories in the EMA database. This analysis shows that articulatory movements follow similar velocity profile distributions across speakers. The shape parameter k of Tau-guided movements is identically distributed across speakers, following a unimodal distribution. The statistical mode is very close to the value of k for which Tau-guided movements minimize a cost function based on forces acting on the moving articulator, providing new evidence that articulatory effort is optimized during speech production.

1. References

- [1] D. N. Lee, "Guiding movement by coupling taus," *Ecological psychology*, vol. 10, no. 3-4, pp. 221–250, 1998.
- [2] D. N. Lee and P. E. Reddish, "Plummeting gannets: A paradigm of ecological optics," *Nature*, vol. 293, no. 5830, pp. 293–294, 1981.
- [3] D. Lee, D. Young, P. Reddish, S. Lough, and T. Clayton, "Visual timing in hitting an accelerating ball," *The Quarterly Journal of Experimental Psychology*, vol. 35, no. 2, pp. 333–346, 1983.
- [4] C. M. Craig and D. N. Lee, "Neonatal control of nutritive sucking pressure: evidence for an intrinsic τ -guide," *Experimental Brain Research*, vol. 124, no. 3, pp. 371–382, 1999.
- [5] B. Schogler, G.-J. Pepping, and D. N. Lee, "TauG-guidance of transients in expressive musical performance," *Experimental Brain Research*, vol. 189, no. 3, pp. 361–372, 2008.
- [6] M. W. Rodger, S. O'Modhrain, and C. M. Craig, "Temporal guidance of musicians' performance movement is an acquired skill," *Experimental brain research*, vol. 226, no. 2, pp. 221–230, 2013.
- [7] E. Saltzman, "Task dynamic coordination of the speech articulators: A preliminary model," *Experimental brain research series*, vol. 15, pp. 129–144, 1986.
- [8] B. J. Kröger, G. Schröder, and C. Opgen-Rhein, "A gesture-based dynamic model describing articulatory movement data," *The Journal of the Acoustical Society of America*, vol. 98, no. 4, pp. 1878–1889, 1995.
- [9] P. Birkholz, B. J. Kroger, and C. Neuschaefer-Rube, "Model-based reproduction of articulatory trajectories for consonant-vowel sequences," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1422–1433, 2010.
- [10] J. M. Scobbie, A. Turk, C. Geng, S. King, R. Lickley, and K. Richmond, "The Edinburgh speech production facility DoubleTalk corpus," in *INTERSPEECH 2013: Proceedings of the 14th Annual Conference of the International Speech Communication Association (ISCA)*, 25-29 August 2013, Lyon, France, International Speech Communication Association, 2013.
- [11] C. Geng, A. Turk, J. M. Scobbie, C. Macmartin, P. Hoole, K. Richmond, A. Wrench, M. Pouplier, E. G. Bard, Z. Campbell, *et al.*, "Recording speech articulation in dialogue: Evaluating a synchronized double electromagnetic articulography setup," *Journal of Phonetics*, vol. 41, no. 6, pp. 421–431, 2013.

Multi-sentence TTS with Expressive and Coherent Prosody

Marcel Granero-Moya, Amith Nagaraj, Peter Makarov, Ammar Abbas, Mateusz Lajszczak, Arnaud Joly, Sri Karlapati, Alexis Moinet, Thomas Drugman, Penny Karanasou

Alexa AI, Amazon

moymarce@amazon.com

Introduction - Generating expressive and contextually appropriate prosody remains a challenge for modern text-to-speech (TTS) systems. This is particularly evident for TTS for long, multi-sentence inputs. We present BLT (BERT Long-Transformer), a multi-sentence TTS model based on four premises. First, transformer architectures based on FastSpeech [1] achieve high speech quality, efficiency, and are generally expected to perform best on longer inputs. Second, long context makes prosody less ambiguous, simplifying the task. Third, contextual word embeddings from large language models are important for expressive and contextually appropriate speech [2]. Lastly, multi-speaker modelling facilitates transfer learning and helps avoid overfitting due to larger dataset sizes.

Our contributions are twofold. (1) We examine three simple extensions to a state-of-the-art TTS system, and (2) we see gains from all three extensions.

Methods - We take a Transformer-based FastSpeech-like system [1] as our baseline. The system is comprised of an acoustic model and a duration model. Both models use Feed-Forward Transformer (FFT) blocks as feature encoders. The duration model predicts phoneme durations (in frames) and is a regressor on top of the phoneme FFT encoder. The acoustic model predicts mel-spectrogram frames and consists of the phoneme FFT encoder and frame-level feature FFT encoder. The output embeddings of the phoneme FFT encoder are up-sampled to the frame level using the phoneme durations predicted by the duration model and then fed into the frame-level feature FFT encoder. The acoustic and duration models use separate phoneme encoders. The architecture of the duration model and the fact that we use force-aligned phoneme durations as training targets are the two main differences from FastSpeech.

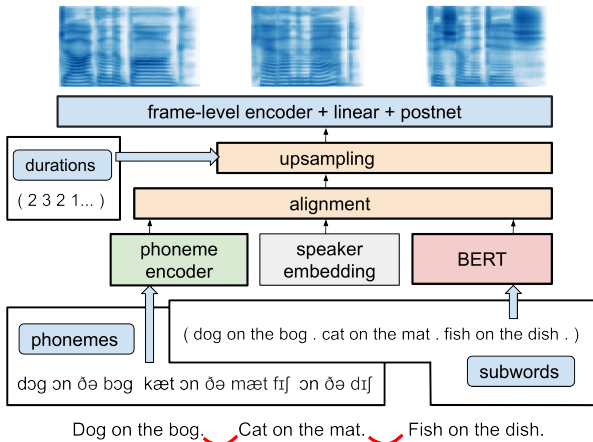


Figure 1: BLT diagram with multi-speaker modelling, BERT word embeddings and long context.

We examine three extensions to this baseline with the goal of improving expressiveness and prosodic coherence on multi-sentence inputs. First, we train the system on multiple speakers as it reduces the risk of overfitting the language model during fine-tuning simply due to the sheer amount of training data. To turn the baseline into a multi-speaker model, we concatenate pre-trained speaker embeddings to phoneme encodings as shown in Figure 1. Second, we use linguistic features since prosody closely tracks syntax. Concretely, we concatenate BERT [3] word embedding to the phoneme encodings. Third, we propose training and synthesizing on longer context. To this end, we concatenate sentences that come after each other in the original data.

Evaluations - We conduct experiments on internal datasets of three female English speakers. Data consist of multi-sentence chunks with total durations of about 24 h, 32 h, and 22 h. We conduct MUSHRA evaluations to compare our models combinations. We vocode mel-spectrograms extracted from the reference audio and use it as reference. We observe consistent statistically significant improvements in all ablations (Table 1).

Mean MUSHRA scores			
	left-most	BLT	reference
FastSpeech-like baseline	70.0 $\pm$ 0.8	73.6 $\pm$ 0.8	75.1 $\pm$ 0.8
+ multi-speaker	66.9 $\pm$ 0.8	72.9 $\pm$ 0.7	73.7 $\pm$ 0.7
+ multi-speaker, BERT	72.1 $\pm$ 0.8	73.4 $\pm$ 0.7	74.3 $\pm$ 0.7

Table 1: MUSHRA results for BLT compared against baseline with a combination of extensions. Left-most stands for left-most column system.

Conclusion - This work looks at multi-sentence TTS. To increase the coherence and contextual appropriateness of prosody we examine three extensions to a Transformer-based FastSpeech-like baseline: long context, contextual word embeddings, and multi-speaker modelling. We observe large improvements over the baseline with all extensions. Full article of this work has been accepted at Interspeech 2022 and published on arXiv [4].

References

- [1] Y. Ren *et al.*, "FastSpeech: Fast, robust and controllable text to speech," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [2] T. Kenter *et al.*, "Improving prosody of RNN-based English text-to-speech synthesis by incorporating a BERT model," *INTER-SPEECH*, 2020.
- [3] J. Devlin *et al.*, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [4] P. Makarov *et al.*, "Simple and Effective Multi-sentence TTS with Expressive and Coherent Prosody," *INTER-SPEECH*, 2022.

Investigating perception of spoken dialogue acceptability through surprisal

Sarenne Wallbridge, Peter Bell, Catherine Lai

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

{s1301730, peter.bell, c.lai}@ed.ac.uk

Recent developments in automatic language modelling have made it possible to test (and extend) psycholinguistic models of human language comprehension. In particular, autoregressive Language Model (LM) estimates of surprisal have been found to correlate with aspects of language perception, including reading times and grammatical acceptability judgements of text. Although spoken interaction is arguably the primary form of language use, most studies of surprisal are based on monological, written data.

Progress in developing models of perceptual salience for the speech signal has been markedly slower than for its textual counterpart. Towards the goal of understanding perception in spontaneous, natural language, we present an exploratory investigation into *whether the relationship between human comprehension behaviour and LM-estimated surprisal holds when applied to dialogue*, considering both written dialogue, and the lexical component of spoken dialogue. We use a novel judgement task of dialogue utterance acceptability where the amount of information in contextual anchors is controlled-for to ask two questions: “How well can people make predictions about written dialogue and transcripts of spoken dialogue?” and “Do LM estimates of surprisal correlate with these acceptability judgements?”.

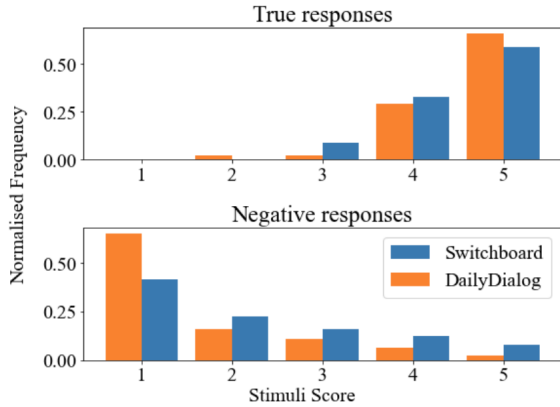


Figure 1: Human score distributions between written (Daily Dialog) and spoken (Switchboard) corpora: people score true responses and randomly sampled negative responses with more certainty for the written data compared to spoken transcripts.

We demonstrate that people can make accurate predictions about upcoming dialogue. Their ability differs between spoken transcripts and written conversation, indicating that lexical information is distributed differently between these two modalities. Using state-of-the-art dialogue LMs, we investigate the relationship between global and local operationalizations of surprisal and human acceptability judgements. We find weak but statistically significant correlations between different LM estimates of surprisal and acceptability judgements. We find a range of correlations for individual operationalizations and that a combination of both global and local provides the most predictive power, motivating future work into how to represent information distribution in the linguistic signal of communication.

$$S_{total}(\mathbf{r}|\mathbf{c}) = \sum_{n=1}^N [S(r_n|\mathbf{r}_{<n}, \mathbf{c})]$$

$$S_{mean}(\mathbf{r}|\mathbf{c}) = \frac{1}{N} \sum_{n=1}^N [S(r_n|\mathbf{r}_{<n}, \mathbf{c})]$$

$$S_{relative}(\mathbf{r}|\mathbf{c}) = S_{mean}(\mathbf{r}|\mathbf{c}) - S_{mean}(\mathbf{r})$$

$$S_{max}(\mathbf{r}|\mathbf{c}) = \max[S(r_n|\mathbf{r}_{<n}, \mathbf{c})]$$

$$S_{var}(\mathbf{r}|\mathbf{c}) = \frac{1}{N-1} \sum_{n=2}^N [S(r_n|\mathbf{r}_{<n}, \mathbf{c}) - S(r_{n-1}|\mathbf{r}_{<n-1}, \mathbf{c})]^2$$

Table 1: Surprisal measure definitions: r and c are response and context word sequences. S_{total} , S_{mean} , and $S_{relative}$ are global operationalisations while S_{max} and S_{var} are local.

Peter 2.0: Building a Cyborg

Matthew P. Aylett¹, Ari Shapiro², Sai Prasad³,
Lama Nachman³, Stacy Marsella⁴, Peter Scott-Morgan⁵

¹CereProc Ltd., ²Embodiment Digital, ³Human & AI Systems Research Lab at Intel Corporation

⁴Northeastern University, ⁵Scott-Morgan Foundation

matthewaylett@gmail.com

Abstract

Peter Scott-Morgan has MND/ALS. He is now paralyzed and depends on technology to keep him alive and communicate with others. In this paper we outline the design and creation of a unique communication system driven by an open source eye-tracking interface (ACAT) which aims to preserve Peter's character through: 1. A cloned artificial voice, 2. An animated avatar. We describe the AI techniques adopted, the interface design and integration process for what is fundamentally an applied accessibility AI project. Finally we discuss Peter's use of the word "transitioning cyborg" to describe his take-up of AI support technology.

Index Terms: accessibility, eye-gaze interfaces, speech synthesis, avatars, artificial intelligence

1. Introduction

"It is true that AI technologies have the potential to dramatically impact the lives of people with disabilities. However, widely deployed AI systems do not yet work properly for disabled people, or worse, may actively discriminate against them." - Peter and Laura Smith¹

When Peter Scott-Morgan got in touch with the authors in 2018 he had been diagnosed with MND (termed ALS in the United States, MND in the UK). With MND, motor neurons gradually stop sending messages to the muscles. This leads the muscles to weaken, stiffen and waste. There is no known cure. However *"This was to be Terminal Disease like no one had seen it before!"*² and Peter believed that AI technology could support him in his illness, not just to keep him alive, but to thrive. A crucial part of this support would be:

1. Spontaneous Communication: For Peter to be able to communicate spontaneously to those around him, from his loved ones to his research colleagues after the muscles required to speak stopped working.
2. Personality Retention: That Peter would not just communicate with others but would do so in a way which conveyed is unique personality and self.

In this paper we focus on personality retention.

Despite commercial voice cloning being available since 2014 there was no support for personalized voices in current AAC systems. Furthermore, Peter wanted to have an avatar to speak for him as his face and head muscles would be paralyzed. In addition, the planned system was required to mimic the behavior of an individual as well as offering appropriate control of this behavior for an AAC user.

¹P. Smith and L. Smith, "Artificial intelligence and disability: too much promise, yet too little substance?" AI and Ethics, vol. 1, no. 1, pp. 81–86, 2021.

²Peter Scott-Morgan, Interview at the Hay Festival 2021

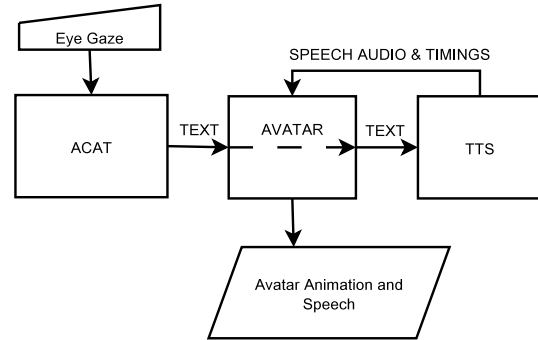


Figure 1: Avatar AAC system

This paper³ describes and discusses the custom system built for Peter Scott-Morgan. This involves an open source software (S/W) platform that enables full PC control through various sensors (eye-tracker in the case of Peter) which Intel Labs first created for Professor Stephen Hawking, a TTS system which clones Peter's voice and offers a rich set of expressive functionality created by CereProc Ltd., and an avatar based on a movie quality 360 degrees body scan created by Embodiment Digital Inc. .

2. Conclusion

The technology discussed here can be regarded as a sub-set of the requirements for a social robot or animated agent. A robot/agent needs to output speech and control its visual appearance. Having a unique personality is often regarded as a requirement. The technology that will be added to Peter's system is also required by such systems: the ability to monitor conversation and suggest sensible responses, the ability to respond rapidly in a conversation, the ability to control conversational dynamics, interrupting, allowing interruption, giving feedback etc. Speech synthesis can be seen as a well described technology that takes text in and produces audio out. But for both AAC use and for social robots/agents, this is not a suitable framework. When interaction becomes important, the way such synthesis integrates with movements and responds to outside events becomes paramount. By doing so we may improve our social robots/agents but more important we can completely rewrite the future of disability. *"As a scientist, AND as a prototype, I'm VERY optimistic about the power of AI and robotics to transform our expectations of what it means to be old. Even in terms of becoming forgetful, or getting dementia. We're at the early dawn of escaping the fear of becoming infirm, of being powerless, of feeling trapped in an inadequate body."* – Peter Scott-Morgan, Fred Hood Memorial Lecture, Edinburgh Book Festival 2021.

³published in full at PETRA '22

Monitoring sleep disordered breathing of long-Covid patients at home using acoustic AI technology

Gerardo Roa Dabike, Guy J. Brown, Ning Ma

Speech and Hearing Research Group (SPandH), University of Sheffield, Sheffield, UK

{g.roadabike, g.j.brown, n.ma}@sheffield.ac.uk

1. Introduction

Obstructive sleep apnoea (OSA) is a chronic and prevalent condition that severely interrupts breathing during sleep, leading to fatigue and increased risk of stroke, heart attack, high blood pressure and diabetes. However, many OSA sufferers remain undiagnosed until these other medical problems become apparent [1, 2]. Polysomnography (PSG), the gold standard for OSA diagnosis, requires patients to sleep overnight in a clinic while wearing many wired sensors, which is uncomfortable, time-consuming and expensive.

There is a recognised prevalence of OSA in patients recovering from Covid-19. And patients that already suffered from OSA are more likely to develop severe Covid-19 [3].

We have collected pre-Covid data and developed a robust machine learning technique for screening OSA acoustically [4, 5], but this has not been evaluated on patients in which OSA is compounded with the symptoms of long-Covid. In this project, we aim to determine if the acoustic of OSA from non-Covid patients differs from those from long-Covid patients and if the developed system is still robust for long-Covid patients.

2. Data Collection

The data is collected by asking participants with long-Covid and OSA to sleep two nights wearing a Home Sleep Apnoea Testing (HSAT) device while recording their sleeping sounds using a smartphone. The HSAT data is scored by a Registered Polysomnographic Technologist and is used as reference for labelling audio during an OSA episode. To synchronise the audio recordings and the HSAT data, the crosscorrelation function of the HSAT snore channel signal and the audio signals containing snore sound was computed over a 20-min segment. The HSAT data was then time-shifted according to the largest delay peak in the crosscorrelation function.

3. OSA Classifier and Dataset

The classifier consists of a convolutional neural network (Figure 1) trained using 30-second audio segments (an apnoea event typically lasts 15-30 seconds) with 20 second overlap. The model was constructed using SpeechBrain and speed augmentation for data augmentation.

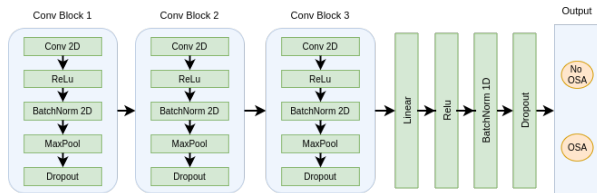


Figure 1: Convolutional Classifier

Table 1 shows the demographics of the Pre-Covid (for training and development) and Long-Covid corpus (for evaluation). So far, 57 long-Covid participants have been recorded and we are targeting 75 participants.

Table 1: OSA Datasets

Pre-Covid	Gender	108 Patients	168 Nights
	Age	60 males	39 females
	Duration	25 - 71	avg: 45 - std: 13
	AHI	3.0 - 9.8	avg: 7.0 - std: 1.4
Long-Covid	Gender	1.2 - 114	avg: 25 - std: 25
	Age	57 Patients	103 Nights
	Duration	29 males	28 females
	AHI	19 - 68	avg: 33 - std: 12
	Gender	6.2 - 9.3	avg: 7.6 - std: 1.2
	Age		
	Duration		
	AHI		

Table 2 shows the preliminary results obtained training the model with Pre-Covid data and ten folds using three different Apnoea Hypopnea Index (AHI) (events per hour) cut-off points.

Table 2: Results using different AHI cut-off points.

AHI cut-off points	5	15	30
Sensitivity	0.80	0.76	0.81
Specificity	0.53	0.63	0.83
AUC	0.72	0.75	0.82

4. Conclusions

Acoustic-based assessment of OSA is promising, offering a robust and less invasive solution for sleep management [4, 5]. Such a solution allows continuous monitoring of sleep at home. This is particularly relevant during Covid-19 when a sleep test is much more likely to take place at home. This study presents some preliminary findings from the long-Covid participants.

5. References

- [1] K. K. Motamedi, A. C. McClary, and R. G. Amedee, "Obstructive sleep apnea: a growing problem," *Ochsner Journal*, 2009.
- [2] A. Castaneda, E. Jauregui-Maldonado, I. Ratnani, J. Varon, and S. Surani, "Correlation between metabolic syndrome and sleep apnea," *World journal of diabetes*, vol. 9, no. 4, p. 66, 2018.
- [3] S. Strausz, T. Kiiskinen, M. Broberg, S. Ruotsalainen, J. Koskela, A. Bachour, A. Palotie, T. Palotie, S. Ripatti, and H. M. Ollila, "Sleep apnoea is a risk factor for severe covid-19," *BMJ Open Respiratory Research*, vol. 8, no. 1, 2021.
- [4] H. E. Romero, N. Ma, G. J. Brown, A. V. Beeston, and M. Hasan, "Deep learning features for robust detection of acoustic events in sleep-disordered breathing," in *ICASSP 2019*, 2019.
- [5] H. E. Romero, N. Ma, G. J. Brown, and E. A. Hill, "Acoustic screening for obstructive sleep apnea in home environments based on deep neural networks," *IEEE J Biomed Health Inform*, 2022.

Incremental Disfluency Detection for Spoken Learner English

Lucy Skidmore, Roger K. Moore

Speech and Hearing Research Group, University of Sheffield, UK

{lskidmorel, r.k.moore}@shef.ac.uk

Abstract

Incremental disfluency detection provides a framework for computing communicative meaning from hesitations, repetitions and false starts found in speech. One application of this research is in dialogue-based computer-assisted language learning (CALL), where detecting learners’ production issues word-by-word facilitates timely and pedagogically driven responses from an automated conversational system. Systematic in their structure, disfluencies comprise of a reparandum phrase, optional interregnum phrase and repair phrase [1]:

I’d like a [$\underbrace{\text{coffee}}_{\text{reparandum}} + \underbrace{\{\text{uh}\}}_{\text{interregnum}} \underbrace{\text{tea}}_{\text{repair}}]$ please.

The disfluency behaviour found in learner speech differs to that of native speech. Disfluencies occur at a higher rate [2], contain longer reparandum phrases [3] and more frequently occur with speaker grammatical errors. With the knowledge that such features can cause performance degradation in existing models [4, 5], it is of interest to explore how these models can be better adapted for learner speech. Following [6], the model used for disfluency detection combines an LSTM with an HMM decoder. Using the NICT-JLE dataset [7], features are extracted incrementally as inputs to the LSTM for training. During testing, the LSTM softmax output layer is used as an input to the HMM where outputs are updated incrementally with the best sequence hypothesis from Viterbi decoding. The baseline model uses word embeddings and part-of-speech tags as features. During experimentation, five model adaptations were tested: dataset lemmatization, automatic edit term removal as well as the inclusion of character embeddings, prosodic cues and learner proficiency scores as input features. Each model iteration was evaluated in terms of its overall performance as well as its detection accuracy for five difficult-to-detect disfluency types (see Table 1 below).

Table 1: Overall repair start (rpS) and reparandum start (rmS) detection F-scores alongside average performance of difficult-to-detect cases for the baseline model compared to subsequent iterations of model adaptation.

Model iteration	overall		length 2+		non-rep		nested		with edits		with errors	
	rpS	rmS	rpS	rmS	rpS	rmS	rpS	rmS	rpS	rmS	rpS	rmS
baseline	0.74	0.67	0.70	0.60	0.66	0.57	0.75	0.68	0.72	0.62	0.55	0.49
+ lemmatization	0.76	0.69	0.70	0.59	0.74	0.68	0.77	0.70	0.74	0.63	0.59	0.53
+ pause & laughter	0.76	0.70	0.71	0.61	0.75	0.68	0.78	0.71	0.75	0.64	0.59	0.53
+ char embeddings	0.77	0.71	0.73	0.62	0.76	0.69	0.78	0.71	0.75	0.64	0.60	0.53
- edit terms	0.78	0.73	0.75	0.65	0.77	0.72	0.80	0.74	0.80	0.72	0.59	0.53
+ learner level	0.79	0.74	0.76	0.66	0.78	0.72	0.81	0.75	0.79	0.71	0.60	0.54

As the results from Table 1 show, when combined, these adaptations not only lead to an overall improvement of model performance but also show increased accuracy for difficult-to-detect disfluencies. This is particularly true for lemmatization, which shows the highest overall improvement as well as significant improvement for non-repetitious disfluencies and disfluencies that co-occur with learner errors. The latter still remain a challenge for the model to detect with the lowest detection rates overall. The results from this work add to the growing body of research on incremental disfluency detection and provide several new avenues for further exploration.

1. References

- [1] W. J. Levelt, “Monitoring and self-repair in speech,” *Cognition*, vol. 14, no. 1, pp. 41–104, 1983.
- [2] N. H. De Jong, M. P. Steinel, A. Florijn, R. Schoonen, and J. H. Hulstijn, “Linguistic skills and speaking fluency in a second language,” *Applied Psycholinguistics*, vol. 34, no. 5, pp. 893–916, 2013.
- [3] Y. Lu, M. J. Gales, and Y. Wang, “Spoken Language ‘Grammatical Error Correction’,” in *Proceedings of Interspeech 2020*, 2020, pp. 3840–3844.
- [4] J. Hough and M. Purver, “Strongly incremental repair detection,” in *Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference (EMNLP)*. Doha, Qatar: ACL, 2014, pp. 78–89.
- [5] R. Moore, A. Caines, C. Graham, and P. Buttery, “Incremental dependency parsing and disfluency detection in spoken learner english,” in *International Conference on Text, Speech, and Dialogue*. Springer, 2015, pp. 470–479.
- [6] J. Hough and D. Schlangen, “Joint, incremental disfluency detection and utterance segmentation from speech,” in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 2017, pp. 326–336.
- [7] E. Izumi, K. Uchimoto, and H. Isahara, “The NICT JLE Corpus: Exploiting the language learners’ speech database for research and education,” *International Journal of the Computer, the Internet and Management*, vol. 12, no. 2, pp. 119–125, 2004.

Audio-Based Computational Analysis of Podcast Expressivity

Shahar Elisha^{1,2}, Emmanouil Benetos¹, Jussi Karlgren², Mariano Beguerisse-Diaz²

¹Queen Mary University of London, UK

²Spotify

s.elisha@se21.qmul.ac.uk

Abstract

Podcasts cover a wide range of topics and can take on various tones and styles: from scripted dramas and heated interviews to spontaneous casual conversations between friends. Our goal is to understand whether detecting such nuanced audio features can improve search and recommendations and be helpful in content moderation settings. In this work, we train and evaluate audio-based models for genre, format, and style detection. Specifically, the tasks we focus on are: (1) comedy detection [1], (2) classification of scripted vs spontaneous speech [2], and (3) style classification [3]. We assess their performance on a large English podcast dataset [4] and address the limitations of these baseline models.

Current research in podcast understanding has primarily focused on natural language processing techniques on metadata or speech-to-text transcriptions [5]; yet podcasts remain a relatively unexplored area compared to other formats [6]. A podcast’s audio signal is rich in paralinguistic information that is not accessible in textual representations. Even with perfect transcription, non-verbal information such as the content’s intensity, rhythm and diction is lost. For example, awkward silences, interruptions, hesitation, interjections (e.g., mmhh, mhhm, hmmm, aha) cannot be captured by transcription. Furthermore, in practice, methods based on speech-to-text have important drawbacks: they are poor at handling speech disfluencies, mispronunciations, and overlapping speech; they do not cover the full range of languages, dialects, and sociolects; and are expensive to run them on large podcast collections.

We focus initially on English language podcasts. Because linguistic and spoken word expressivity vary across cultures, analysing a global podcast catalogue requires understanding these differences. This work is a first step towards a cross-cultural and -lingual analysis of audio-based expression.

1. References

- [1] M. K. Hasan, W. Rahman, A. Bagher Zadeh, J. Zhong, M. I. Tanveer, L.-P. Morency, and M. E. Hoque, “UR-FUNNY: A multimodal language dataset for understanding humor,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 2046–2056. [Online]. Available: <https://www.aclweb.org/anthology/D19-1211>
- [2] A. Veiga, D. Celorico, J. Proença, S. Candeias, and F. Perdigão, “Prosodic and phonetic features for speaking styles classification and detection,” in *Advances in Speech and Language Technologies for Iberian Languages*, D. Torre Toledano, A. Ortega Giménez, A. Teixeira, J. González Rodríguez, L. Hernández Gómez, R. San Segundo Hernández, and D. Ramos Castro, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 89–98.
- [3] N. Ryant and M. Liberman, “Automatic Analysis of Phonetic Speech Style Dimensions,” in *Proc. Interspeech 2016*, 2016, pp. 77–81.
- [4] A. Clifton, S. Reddy, Y. Yu, A. Pappu, R. Rezapour, H. Bonab, M. Eskevich, G. Jones, J. Karlgren, B. Carterette, and R. Jones, “100,000 podcasts: A spoken English document corpus,” in *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 5903–5917. [Online]. Available: <https://www.aclweb.org/anthology/2020.coling-main.519>
- [5] S. Reddy, M. Lazarova, Y. Yu, and R. Jones, “Modeling language usage and listener engagement in podcasts,” *CoRR*, vol. abs/2106.06605, 2021. [Online]. Available: <https://arxiv.org/abs/2106.06605>
- [6] J. Karlgren, “Lexical variation in English language podcasts, editorial media, and social media,” *North European Journal of Language Technology*, vol. 8, no. 1, 2022.

Tandem Multitask Training of Speaker Diarisation and Speech Recognition for Meeting Transcription

Xianrui Zheng*, Chao Zhang, Philip C. Woodland

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

{xz396, cz277, pcw}@eng.cam.ac.uk

Multi-party interactions, such as meetings, are natural scenarios for automatic speech recognition (ASR) applications. Since these scenarios often result in long audio streams, individually trained systems are often required prior to applying an ASR system in order to obtain “who spoken when”. One common approach is to first obtain individual speech segments by finding “who spoke when” using a speaker diarisation system. A speaker diarisation pipeline often consists of at least three modules, namely voice activity detection (VAD), speaker embedding extraction, and clustering, which are normally implemented with separate models [1]. A separately trained ASR system can then be used to transcribe each segment found by speaker diarisation, and obtain speaker-attributed ASR output over long audio streams [2].

To achieve speaker diarisation and speech recognition using a single model, a tandem multitask training (TMT) method is proposed. Since self-supervised-learning-based (SSL) pre-trained models for speech data, such as Wav2Vec 2.0 (W2V2), have become the backbone of many speech tasks, the TMT will take the advantage of the pre-trained knowledge and use a single W2V2 encoder for all tasks. The multitask framework implements voice activity detection (VAD), speaker classification (SC), and ASR with connectionist temporal classification (CTC) using an early layer, middle layer, and late layer of W2V2, which coincides with the order of segmenting the audio with VAD, clustering the segments based on speaker embeddings, and transcribing each segment with ASR.

As shown in Fig. 1, the input to the W2V2 encoder is an entire utterance. While the outputs of the encoder is mapped to the distribution over the subword recognition vocabulary for the ASR task, these outputs can be sampled using a fixed size window and then fed into the average pooling layer in parallel before projecting to the speaker embedding dimension and the final distribution over speakers. In order to include VAD in the multitask training framework, we propose training the VAD task at odd training steps, and training ASR and SC tasks at even training steps. Only feed-forward layers are added for these three tasks to get the desired output dimensions.

Experimental results on the augmented multi-party (AMI) dataset are shown in Table 1. The first line fine-tunes W2V2 encoders for three tasks separately; the second line jointly fine-tunes SC and ASR (SC+ASR); the third line fine-tunes all three tasks together (VAD+SC+ASR). The results are evaluated by diarisation error rates (DER) and cpWER-us where the latter is a modification of cpWER [2] to take into account the mismatch between the estimated and actual number of speakers. Both SC+ASR and VAD+SC+ASR can save computation and stor-

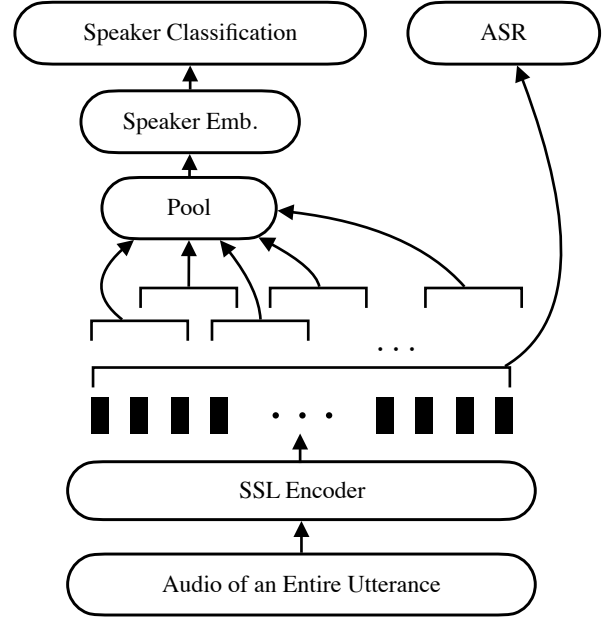


Figure 1: Illustration of the proposed multitask fine-tuning for ASR and SC tasks.

age costs, and performed better than models trained separately. SC+ASR gives relative reductions of 11% and 1% for Eval DER and cpWER-us respectively, and VAD+SC+ASR gives relative reductions of 17% and 3% for Eval DER and cpWER-us respectively.

Table 1: %DER and %cpWER-us with automatic segmentation. ‘ L_n ’ means to use the first to the n -th W2V2 transformer layers for the task. DERs inside the parentheses scored including overlapped regions.

TMT Tasks			%DER		%cpWER-us	
V	S	A	Dev	Eval	Dev	Eval
–	–	–	14.5 (17.9)	12.3 (14.8)	45.3	43.8
–	L_5	L_{12}	11.5 (14.8)	11.0 (13.4)	42.5	43.3
L_1	L_3	L_{12}	11.7 (15.1)	10.2 (12.6)	42.9	42.3

References

- [1] G. Sun, C. Zhang, and P. Woodland, “Combination of deep speaker embeddings for diarisation,” *Neural Networks*, vol. 141, 2021.
- [2] S. Watanabe *et al.*, “CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings,” in *Proc. The 6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020)*, 2020.

*Supported by an Amazon Studentship. This work has been performed using resources provided by the Cambridge Tier-2 system operated by the University of Cambridge Research Computing Service (www.hpc.cam.ac.uk) funded by EPSRC Tier-2 capital grant EP/T022159/1

Comparing Human and Machine Perceptions of Voice Anonymisation

Farida Yusuf⁽¹⁾, Dan Kumpik⁽²⁾, Matt Clifford⁽²⁾, Jonathan Erskine⁽²⁾ and Jennifer Williams⁽¹⁾

⁽¹⁾ University of Southampton

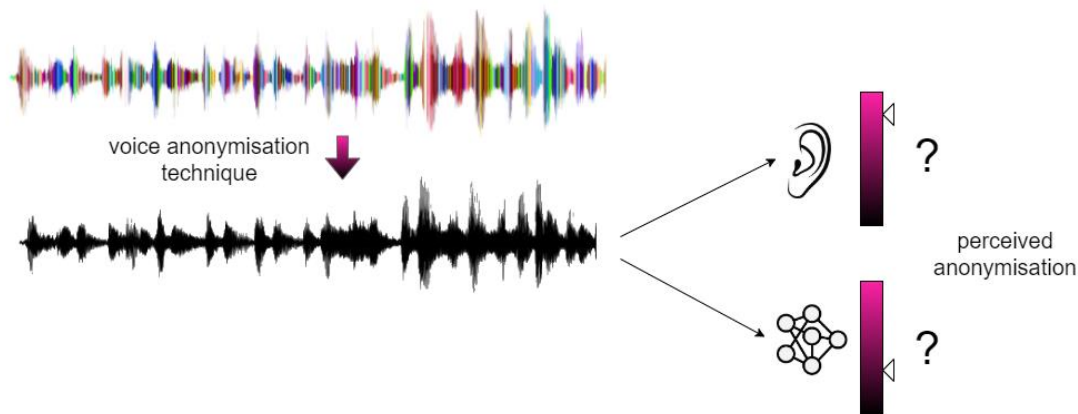
⁽²⁾ University of Bristol

{fy1n18, j.williams}@soton.ac.uk, {dan.kumpik, matt.clifford,
jonathan.erskine}@bristol.ac.uk

Abstract

Voice anonymisation attempts to address issues of privacy. The removal or obfuscation of identifying acoustic characteristics (e.g. features correlated to identity, age, gender, nationality, heritage, location and medical status) from speech is a useful application for several corners of society, e.g. protecting the identity of a legal witness, avoiding copyright and privacy issues over use or reuse of voice in creative productions, preserving confidentiality of a speaker's medical data (due to speech cues of some medical conditions), or storing voice data in compliance with privacy regulations like GDPR.

But who decides if anonymity has been achieved—human or machine? Our proposed work will explore fundamental results about human/AI alignment on perceptions of voice anonymity. To do so, we identify two specific research questions: a) do humans and machines “perceive” anonymised speech in the same way? b) does perceived anonymity (from human judgements) correlate strongly with measured anonymity (from machine AI algorithms)? These are significant questions given that popular uses of voice anonymisation intend to implement privacy to protect voice as a biometric, without considering if voice is also protected from unauthorised human listeners. Alongside the impact of answering to knowledge on voice privacy, the previous examples would benefit from validating voice anonymity between human and machine perceptions.



Results are forthcoming; we will present work in progress and discuss our hypotheses regarding human performance across different discrimination tasks, human versus machine performance on perception, effects of low-level signal modifications versus high-level anonymisation tools, and whether humans are biased in their judgements by AI input. Our proposed methodology will use psychological evaluations obtained through surveys where participants are presented with speech that has been anonymised using a small variety of anonymisation techniques. We will also use a state-of-the-art voice recognition system as a “machine AI algorithm” that provides us with a machine “perception” of anonymisation, through extracted voice prints.

Our goal is to introduce a baseline association of human perception of anonymisation with machine AI anonymisation decisions, using anonymisation techniques that are considered to be competitive within the speech research community. We will also draw from methodologies in psychology to see if it is possible to associate psychometric perception with the performance of machine AI. The impact of this work would thus be, for the first time, to address voice anonymisation in a holistic manner. We expect this to widely inspire further work and research directions concerned with understanding the similarities and differences between human and machine processing of voice anonymisation.

ABAIR-ÉIST: recent progress in Irish language low-resource ASR development

Liam Lonergan¹, Christian Saam^{1,2}, Mengjie Qian³, Neasa Ní Chiaráin¹, Christer Gobl¹, Ailbhe Ní Chasaide¹

¹Phonetics and Speech Laboratory, Centre for Language and Communication Studies, Trinity College Dublin

²ADAPT Centre, Trinity College Dublin

³Engineering Department, University of Cambridge, UK

Abstract

Up until recent months, the focus of Automatic Speech Recognition development for the Irish language as carried out by the AB AIR project in the Phonetics and Speech Laboratory has been on developing hybrid systems. This work has brought very promising results, with Word Error Rates reaching as low as 8%, using a test set partitioned from the training set. In keeping with the current trends of research however, the potential of End-to-End approaches are now being explored. Obstacles relating to the size and nature of the datasets has slowed progress thus far, nonetheless, applying different approaches has seen a steady improvement in the E2E systems' WERs. The issues faced with the data are discussed, as well as the progression of system architectures employed.

Data

The initial source of data for Irish ASR was our in-house recordings for synthesis, which comprised of 18h and 7 speakers of different genders and dialects. To supplement this, a further 20h of recordings were collected from 280 speakers, however, the textual materials used for these recordings were quite limited. This led to a significant amount of utterance texts being recorded multiple times. Therefore, not only is the amount of training data an issue, but also that the data itself is repetitive, which proved to be detrimental to E2E systems.

Efforts were then made to increase the store of acoustic material for training, which has brought some improvements to our systems.

System Architecture Set-ups

- Transformer, character-based
- Conformer, character-based
- Conformer importing decoder from a language model trained on textual data
- Initialising from a pretrained model trained on the Russian Open TTS dataset
- XLSR front-end Conformer

The problems in terms of amount and nature of data faced when transitioning from hybrid systems to End-to-End systems and strategies employed to overcome them will be of interest to the wider community, particularly for low-resource contexts of development.

	Unit	Dur (hrs)	WER: dev/eval
Transformer	char	43h	44.4 / 44.2
Conformer	char	43h	38.0 / 38.1
Conformer with LM Decoder	bpe200	43h	42/42.1
Russian initialized conformer	Bpe100	43h	26.5/26.8
Russian initialized conformer with LM	Bpe100	43h	23.6/24.0
XLSR frontend Conformer	Bpe100	43h	27.1/28.4
XLSR frontend with 130h	Bpe100	130h	19.6/20.9

A summary of the GENE Challenge 2022 on co-speech gesture generation

Youngwoo Yoon^{*1}, Pieter Wolfert^{*2}, Taras Kucherenko^{*3}, Carla Viegas⁴,
Teodor Nikolov⁵, Mihail Tsakov⁵, Gustav Eje Henter⁶

¹ETRI, Republic of Korea ²Ghent University – imec, Belgium ³SEED – Electronic Arts, Sweden
⁴CMU, USA ⁵Umeå University, Sweden ⁶KTH Royal Institute of Technology, Sweden

Humans use both verbal (text, speech) and non-verbal behaviour (facial expression, gestures, stance) to communicate. Embodied conversational agents like virtual avatars and social robots therefore also require the ability to gesticulate, and data-driven gesture generation is currently a highly active research topic. However, research efforts are hard to compare, since there are no established benchmarks and each study tends to use its own dataset, motion visualisation, and evaluation method.

Challenges like the Blizzard Challenges have been of great use for advancing text-to-speech technology [1], inspiring us to conduct the first [2] and second [3] challenge in speech-driven gesture generation, the GENE (Generation and Evaluation of Non-verbal Behaviour for Embodied Agents) Challenges. In these challenges, participating teams built automatic gesture-generation systems using a common dataset. Motion produced by these systems was then evaluated in several large, crowd-sourced user studies using the same motion-rendering pipeline. Differences in evaluation outcomes are then attributable only to differences between the motion-generation methods.

The 2022 challenge dataset was based on 18 hours of full-body motion capture, including fingers, of different persons engaging in dyadic conversations. Only one side of the conversation was considered at a time. Ten teams participated in the evaluation across two tiers: full-body (‘F’) and upper-body gesticulation (‘U’). Not all teams participated in all tiers, and (like in the Blizzard Challenges) results are reported using anonymous labels, prefixed ‘FS’ and ‘US’ for the tiers. The evaluation also included natural motion capture (FNA/UNA) and two baseline systems (UBA and FBT/UBT). For each tier we evaluated both the human-likeness (i.e., quality) of the gesture motion and its appropriateness for the specific speech.

Human-likeness was evaluated by 121 and 150 crowd-sourced test-takers (F and U, respectively) using the MUSRHA-like setup HEMVIP [4], with audio removed from the videos, to control for the effect of the speech on gesture perception. Appropriateness was assessed by 247 and 304 test-takers based

on a novel matched-mismatched paradigm. Here, two videos with the same speech audio but different motion were presented. The motion always came from the same system, but from different speech input: in one case, it came from the speech in the video, but in the other case, it was motion associated with some other, unrelated speech. How often subjects selected the matched video as more appropriate gives an idea of how specific the motion is to the speech. This methodology decouples gesture appropriateness from the human-likeness of the motion, which previously was a major confounder in evaluations [2].

The evaluation results are both surprising and revealing. In each tier, one synthetic condition is rated as significantly more human-like than human motion capture. We believe this has not been shown before on a high-fidelity avatar. On the other hand, all synthetic motion is found to be vastly less appropriate for the speech than the original motion-capture recordings are, even though our new paradigm controls for the influence of motion quality on the evaluation. This mirrors the situation in text-to-speech, where synthetic speech has attained human-like quality, but contextual appropriateness remains to be solved. Data and materials are available through the challenge webpage at <https://youngwoo-yoon.github.io/GENEAchallenge2022/>.

1. References

- [1] S. King, “Measuring a decade of progress in text-to-speech,” *Loquens*, vol. 1, no. 1, p. e006, 2014.
- [2] T. Kucherenko, P. Jonell, Y. Yoon, P. Wolfert, and G. E. Henter, “A large, crowdsourced evaluation of gesture generation systems on common data: The GENE Challenge 2020,” in *Proc. IUI*, 2021.
- [3] Y. Yoon, P. Wolfert, T. Kucherenko, C. Viegas, T. Nikolov, M. Tsakov, and G. E. Henter, “The GENE Challenge 2022: A large evaluation of data-driven co-speech gesture generation,” in *Proc. ICMI*, 2022.
- [4] P. Jonell, Y. Yoon, P. Wolfert, T. Kucherenko, and G. E. Henter, “HEMVIP: Human evaluation of multiple videos in parallel,” in *Proc. ICMI*, 2021.

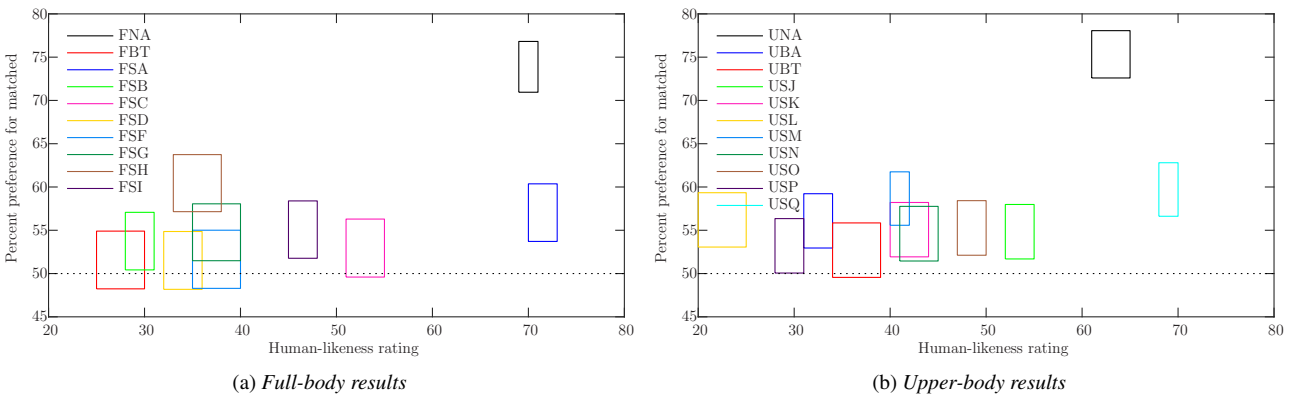


Figure 1: Evaluation results for each tier. Each box is a system. Box widths show 95% confidence intervals for the median human-likeness rating. Box heights show appropriateness, as 95% confidence intervals for the preference for matched motion in percent.

Neural formant synthesis – a proving ground for speech-synthesis control

Gustavo Teodoro Döhler Beck, Ulme Wennberg, Zofia Malisz, Gustav Eje Henter

Division of Speech, Music and Hearing, KTH Royal Institute of Technology, Stockholm, Sweden

1. Technologists and scientists want control

Synthetic speech is approaching a signal quality that is virtually indistinguishable from human speech. This has come to pass by replacing signal processing with deep learning, in effect ceding control over the generated speech to the machine. Human beings, however, exhibit great control over their speech. Therefore, speech technologists have shifted research focus towards regaining control of the output speech – to improve its applicability and appropriateness in the communicative context.

A particularly interesting application of controllable synthesis is the speech sciences, in which synthetic speech has long been an important tool for stimulus creation [1]. However, such applications require highly accurate control over the output. Modern, neural speech synthesis generally does not provide such control. As a result, speech scientists rely on legacy tools e.g., formant synthesizers. Unfortunately, such speech also sounds artificial and is processed and perceived differently from natural speech by humans [2]. This casts doubt on the universality of research findings derived from such stimuli.

2. A first neural formant synthesiser

Our belief [1] is that controllable speech synthesis for the phonetic sciences is a compelling research problem for speech technology and machine learning. To this end, we have published a proof-of-concept method [3] that tries to marry the high quality of neural speech synthesis with phonetically relevant control.

Although current neural-vocoder technologies in principle can learn to convert an arbitrary set of speech features into high-quality waveforms, they require a lot of data and computation to train. In our recent ICASSP paper [3], we simplify the problem by training a small network with CNNs and adversarial losses to generate acoustic features from perceptually relevant speech parameters. One can then leverage a pre-trained neural vocoder to convert these acoustics into audio. This should be easier for resource-constrained applications than training a neural vocoder from scratch. We call our system *Wavebender GAN*.

Wavebender GAN can be trained to create acoustic features from any set of control parameters, but we specifically demonstrate its use by creating the first (to our knowledge) *neural formant synthesiser*, controlled by a minimalist parameter set:

- fundamental frequency (linearly interpolated log f_0 and a binary voicing flag)
- formant frequencies (F1 and F2)
- two measures of voicing quality (spectral centroid and spectral slope)

HiFi-GAN [4] is used as the neural vocoder. Subjective and objective experiments demonstrate impressive speech quality and promising results in terms of control accuracy [3]. Data aug-

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. ZM was supported by the Swedish Research Council grant no. 2017-02861 “Multimodal encoding of prosodic prominence in voiced and whispered speech”.

The authors thank Simon King, Alice Turk, Jens Edlund, and Jonas Beskow for helpful discussions.

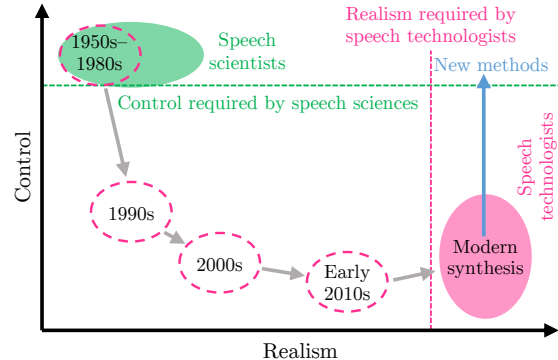


Figure 1: *Speech technologists (pink) and speech scientists (green) have grown apart due to having different priorities. New methods like Wavebender GAN aim to advance both communities, by combining high realism with precise control.*

mentation was found to help disentangle the effect of different speech parameters on the signal, improving the control. For more information, see our ICASSP paper [3] and the project webpage at <https://gustavo-beck.github.io/wavebender-gan/>.

3. Why should you care?

We think neural formant synthesis is interesting as more than an exercise in speech-reconstruction minimalism. In fact, it – and other speech-synthesis problems with phonetically relevant control – 1) present highly demanding test cases suitable for benchmarking the accuracy of controllable machine learning, 2) has the potential to reshape how phonetic research is done, and 3) asks hard questions about the nature of “independent” control and “individual” cues when dealing with complex and highly entangled phenomena such as speech. For example, f_0 and energy are closely linked in human speech production, so it is not well defined what energy should do when pitch is manipulated. The answer depends on the application. A versatile tool should offer multiple options, cf. how formant amplitude can be dependent or independent of frequency in different versions of the Klatt synthesiser [5]. Grappling with these questions of what we mean by “control” is not only of importance to speech technology and sciences, but to machine learning as a whole.

4. References

- [1] Z. Malisz, G. E. Henter, C. Valentini-Botinhao, O. Watts, J. Beskow, and J. Gustafson, “Modern speech synthesis for phonetic sciences: a discussion and an evaluation,” in *Proc. ICPhS*, 2019.
- [2] S. J. Winters and D. B. Pisoni, “Perception and comprehension of synthetic speech,” *Research on Spoken Language Processing Progress Report*, no. 26, pp. 95–138, 2004.
- [3] G. T. Döhler Beck, U. Wennberg, Z. Malisz, and G. E. Henter, “Wavebender GAN: An architecture for phonetically meaningful speech manipulation,” in *Proc. ICASSP*, 2022.
- [4] J. Kong, J. Kim, and J. Bae, “HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis,” in *Proc. NeurIPS*, 2020.
- [5] D. H. Klatt, “Software for a cascade/parallel formant synthesizer,” *J. Acoust. Soc. Am.*, vol. 67, no. 3, pp. 971–995, 1980.

Empowering neural TTS with HMMs to get the best of both worlds

Shivam Mehta, Harm Lameris, Éva Székely, Jonas Beskow, Gustav Eje Henter

Division of Speech, Music, and Hearing, KTH Royal Institute of Technology, Stockholm, Sweden

In the last decade, text-to-speech (TTS) technology has made tremendous technological advances. The use of more powerful function approximators like deep neural networks as the modelling backbone for speech synthesis has lead to high performance in terms of both intelligibility and naturalness, surpassing the previous paradigms of concatenative synthesis and statistical parametric speech synthesis (SPSS) based on hidden Markov models (HMMs) [1]. Previously, the integration of deep learning with HMM-based TTS systems sacrificed the ability to jointly learn and align, instead relying on an external aligner [2]. However, sequence-to-sequence models with neural attention, like Tacotron [3], reintroduced the ability to jointly learn and align and, leveraging neural vocoders, can generate remarkably natural-sounding synthetic speech.

Unfortunately, neural attention mechanisms are often non-monotonic in nature, and do not enforce the correct order of the spoken phonemes. The synthesis is then susceptible to skipping and stuttering artefacts, and might even break down into unintelligible gibberish [4]. Additionally, neural attention is a data-hungry mechanism that requires a substantial amount of training data to learn the alignments. While most components of a neural TTS system – the front-end encoder, the intermediate audio representation (mel-spectrograms), and acoustic feedback provided by autoregression – grant a major performance boost in the synthesis quality, attention rather impedes the ability to learn to generate high-quality speech quickly [4].

In our recent ICASSP paper [5], we propose to fuse the best components of the two paradigms, SPSS and neural TTS so that they complement each other. Specifically, **we replace neural attention** (as in, e.g., Tacotron 2 [6]) **with a left-to-right no-skip HMM defined by a neural net**. This enables a *neural HMM TTS* approach with several beneficial features:

1. A learnt front-end encoder, like in neural TTS. This replaces SPSS rule-based text processing and improves prosody and pronunciation of out-of-vocabulary words.
2. Mel-spectrograms as the acoustic features. This allows using pre-trained neural vocoders for better-quality waveforms and also avoids overly flat prosody due to oversmoothing of f_0 , which happens for vocoders with an explicit pitch feature [4, 5].
3. Autoregression for acoustic feedback, like in Tacotron 2 [6]. This means that each acoustic frame can be different (have different statistics) without band aids like dynamic features or frame-level positional features. It also enables much more powerful duration modelling [7].
4. It retains the primary advantage classic SPSS has over TTS with neural attention, namely the left-to-right no-skip HMMs at the core. This enforces a monotonic alignment between phonemes and audio. This inductive bias helps learn to align the text and audio much faster and works better in a low-resource setting.

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

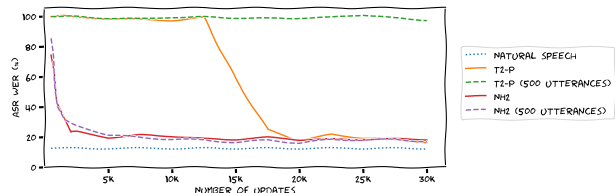


Figure 1: Average ASR WER of validation-set resynthesis.

As an example of the approach, our ICASSP paper [5] demonstrates the effect of replacing the neural attention in Tacotron 2 with a neural HMM. We contrast the resulting system (NH2) against a comparable Tacotron 2 configuration (T2-P), both trained on LJ Speech. A subjective evaluation finds no difference in speech naturalness between the two. To evaluate intelligibility, we synthesised the validation-set sentences every 500 updates during training, transcribed this output with Google Cloud ASR, and show the resulting word error rate (WER) over time in Figure 1. We see that the system with the neural HMM learns to speak after 10 times fewer updates, and managed to learn to speak intelligibly even when the dataset was reduced to a mere 500 utterances. Our example system furthermore has 36% fewer parameters, allows control over speaking rate, and does not break into gibberish on long or short utterances.

Our main takeaway is that speech synthesis using neural HMMs can speed up TTS research and development. It not only enables faster iteration but also improves performance in low-resource scenarios. Audio and code can be found on our webpage at <https://shivammehta007.github.io/Neural-HMM/>.

1. References

- [1] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, “A survey on neural speech synthesis,” *arXiv preprint arXiv:2106.15561*, 2021.
- [2] O. Watts, G. E. Henter, T. Merritt, Z. Wu, and S. King, “From HMMs to DNNs: where do the improvements come from?” in *Proc. ICASSP*, 2016, pp. 5505–5509.
- [3] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyriannakis, R. Clark, and R. A. Saurous, “Tacotron: Towards end-to-end speech synthesis,” in *Proc. Interspeech*, 2017, pp. 4006–4010.
- [4] O. Watts, G. E. Henter, J. Fong, and C. Valentini-Botinhao, “Where do the improvements come from in sequence-to-sequence neural TTS?” in *Proc. SSW*, 2019, pp. 217–222.
- [5] S. Mehta, É. Székely, J. Beskow, and G. E. Henter, “Neural HMMs are all you need (for high-quality attention-free TTS),” in *Proc. ICASSP*, 2022, pp. 7457–7461.
- [6] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyriannakis, and Y. Wu, “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” in *Proc. ICASSP*, 2018, pp. 4799–4783.
- [7] S. Ronanki, O. Watts, S. King, and G. E. Henter, “Median-based generation of synthetic speech durations using a non-parametric approach,” in *Proc. SLT*, 2016, pp. 686–692.

UNSUPERVISED DATA SELECTION FOR SPEECH RECOGNITION WITH CONTRASTIVE LOSS RATIOS

Chanh Park, Rehan Ahmad, Thomas Hain

Speech and Hearing Research Group, University of Sheffield, Sheffield, UK
cpark12, r.ahzzmad, t.hain@sheffield.ac.uk

ABSTRACT

This paper proposes an unsupervised data selection method by using a submodular function [1] based on contrastive loss ratios [2] of target and training data sets. A model using a contrastive loss function is trained on both sets. Then the ratio of frame-level losses for each model is used by a submodular function. By using the submodular function, a training set for

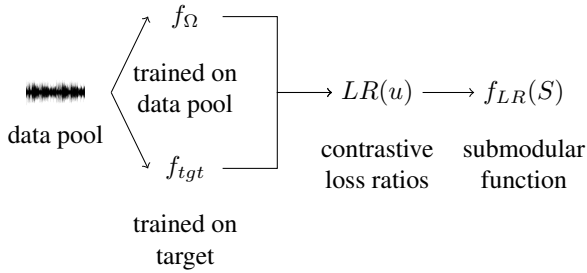


Fig. 1: Contrastive loss ratios and Submodular function

automatic speech recognition matching the target data set is selected. Experiments show that models trained on the data sets selected by the proposed method outperform the selection method based on log-likelihoods produced by GMM-HMM models, in terms of word error rate (WER). When selecting a fixed amount, e.g. 10 hours of data, the difference between the results of two methods on Tedtalks was 20.23% WER relative. The method can also be used to select data with the aim of minimising negative transfer, while maintaining or improving on performance of models trained on the whole training set. Results show that the WER on the AMI and WS0 data set was reduced by 2.92% and 3.93% relative when selecting 90% from the whole data set.

Acknowledgements This work was conducted at the VoiceBase Research Centre for Speech and Language Technologies at the University of Sheffield, which is funded by VoiceBase Inc.

Table 1: Data pool

	AMI	Fisher	Tedtalk s	Wsjcam0
Hours	10	10	10	10
Segments	3526	3330	3244	3685

Table 2: Numbers of selected segments by contrastive loss ratios (CLR) and log-likelihood (LL).

target data set	hours of subset (CLR/LL)			selected data set
	10h	20h	30h	
TD	103/1385	1524/2250	2797/2899	AMI
	362/162	1789/781	2686/1807	FS
	2773/1100	3181/2099	3219/2779	TD
	0/720	152/1662	1471/2781	WS0

Table 3: WERs(%) on selected data sets.

Method	target	10h	20h	30h	40h
CLR	TD	28.07	25.54	24.43	24.58
LL	TD	35.19	28.37	26.42	24.58

Table 4: WERs(%) on selected data sets for negative transfer minimisation by contrastive loss ratios.

Target	80%	85%	90%	95%	100%
AMI	26.98	26.79	25.91	26.35	26.69
FS	35.83	36.96	35.83	35.72	35.72
TD	24.97	25.25	24.94	24.34	24.58
WS0	9.66	9.71	9.51	9.66	9.90

1. REFERENCES

- [1] A. Krause, and D. Golovin, “Submodular function maximization,” in *Tractability: Practical approaches to hard problems*, L. Bordeaux, Y. Hamadi and P. E. Kohli, Eds., p. 71–104. Cambridge University Press, 2014.
- [2] S. Schneider, A. Baevski, R. Collobert and M. Auli, “wav2vec: Unsupervised pre-training for speech recognition,” in *Proc. Interspeech 2019*, Graz, Austria, pp. 3465–3469.

Domain-Informed Probing of wav2vec 2.0 Embeddings for Phonetic Features

Patrick Cormac English¹, John D. Kelleher², Julie Carson-Berndsen¹

SFI Centre for Research Training in Digitally-Enhanced Reality (d-real)

¹ADAPT Research Centre, School of Computer Science, University College Dublin, Ireland

²ADAPT Research Centre, Technological University Dublin, Ireland

patrick.english@ucdconnect.ie, john.d.kelleher@tudublin.ie, julie.berndsen@ucd.ie

Introduction

In recent years large transformer model architectures have become available which provide a novel means of generating high-quality vector representations of speech audio. These transformers make use of an attention mechanism to generate representations enhanced with contextual and positional information from the input sequence. Previous works have explored the capabilities of these models with regard to performance in tasks such as speech recognition and speaker verification, but there has not been a significant inquiry as to the manner in which the contextual information provided by the transformer architecture impacts the representation of phonetic information within these models. In this paper, we report the results of a number of probing experiments on the representations generated by the wav2vec 2.0 model’s transformer component, with regard to the encoding of phonetic categorization information within the generated embeddings. We find that the contextual information generated by the transformer’s operation results in enhanced capture of phonetic detail by the model, and allows for distinctions to emerge in acoustic data that are otherwise difficult to separate.¹

1. Experimental Overview

The experimental set-up for this work can be seen in Figure 1. From the TIMIT spoken-English corpus and the outputs of the wav2vec 2.0 transformer component (provided with audio from the TIMIT dataset) a composite dataset is generated containing both the TIMIT phoneme label and the vector representation of the respective audio timestep generated by the transformer. The vector representation is the average of all the timestep representations that occurred within the phoneme window as per TIMIT’s timing information. This dataset is then provided to a single-layer MLP network with 200 RELU activations, which is trained for the task of identifying the phoneme label from the vector representations. This forms the basis of a probing task (in the vein of Conneau et al. 2018) whereby the model’s predictive accuracy is taken as indicative of whether the embeddings encode information relevant to the task target (phoneme identification, in this instance).

2. Results

From the above probing task a number of outputs were derived: 1) Prediction accuracies across 3 categories (Manner of Articulation, Place of Articulation, Phoneme Labels), the former two of which were derived via mapping from the phoneme labels; 2)

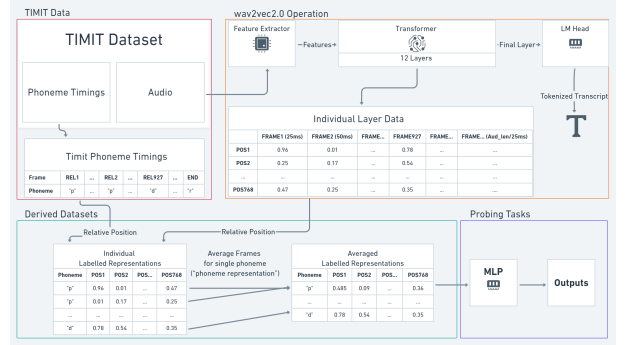


Figure 1: Overview of Experimental Set-up

Confusion Matrices, visualised as heatmaps, which display the frequency of phoneme label prediction errors; and 3) Hierarchical Clustering Dendrograms computed via Ward’s method from the MLP confusions (with the confusions input as probabilities), in which proximity in the hierarchy can be understood as representing “similarity” between two phoneme labels in the probe’s output. As an example of these outputs, Figure 2 illustrates the dendrogram visualisation of the MLP outputs generated at the final layer of the transformer.

While the specific nature of the phonetic information captured by modern large transformer models will require significant further work to adduce, we have demonstrated that there is significant evidence to suggest that transformer architectures are capable of capturing significant levels of phonetic detail that accords with domain-informed understandings of phoneme relationships, and that permit distinction between less separable phonemes. There are several patterns of interest captured in the hierarchical view in particular, with one salient example being the initial clustering of the “acoustically” similar closures for stops such as “dcl” and “tcl” in layer 0 of the transformer, which had migrated to co-locate with their respective burst components in the final layer of the transformer (with final positions visible in figure 2 below).

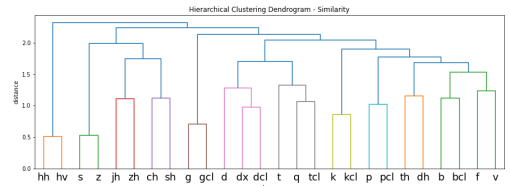


Figure 2: Dendrogram visualisation of obstruent similarities at final layer of transformer

¹This paper has been published at the 19th SIGMORPHON Workshop 2022. This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224

Self-supervised Graphs for Audio Representation Learning with Limited Labeled Data

Amir Shirian¹, Krishna Somandepalli², Tanaya Guha^{1,3}

¹University of Warwick ²Google Research ³University of Glasgow

1. Executive Summary

Motivation: Large databases with high-quality manual labels are scarce in audio domain. For tasks such as speech-based emotion analysis, manual labels are often difficult to acquire due to the subjectivity involved in the perception and expression of emotion across speakers, language and culture. On the other hand, for tasks such as acoustics event classification, manually labeling a large volume of audio data is simply impractical. Thus a core challenge in audio analysis is to learn from a limited amount of labeled data while taking advantage of larger amount of unlabeled training samples.

The idea: We propose a Self-Supervised Learning (SSL) approach on *graphs* to learn effective audio representations from limited amount of labeled data. Considering each audio sample as a node in a graph, we cast audio classification as a node labeling task.

Why graphs: Graphs are a compact, efficient, and scalable way to represent data. The motivation behind adopting a graph approach to model speech is two-fold: (i) Leads to compact, high performing models as compared to commonly used recurrent models as shown in recent works (e.g., [1]); (ii) A graph structure, if properly constructed, can efficiently capture the relationship between the small number of available labeled nodes and a larger number of unlabeled nodes.

Contributions: (1) A subgraph-based learning framework for semi-supervised, self-supervised audio representation learning with limited labeled data. (2) A new graph SSL task, called graph shuffling, and a new variant of graph denoising SSL task. (3) SOTA performance for *acoustic event classification* and *speech emotion recognition* on 3 benchmark databases that is better or on par with fully supervised models.

2. Approach

Our model consists of an audio feature encoder, a subgraph construction step and a multitask-SSL architecture with new pretext tasks and loss functions. Below is a very brief explanation of our approach; for details please see [2].

Audio feature encoder f : $\mathcal{S} \rightarrow \mathcal{Z}$ takes raw audio $\mathcal{S}$ as input and returns embedding $\mathcal{Z}$. The embeddings $\mathcal{Z}$ are used as node attributes in subgraphs $\mathcal{G}$ that we later construct. We use, $\mathcal{Z} = \text{low-level descriptors (LLDs)}$ for speech and $\mathcal{Z} = \text{log-spectrogram convolutional features}$ for non-speech audio data.

(Sub)graph construction: Instead one large graph containing all training samples, we propose to construct and train on sub-graphs $\mathcal{G}_s$. To construct $\mathcal{G}_s$, we randomly select a subset of labeled nodes (equal number of samples from each class) and a subset of unlabeled training nodes. This ensures the degrees of the nodes do not vary too much and class balance is maintained in each subgraph. The edges are connected by computing the

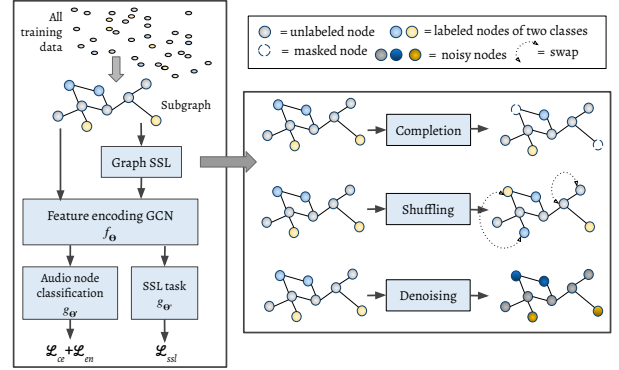


Figure 1: Our approach: A subgraph-based audio representation learning framework with SSL task.

nearest neighbors using $\mathcal{Z}$.

Subgraph SSL training: We adopt an auxiliary learning paradigm to merge self-supervision into the main task of audio classification. This is done by jointly optimizing for node classification and an auxiliary graph SSL task. Our model is *inductive*, i.e., neither attributes nor edges of the test nodes are present during the training process. Loss functions used are cross-entropy, entropy regularization (proposed) and SSL loss.

Table 1: Acoustic event classification results on *AudioSet*.

Model	mAP	Params
<i>Semi-supervised (10% labels)</i>		
Ours w/o SSL	0.23 ± 0.01	218K
Ours w/ denoise	0.26 ± 0.00	260K
Ours w/ completion	0.27 ± 0.01	260K
Ours w/ shuffle	0.24 ± 0.00	219K
Spectrogram-VGG	0.16 ± 0.05	6M
AST2021	0.22 ± 0.01	88M
<i>Fully supervised</i>		
Ours w/o SSL	0.42 ± 0.02	218K
Spectrogram-VGG	0.26 ± 0.01	6M
DaiNet2017	0.25 ± 0.07	1.8M

Code: github.com/AmirSh15/VAED-HeterGraph

3. References

- [1] A. Shirian and T. Guha, “Compact graph architecture for speech emotion recognition,” in *ICASSP*. IEEE, 2021, pp. 6284–6288.
- [2] A. Shirian, K. Somandepalli, and T. Guha, “Self-supervised graphs for audio representation learning with limited labeled data,” *IEEE Journal of Selected Topics in Signal Processing (To appear)*, 2022.

Oral Session A

Evaluating watchability for video localisation

*Zack Hodari¹, Tian Huey Teh¹, Vivian Hu¹, Tomás Gómez Ibarondo¹,
Devang S Ram Mohan¹, Alexandra Torresquintero¹, Chris Wallis¹,
James Leoni¹, Simon King^{1,2}*

¹Papercup, London, UK

²The Centre for Speech Technology Research, University of Edinburgh, UK

zack@papercup.com

Abstract

Video is the most consumed form of media. Unfortunately, most video content is locked in one language, making it inaccessible to at least 80% of people. While video localisation exists to bridge this gap, traditional dubbing is expensive, meaning most content remains stuck in its source language. Machine learning promises a more scalable solution to this, though video localisation poses many challenging problems, including: translation, localisation of translations, voice selection, isochrony (translation length and speech timing), prosody modelling and control, placement of non-speech like breaths and laughter, matching background acoustic qualities like room tone and background audio, on-screen text translation, and lip syncing.

In order to apply machine learning to solve these challenges it is important to measure the impact on end-viewer experience. We call this metric watchability. Watchability includes all aspects that contribute to the end-viewer experience. Like many subjective concepts, watchability is difficult to measure. We have explored approaches across two paradigms to evaluate watchability: narrow evaluations and in-context evaluations. The narrow paradigm relies on focused studies that evaluate specific concepts or single components of the system, such as naturalness, acoustic quality, prosody quality/appropriateness, pronunciation accuracy, and voice likability. While these are precise, it is hard to assess the relative importance of each concept across evaluations. The second paradigm, in-context evaluation, evaluates fully dubbed and produced videos. However, full-length videos pose challenges relating to cognitive biases such as the recency effect and anchoring.

In addition to measuring watchability, it is important to understand the relative contribution of the factors that make up watchability. These relative contributions can help inform which challenges are most worth investing research effort into. At Papercup, we make the assumption that the lowest-hanging fruit to improve watchability is controllable TTS. We believe other challenges, such as paraphrasing for localisation of translations, are better solved by humans in the short term. We present two research paths that could be pursued to uncover relative contributions to watchability, and we invite discussion to further explore other evaluation paradigms.

Transforming adult to child speech for dubbing

Protima Nomo Sudro, Anton Ragni, Thomas Hain

The University of Sheffield, UK

{p.nomo.sudro,a.ragni,t.hain}@sheffield.ac.uk

Abstract

Child actors play an important role in movies, cartoons, and dubbing. Dubbing typically involve translation of original dialogue from the films based on lip movement, tone and script. Though human voice is the highest standard with respect to translated audio quality. However, dubbing is an expensive and time-consuming mode of audio translation due to complexities and many dependent professionals it requires. With the advances in technology, dubbing became easier and cost-effective. In this direction, we study whether voice conversion (VC) method can benefit dubbing adult to child speech. This work exploit CycleGAN based VC framework and CMU kids corpora is considered for child speech data. Analysis of child speech data showed that it consists of non-speech segments like laugh, hesitation, false start, disfluencies, long pause in between the utterance. The investigation of the dataset led us to perform speech segmentation for training VC model. The subset of CMU corpus is created by removing non-speech segments from each of the utterance. For the pre-processing, we utilize the existing transcriptions released with the dataset. The speech files having non-speech segments in word boundaries greater than 100 ms are splitted. If the non-speech segments exceeds 200 ms in between words, then those regions are also segmented. The subset of CMU corpus after segmentation contained 6.30 hours of speech data instead of 9.1 hours. Collecting parallel data from professional adult and child actors is a challenging task. Initially, parallel adult speech data is acquired using a text-to-speech model Fastspeech 2. Additionally, parallel adult speech data is also obtained from a female voice actor, recorded in Zoo Digital studio, London. The effectiveness of the VC system for different input combinations are evaluated using WER and MCD values. When segmented speech is used to train the VC model, we observe improved performance compared to unprocessed child speech data ¹.

¹The authors would like to thank ZOO Digital and Knowledge Transfer Partnerships, Innovate UK for the funding

Deciphering Speech: a Zero-Resource Approach to Cross-Lingual Transfer in ASR

Ondřej Klejch, Electra Wallington, Peter Bell

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

{o.klejch, electra.wallington, peter.bell}@ed.ac.uk

We present a method for cross-lingual training an ASR system using absolutely no transcribed training data from the target language, and with no phonetic knowledge of the language in question. Our approach uses a novel application of a decipherment algorithm, which operates given only unpaired speech and text data from the target language. We apply this decipherment to phone sequences generated by a universal phone recogniser trained on out-of-language speech corpora, which we follow with flat-start semi-supervised training to obtain an acoustic model for the new language. To the best of our knowledge, this is the first practical approach to zero-resource cross-lingual ASR which does not rely on any hand-crafted phonetic information. We carry out experiments on read speech from the GlobalPhone corpus, and show that it is possible to learn a decipherment model on just 20 minutes of data from the target language. When used to generate pseudo-labels for semi-supervised training, we obtain WERs that range from 32.5% to just 1.9% absolute worse than the equivalent fully supervised models trained on the same data.

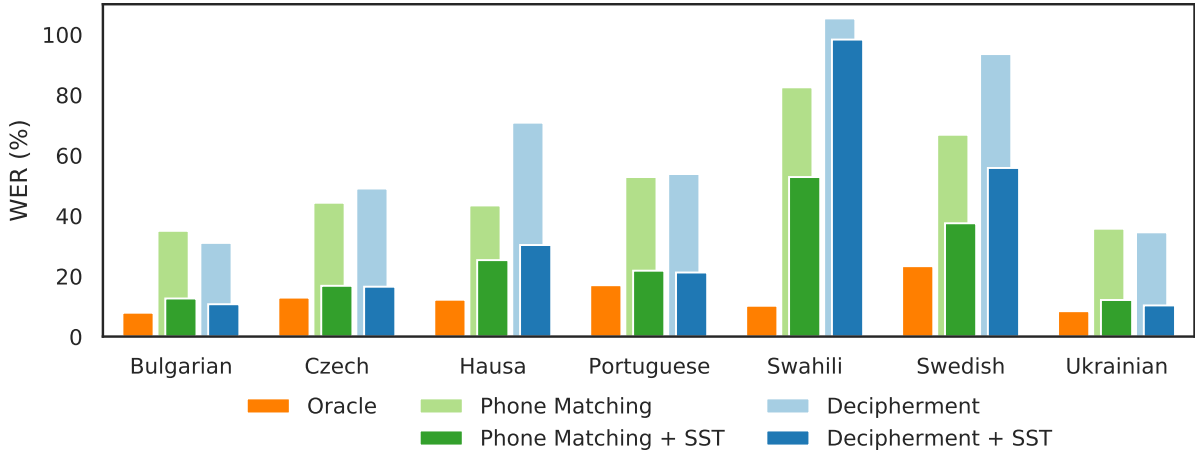


Figure 1: Word error rate (WER) for experiments conducted on the GlobalPhone corpus. Phone Matching corresponds to experiment in which we used linguistic knowledge to find the closest phone recognised by an universal phone recogniser for each phoneme in the target language.

Poster Session B

Person-specific automatic speaker recognition: understanding the behaviour of individuals for applications of ASR

Vincent Hughes¹, Paul Foulkes¹, Philip Harrison¹, Jessica Wormald¹, Chenzi Xu¹, David van der Vloed² and Finnian Kelly³

¹Department of Language and Linguistic Science, University of York, UK.

²Netherlands Forensic Institute, Netherlands.

³Oxford Wave Research, Oxford, UK.

{forename.surname}@york.ac.uk, d.van.der.vloed@nfi.nl, finnian@oxfordwaveresearch.com

Abstract

This paper introduces a new ESRC-funded project called [*Person-specific automatic speaker recognition: understanding the behaviour of individuals for applications of ASR*](#) (ES/W001241/1). The project will run from 2022 to 2025 and involves collaboration between the University of York, the Netherlands Forensic Institute and Oxford Wave Research.

1. Person-specific ASR

The project will further our understanding of why some speakers are particularly easy or difficult for automatic speaker recognition (ASR) systems to handle, particularly in the context of forensic voice comparison. In doing so, we will assess the performance of different systems for individual speakers and develop methods to process problematic types of speakers. The project has four central research questions:

1. What systematic properties of speakers affect the ASR scores, in terms of voice (e.g. pitch, voice quality) and demographic factors (e.g. accent, ethnicity, age, sex)? And how do the magnitudes of these effects compare to known technical effects?
2. How consistent are results for individual speakers within and across ASR systems?
3. How do results produced by techniques that combine ASR and linguistic methods on a person-specific basis compare with the current one-size-fits-all approach?
4. How generalisable are methods and results across datasets and languages?

2. Workpackages

The project is organised around three workpackages. In workpackage (1) we will collect controlled recordings of phoneticians in an anechoic chamber. Phoneticians will vary aspects of their vocal output across a number of conditions, including long term vocal settings (e.g. larynx height and velopharyngeal settings) and accent guises. Additional vocal conditions may be generated after the recording session (e.g. pitch manipulation). The vocal conditions are repeated within sessions, and for each participant at least three sessions are completed, minimally a week apart. Material will be recorded simultaneously with mobile-to-VOIP and three microphones at different distances from the participant to recreate common forensic conditions. We will compare within- and between-session scores to assess the relative effects of speaker

variation, technical variation, and random occasion-to-occasion variability on ASR results.

In workpackage (2), we will conduct much larger scale testing of a range of different ASR systems varying elements of the ASR processing and calibration using databases of English (Home Office CONTEST Collection) and Dutch (NFI FRIDA, see e.g. [1]). Alongside this, we will extract acoustic information from the English speech to characterise speakers and varieties within this set of corpora. We hope to extract features including by-vowel formant distributions (F1-F2), long term F3 distributions and f_0 measures. This will allow us to assess distributions of features frequently analysed in an auditory-acoustic speaker comparison and assess whether any results can help us to account for observed patterns in the ASR output. We also hope to provide access to the acoustic information for practitioners and researchers to improve typicality assessments in real casework (for example, through an RShiny application [2]).

Finally, in workpackage (3), we will develop novel methods to systematically integrate ASR and expert linguistic analyses. This will involve flagging comparisons containing problematic speakers for the ASR system, subjecting them to more detailed linguistic analysis, and then validating the entire approach.

3. References

- [1] D. van der Vloed, F. Kelly, and A. Alexander, “Exploring the effects of device variability on forensic speaker comparison using VOCALISE and NFI-FRIDA, A forensically realistic database,” presented at the The Speaker and Language Recognition Workshop (Odyssey 2020), Nov. 2020. doi: 10.21437/odyssey.2020-57.
- [2] Winston Chang and Joe Cheng and JJ Allaire and Carson Sievert and Barret Schloerke and Yihui Xie and Jeff Allen and Jonathan McPherson and Alan Dipert and Barbara Borges, “shiny: Web Application Framework for R.” 2022. [Online]. Available: <https://CRAN.R-project.org/package=shiny>

Speaker identification in courtroom contexts: performance of human listeners compared to a state-of-the-art forensic voice comparison system

Philip Weber¹, Nabanita Basu¹, Agnes S. Bali², Claudia Rosas-Aguilar^{1,3}, Gary Edmond⁴, Kristy A. Martire², Geoffrey Stewart Morrison¹

¹ Forensic Data Science Laboratory, Aston University, Birmingham, UK

² School of Psychology, University of New South Wales, Sydney, Australia

³ Instituto de Lingüística y Literatura, Universidad Austral de Chile, Valdivia, Chile

⁴ School of Law, Society & Criminology, University of New South Wales, Sydney, Australia

p.weber1@aston.ac.uk

Abstract

Expert testimony in court is only admissible if it will assist the trier of fact (e.g. judge or jury) to make a decision that the trier of fact would not be able to make unaided. This research addresses the question of whether speaker identification by a judge listening alone (to audio recordings) or by a jury listening as a collaborative group would be more or less accurate than the output of an expertly operated forensic-voice-comparison system that is based on state-of-the-art automatic-speaker-recognition technology.

Individuals listening alone and groups of collaborating listeners made judgements on pairs of recordings, each of which was either a same-speaker pair or a different-speaker pair. The recordings were taken from the *forensic_eval_01* dataset [1]. They were of adult male Australian English speakers and reflected the conditions of the recordings in an actual forensic case. Individual listeners came from three groups differing in their degree of familiarity with the language and accent spoken: 1) Australian-English listeners who are familiar with the language and accent spoken on the recordings; 2) North-American-English listeners who are familiar with the language but less familiar with the accent; and 3) Spanish-language listeners who are less familiar with the language. Collaborating groups of listeners will consist only of listeners from Australia. (This part of the research is yet to be completed.)

Approximately 60 listeners from each listener group were presented with 61 pairs of recordings. Each recording was reduced to a randomly-selected 15-second section. For each pair, listeners were required to record a numeric response indicating how much more likely they thought the properties of the voices on the recordings were “if they are both recordings of the same adult male Australian-English speaker” vs “if they are recordings of two different adult male Australian-English speakers”. These responses were converted to log likelihood ratios (LLR). Listeners’ responses were compared with LLRs output by the E3 Forensic Speech Science System (E³FS³ [2]). The system was trained and calibrated-cross-validated in accordance with the procedures described in [2], to obtain LLRs for the same set of pairs of recordings.

For each listener and for the automated system the Cost of Log Likelihood Ratio (C_{llr}) metric was calculated. The C_{llr} penalizes a system (human or automated) for misleading LLRs (supporting the incorrect hypothesis) and weak LLRs (close to 0). A C_{llr} value closer to 0 indicates a better-performing system. Metrics were also calculated to compare the scale and shift

(bias) of the listener’s LLRs relative to those of the forensic-voice-comparison system.

Figure 1 summarises the C_{llr} values obtained for each listener group. All groups exhibited large inter-listener variability, and all performed worse than the system ($C_{llr} = 0.42$). Broadly-speaking, Australian-English listeners performed better than North Americans, who performed better than Spanish-language listeners. Listeners in general were biased towards giving responses that favoured the different-speaker hypothesis, while the system was well calibrated.

This research is ongoing, to examine the behaviour of collaborating groups of listeners (proxies for juries), the effect of various biasing factors, and to further analyse the results in the context of listeners’ beliefs about their own speaker-identification abilities.

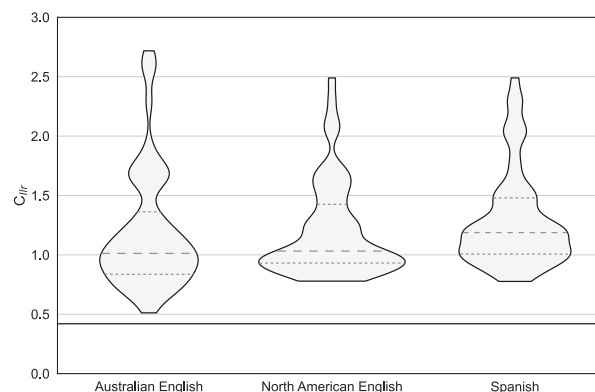


Figure 1: C_{llr} responses from humans (violin plots) and the E³FS³ system (horizontal line at $C_{llr} = 0.42$).

Acknowledgements: This research was funded by Research England’s Expanding Excellence in England fund.

References

- [1] Morrison G.S., Enzinger E. (2016). Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (*forensic_eval_01*) – Introduction. *Speech Communication*, 85, 119–126.
- [2] Weber P., Enzinger E., Labrador B., Lozano-Díez A., Ramos D., González-Rodríguez J., Morrison G.S. (2022). Validation of the alpha version of the E3 Forensic Speech Science System (E³FS³) core software tools. *Forensic Science International: Synergy*, 4, article 100223.

Automatic generation of accented speech using phonetic features

Margot Masson, Anthony Ventresque, Julie Carson-Berndsen

SFI Centre for Research Training in Digitally-Enhanced Reality (d-real)¹,
School of Computer Science, University College Dublin, Dublin, Ireland

margot.masson@ucdconnect.ie, anthony.ventresque@ucd.ie, julie.berndsen@ucd.ie

Overview. It is well known that despite the great progress made in the field, the accuracy of automatic speech recognition (ASR) for accented speech is still lower than that of unaccented speech. The lack of accented speech data has become the main challenge to overcome in the development of accent-robust ASR systems.

Since the task of collecting and labelling accented speech data is laborious and expensive, a currently explored solution is the automatic synthesis of accented speech. This work is part of this approach, and proposes to mimic accented speech by introducing small variations in speech. The work presented in this paper focuses on non-native accents, where variations are due to the approximation by speakers of phonemes which do not exist in their native language, by existing ones perceived as similar. For instance, replacing the [ð] sound - corresponding to the grapheme sequence “th” like in “those” - by [z] or [d] mimics the way French people tend to pronounce [ð], which does not exist in French.

The implementation of this method requires the definition of a similarity measure between phonemes, as well as a definition of the target accent specific phonemes. The similarities between phonemes can be defined in several ways, from knowledge-based methods to learned embeddings. In the work presented in this paper, we chose to use a similarity measure based on the phonetic features of the IPA chart. This method, described in the next paragraphs, has been implemented and is currently being evaluated, as outlined in the last paragraphs.

Methodology. Given an initial text dataset, our speech synthesis method consists of, first, transforming the text into phonemes, applying variations to the phoneme sequence according to the target accent and finally giving the varied sequence to the text to speech (TTS) system that will generate the audio files. The core of this method lies in the way we apply variations to the phoneme sequence, namely using a compatibility matrix to select the phonemes to vary along the original phoneme sequence, and after that, using a second matrix - the similarity matrix - to choose the replacing phoneme, which is one of the most similar ones to the original phoneme.

The construction of the *phonetic compatibility matrix* is very straightforward. It has been built as a boolean matrix, associating the different languages with all possible phonemes, the values being True if the phoneme exists in the target language and False otherwise. Thus, the False values for English phonemes in the matrix correspond to the phonemes that non-native speakers are very likely to approximate by another phoneme when speaking in English. Besides, False values for other phonemes than the English ones only indicate that these phonemes do not exist in the considered language.

The construction of the *similarity matrix* is based on the phonetic features information of the IPA. Thus, the phonemes have been positioned in a three dimensional space (figure 1), the three axes corresponding to the *place of articulation*, the *manner of articulation* and the *voicing*, and their coordinates in this space have been used to calculate the distance between them, as a similarity measure. This construction highlights the positional similarity of the phonemes. For instance, in this space, the coordinates of [ð] are (3,5,1) and the coordinates of [z] are (4,2,1), which result in a Euclidean distance of 3.16 in a space where biggest distance is 13, resulting in a normalised distance of 0.24 (0.76 in similarity).

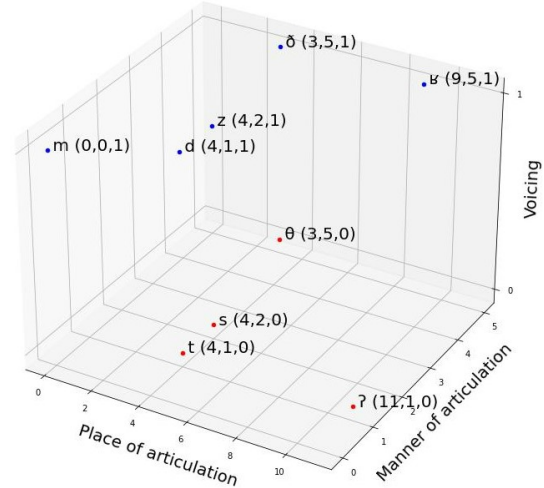


Figure 1: 3D representation of some phonemes

Evaluation. In order to evaluate the quality of the accents that have been simulated using phonetic features, two different experiments are proposed. The first experiment consists of comparing the results - word error rate and word and phoneme alignment - of the recognition for “real accents” versus synthesised ones. This experiment offers a first insight into the quality of the accent by showing the impact of the variations on ASR systems, which should be equivalent to the impact that real accent normally has on speech recognition.

The second experiment will be shaped as a human questionnaire, that will evaluate the credibility of the generated accents, especially by asking respondents to recognize and rate the accents in several audio tracks. For validity purposes, real accents will be proposed and several TTS tools will be used to generate the artificial accents. This experiment will provide more insight into how the generated accents are perceived.

These two experiments are still a work in progress; however, the initial results are promising and show that the artificial data are actually challenging ASR systems.

¹ This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224.

Exploring hidden speech representations of self-supervised automatic speech recognition models

Tamara Soloveva¹, Ramon Sanabria², Peter Bell³

¹MSc Speech and Language Processing

²The Centre for Speech Technology Research

³The Centre for Speech Technology Research

T.Soloveva@sms.ed.ac.uk, R.Sanabria@ed.ac.uk, Peter.Bell@ed.ac.uk

Abstract

Although automatic speech recognition (ASR) models using self-supervised learning (SSL) are extremely successful, they are still considered black box models and are not well understood. In this research we explore two recent models developed by Meta AI: Wav2vec 2.0 and HuBERT. We extract the internal representations of these models, perform layer-wise analysis of the hidden features and discover which linguistic properties are encoded.

We use the L2 norm to measure the distance between frames of hidden speech representations. Based on the distance changes we visualise the contents of data arrays, detect points of sudden content alternations and estimate how the length of representation segments correlates with a range of linguistic units obtained via forced alignment: words, syllables, phones, syllable-like linguistic units and groups of phones (Figure 1).

We identify the layers where certain information is accumulated and compare the extent to which linguistic unit boundaries coincide with points of big distance changes between the hidden features frames. These calculations allow to compare the contents of the hidden layers across the models and the linguistic units.

The main findings of our research are that phones are the most often learned units in SSL ASR, the representation segments also align well with rule-based syllables and groups of 2 phones. Layer 23 of HuBERT Large has the best performance for all the units while different layers of Wav2vec 2.0 focus on learning units of different length.

We constructed word embeddings by average-pooling the frames of the representations, they were tested via probing task and showed a reasonably good performance. Ultimately we identified that the representation segments of Wav2vec 2.0 and HuBERT correspond to 1-7 phones and 1-4 respectively (Figure 2).

The findings of this research can be used for improving the performance of the existing models, eliminating biases and avoiding overfitting. Additionally, knowing which layers accumulate certain kinds of linguistic information can help to use the extracted hidden layers for a variety of downstream tasks in ASR and speech synthesis.

Index Terms: speech recognition, self-supervised learning, HuBERT, Wav2vec 2.0

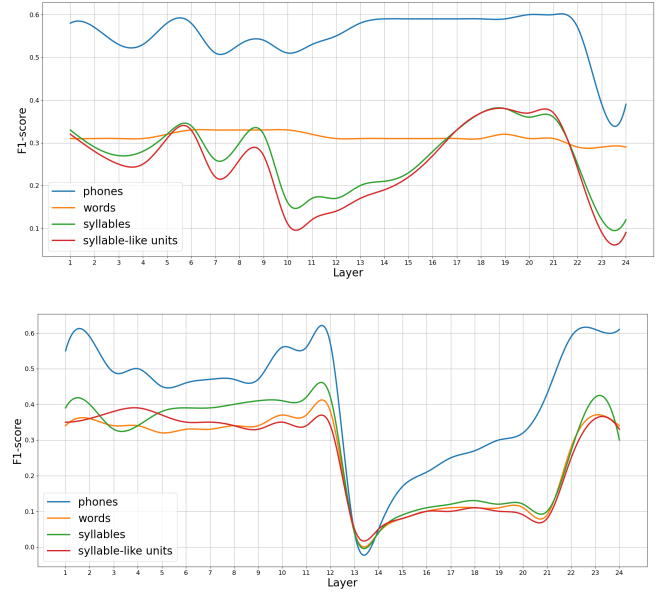


Figure 1: *F1 scores for speech units of different length for every layer of Wav2Vec large (top) and HuBERT Large (bottom)*

	over	the	fireplace	there	was	a	large
W2V	OW1 V ER0 DH AH1	F AY1 ER0	P LEY2 S	DH EH1 R W AH1 Z AH0 L	AA1 R JH		
HB	OW1 V ER0 DH AH1	F AY1 ER0	P LEY2 S	DH EH1 R W AH1 Z AH0 L	AA1 R JH		

	old-fashioned	gilt	mirror
W2V	OW1 L D F AE1 SH AH0 N D	G IH1 L T M IH1 R ER0	
HB	OW1 L D F AE1 SH AH0 N D	G IH1 L T M IH1 R ER0	

Figure 2: *SSL representation segments for the utterance Over the fireplace there was a large old-fashioned gilt mirror.*

AVSE Challenge: Audio-visual Speech Enhancement Challenge

Andrea Lorena Aldana Blanco¹, Cassia Valentini-Botinhao¹, Ondrej Klejch¹, Mandar Gogate², Kia Dashtipour², Amir Hussain², Peter Bell¹

¹University of Edinburgh, Edinburgh, UK

²Edinburgh Napier University, Edinburgh, UK

lorena.aldana@ed.ac.uk

In human communication, speech plays a very important role. However, speech signals can be easily degraded by noises or other competing speech signals making it difficult to preserve the intelligibility and quality of the target message. Moreover, due to aging and genetic or pathological conditions, our human auditory system can deteriorate, leading to difficulties in understanding a speech message.

Speech enhancement models focus on improving speech intelligibility and quality in challenging conditions. For instance, by suppressing background noise or by enhancing a speech signal according to the intra-listener characteristics of the target user (i.e., considering the audiogram as part of the speech enhancement model) [1].

Most current speech enhancement models are audio-only [2], thus they exclusively consider audio signals as input to the model. Nonetheless, speech perception is a multimodal process, in which visual cues also play an important role [3]. For example, lip-reading contributes to intelligibility, particularly in challenging listening conditions. For that reason, in recent years, there has been increasing interest in developing audio-visual speech enhancement models that can outperform results from audio-only models [4].

Furthermore, there is still the need to better understand to what extend audio-only models and audio-visual speech enhancement models (might or might not) outperform intrinsic human auditory abilities of enhancing a speech signal in adverse listening conditions, for that reason it is important to provide a common framework for evaluating the proposed speech enhancement systems, and to evaluate them against human performance.

In this first edition of the AVSE Challenge: Audio-visual Speech Enhancement Challenge, we address the problem of enhancing speech signals in two proposed scenarios that are commonly challenging to audio-only speech enhancement models: (i) target speaker mixed with competing speaker and (ii) target speaker mixed with noise. In the current edition of the challenge we consider normal hearing conditions.

The Target speakers are randomly selected from the LRS3 dataset [5]. The dataset includes thousands of sentences from TED and TEDx talks. Video data is provided as cropped faces of 224x224 resolution and a frame rate of 25 fps. Competing speakers are also derived from the same dataset. Furthermore, to create the mixes involving a noise interferer, we collected noises from three different sources: Clarity Challenge (CEC1 - First edition) [6], Demand Dataset [7] and Deep Noise Suppression (DNS) Challenge (Second edition) [8]. The resulting noise database has fifteen noise categories featuring stationary and non-stationary noises.

To create the target and interferer mixes, we consider different SNR levels. In the competing speaker scenario mixes range from -15 dB to +5 dB, while in the noise interferer conditions

SNR levels are from -10 dB to +10 dB. Audio tracks are single-channel, have a sampling rate of 16 kHz and a 16 bits of bit depth.

We provide participants training and development sets that are disjoint regarding speakers and noise files. Datasets are balanced and therefore have the same amount of scenes in the competing speaker and noise scenarios. An overview of the training and development datasets is presented in table 1.

Dataset	# Scenes	# Target speakers	Interferers
Train	34,524	605	Selected from a pool of 405 competing speakers and 7,346 noise files.
Dev	3,306	85	Selected from a pool of 30 competing speakers and 1,825 noise files.

Table 1: Training and development sets

Although as part of the challenge material we provide participants with scripts to test performance of their systems using objective intelligibility metrics such as STOI and PESQ¹, we plan to conduct the evaluation (and ranking of the systems) based on listening tests taking into account the noisy mixes (without any enhancement) and the enhanced samples submitted by participants.

The evaluation dataset comprises 1389 extracted sentences from 30 speakers (15 females and 15 males). The competing speakers are selected from a pool of 6 competing speakers (3 females and 3 males). As for the noises, we selected four categories that are a subset of those from the training and development sets.

Furthermore, we provide a baseline system consisting of data loading pipeline, baseline deep neural network and scripts for model evaluation. The baseline deep neural network ingests the face crop of the target speaker and short-time-Fourier-transform of the noisy speech to output a masked spectrogram. The masked spectrogram is combined with noisy phase to synthesise the enhanced speech. Participants can decide to use the baseline model, some of its parts or none of it in their own systems.

We expect this first edition of the AVSE Challenge to contribute towards advancing the field of Audio-Visual Speech Enhancement and to provide interest insight about multimodal speech perception in challenging listening conditions.

¹AVSE Challenge scripts available at https://github.com/cogmhear/avse_challenge

1. References

- [1] Z. Tu, N. Ma, and J. Barker, “Optimising hearing aid fittings for speech in noise with a differentiable hearing loss model,” *arXiv preprint arXiv:2106.04639*, 2021.
- [2] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, “An overview of deep-learning-based audio-visual speech enhancement and separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1368–1396, 2021.
- [3] L. D. Rosenblum, “Primacy of multimodal speech perception,” in *The Handbook of Speech Perception*. John Wiley Sons, Ltd, 2005, ch. 3, pp. 51–78. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470757024.ch3>
- [4] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, “Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation,” *ACM Trans. Graph.*, vol. 37, no. 4, jul 2018. [Online]. Available: <https://doi.org/10.1145/3197517.3201357>
- [5] T. Afouras, J. S. Chung, and A. Zisserman, “Lrs3-ted: a large-scale dataset for visual speech recognition,” in *arXiv preprint arXiv:1809.00496*, 2018.
- [6] S. Graetzer, J. Barker, T. J. Cox, M. Akeroyd, J. F. Culling, G. Naylor, E. Porter, and R. V. Muñoz, “Clarity-2021 Challenges: Machine Learning Challenges for Advancing Hearing Aid Processing,” in *Proc. Interspeech 2021*, 2021, pp. 686–690.
- [7] J. Thiemann, N. Ito, and E. Vincent, “DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments,” test, Jun. 2013, supported by Inria under the Associate Team Program VERSAMUS. [Online]. Available: <https://doi.org/10.5281/zenodo.1227121>
- [8] C. K. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, “Icassp 2021 deep noise suppression challenge,” in *ICASSP*, 2021.

Leveraging linguistic knowledge for accent robustness of end-to-end models

Andrea Carmantini, Peter Bell

Centre for Speech Technology Research, University of Edinburgh, Edinburgh EH8 9AB, UK

`a.carmantini@ed.ac.uk`

Abstract

Acoustic models are susceptible to the difference in acoustic characteristics between the training distribution and test distributions. Accent variability is a challenging source of variability, and the variations within one accent often do not generalize to others. Consequently, end-to-end models that have only transcriptions as linguistic information need high amounts of data to learn how different accents realize their sounds.

To aid with recognition of accented speech, we make use of an accent independent abstraction of phonemes, often called metaphonemes, developed to have an accent invariant representation of word pronunciations. We force our models to learn hidden representations that are correlated to metaphonemes using multi-task training. Our aim is to obtain a model that is more robust to accented speech and, can, at the same time, adapt faster to different accents through the learned structure.

Our experiments on the Common Voice corpus show better generalization when making use of this additional linguistic information, with a word error rate reduction of up to 12.6% when compared to the baseline. Furthermore, the relative improvement when adapting an existing model by making use of the metaphonemes is higher than using Byte Pair Encodings alone.

Index Terms: speech recognition, acoustic model adaptation, accent adaptation, end-to-end models

A Biological Understanding of Dramatic Speech Through Synthesis

Emily Lau, Brechtje Post, Kate Knill
University of Cambridge

Dramatic vocal performance - whether in an orator's public speech or an actor's monologue - is a layered and nuanced phenomenon unique to the human experience. It is difficult to pin down the exact aspects of a performance that make it appealing to the human ear. Nonetheless, modelling and synthesising dramatic speech would vastly expand our knowledge of the nature of human expression. Previous work has mainly focused on examining acoustic parameters and their correlations to different emotions (Scherer, 1974; Mozziconacci, 2002).

Alternately, other work such as Morton (1977) and Ohala (1984) has examined emotional expression in the context of the evolution of human communicative functions, positing that there exist rules which determine the correlation between the physical structure of certain animal sounds and specific communicative needs. More recently, Xu et al (2013) have proposed that affective speech is controlled by so-called Bio-informational Dimensions (BIDs) which manipulate the speaker's vocal parameters in order to influence the behaviour of the receiver. The two most investigated BIDs have been size projection (the body size projected by the speaker) and dynamicity (the vigorousness of the speech stream). Noble & Xu (2011) showed that listeners are particularly sensitive to these two BIDs. Similarly, using synthetic manipulations of speech along these BIDs, Xu et al (2013) showed that listeners are particularly sensitive to the parameters of F0 and vocal quality in judging body size and emotion. Given these developments, BIDs are a promising theoretical framework in which to study dramatic speech.

In this work, listeners were asked to listen to pairs of utterances that were meant to simulate dramatic expressions of anger. They were all manipulated along the BIDs of size projection and dynamicity to different degrees. The listening task was divided into two sections - one containing resynthesized speech that already expressed anger, and the other neutral speech that was resynthesized to simulate anger. Listeners rated the utterances' differences in dramatic level on a five point Likert scale. This experiment will provide insights into how the degree of manipulation along the BIDs, specifically size projection and dynamicity, affect judgement of dramatic speech. This will be illuminating as to how well specific acoustic parameters correlate with the BIDs in question, and their dependence on the properties of the source stimuli.

References:

- Scherer, K. R. (1974). Voice quality analysis of American and German speakers. *Journal of Psycholinguistic Research*, 3(3), 281-298.
- Mozziconacci, S. (2002). Prosody and emotions. In *Speech Prosody 2002, International Conference*.
- Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F₀ of voice. *Phonetica*, 41(1), 1-16.
- Morton, E. S. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *The American Naturalist*, 111(981), 855-869.
- Xu, Y., Kelly, A., & Smillie, C. (2013). Emotional expressions as communicative signals. *Prosody and iconicity*, 33-60.
- Noble, L., & Xu, Y. (2011). Friendly Speech and Happy Speech-Are They the Same?. In *ICPhS* (pp. 1502-1505).

Modelling Pronunciation Variation in Different Spoken Englishes

Emma O'Neill¹, Julie Carson-Berndsen¹

¹ ADAPT Research Centre, School of Computer Science, University College Dublin, Ireland

emma.l.oneill@ucdconnect.ie, julie.berndsen@ucd.ie

It is often the case that Automatic Speech Recognition (ASR) systems exhibit better performance on some spoken varieties than on others. Speakers with particular regional dialects, non-native speakers, or those whose variety deviates from what is considered 'mainstream' or 'standard' are typically the ones who suffer from reduced performance of voice technology. One source of this performance degradation lies in pronunciation variation. Specific phonetic realisations that were underrepresented (or absent) in the training data used to develop an ASR system can cause issues during recognition, often resulting in the misrecognition of the intended phoneme and thus higher word error rates.

In order to improve these ASR systems it is important to investigate patterns of pronunciation variation that are common to speakers of a particular variety and, specifically, which phonetic realisations lead to errors in the ASR output and which ones the system is robust to. We present a method of leveraging erroneous ASR output to build simple and explainable models of a speaker's variety of English and demonstrate that the resulting representations do capture commonalities specific to a region.

The self-supervised wav2vec2.0 transformer model, fine-tuned on 100 hours of the Librispeech corpus [1] was used to recognise speech from the Datatang corpus [2]: approximately 200 hours of read speech and the corresponding text prompts from 528 speakers across 10 different regions. The phoneme sequences of a 'canonical' pronunciation of the prompts were then compared with the predicted phoneme sequence of the ASR output. Any edit operations (insertions, deletions, or substitutions) were then used to model the speaker's variety as a 40x40 phoneme confusion matrix which could be reshaped to form a 1600-dimensional vector representation.

If these simple models of a speaker's variety are capable of capturing common features of a region's spoken English, we would expect to see those from the same region clustering together in the multi-dimensional space. We performed K-means clustering ($k=10$) on the speaker vectors and visualised the results in 2 dimensions as seen in Figure 1. Whilst there are naturally some outliers, we do observe the expected clustering effects for most of the regions. Notably, many of the American speakers overlap with the Canadian speakers whilst others appear closer to the Indian or Spanish speakers. We hypothesise that this is due to the high degree of variation that exists within American Englishes and the influence of other languages spoken in the region.

We then looked at the classification potential of these speaker representations. If a speaker's region can be accurately classified using just these phoneme confusions then this can be considered further evidence that the models have captured regional pronunciation features. We first used a KNN approach to classify test speakers based on the regions of the training speakers closest in the multi-dimensional space. Whilst this produced impressive results for most regions, the outlier speakers had a large impact on the accuracy for American and Canadian speakers. To improve on this we then applied a Regional Profile approach where the training speakers from each region were pooled together to create a prototypical representation of the variety. This raised the overall classification accuracy from 70.4% to 81.8%.

Finally, we carried out a qualitative analysis on the resulting region profiles. By reverting the vector representations back to a 40x40 matrix we were able to investigate which phoneme confusions are modelled as characteristic of each region. These observations were shown to align with existing literature on the Englishes of each region as well as with research on first language transference on L2 productions by non-native English speakers. Moreover, this analysis allows us to see exactly which pronunciation features affected the ASR output and which features that we might have expected to encounter but to which the system appears to be robust.

1. Acknowledgements

The ADAPT Centre for Digital Content Technology (www.adaptcentre.ie) is funded under the SFI Research Centres Programme (Grant 13/RC/2106_P2)

2. References

- [1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.
- [2] X. Shi, F. Yu, Y. Lu, Y. Liang, Q. Feng, D. Wang, Y. Qian, and L. Xie, "The accented English speech recognition challenge 2020: open datasets, tracks, baselines, results and methods," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6918–6922.

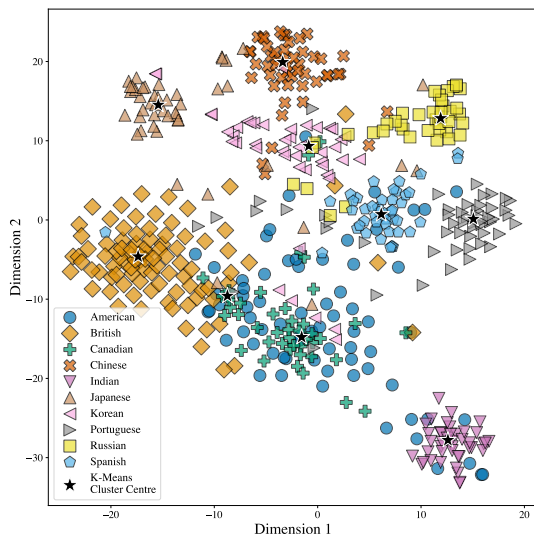


Figure 1: 2-dimensional visualisation of the speaker vectors and cluster centers.

Using Utterance-Specific Dirichlet Priors to Model Uncertainty in Emotion Class Labels

Wen Wu¹, Chao Zhang¹, Xixin Wu², Philip C. Woodland¹

¹Department of Engineering, University of Cambridge

²Stanley Ho Big Data Decision Analytics Research Centre, The Chinese University of Hong Kong

{ww368, cz277, pcw}@eng.cam.ac.uk, wuxx@se.cuhk.edu.hk

Automatic emotion recognition (AER) is a key attribute for artificial intelligence systems that need to naturally interact with humans. However, the task definition is still an open problem due to the inherent ambiguity of emotions. A straightforward definition of AER is to classify each utterance into a set of pre-defined emotion classes (*i.e.* happy, sad, angry). However, since emotion is inherently complex and highly personal, different labels can be assigned to the same utterance by different annotators, which leads to uncertainty in emotion labelling. Majority voting is usually used to remove this uncertainty. This strategy not only causes training utterances without majority agreed labels to be unused, but also effectively assumes that utterances with mixtures of emotions would either not be encountered or not be evaluated at test-time, which is not the case in human interaction [1].

Instead of removing the uncertainty in labels with majority voting, we propose modelling such uncertainty with a novel Bayesian training loss based on utterance-specific Dirichlet prior distributions. In this approach, emotion states are treated as categorical distributions. The emotion class labels provided by different annotators of each utterance are considered as one-hot categorical distributions sampled from an utterance-specific Dirichlet prior distribution. A separate prior distribution is estimated for each utterance by an improved Dirichlet prior network (DPN) [2], which predicts the concentration parameter α of the Dirichlet prior. As shown in Figure 1, given α , the categorical distribution μ is a sample drawn from $\text{Dir}(\mu|\alpha)$, and hard label of emotion class ω_k is a sample drawn from μ . The DPN is trained by minimising the negative log likelihood

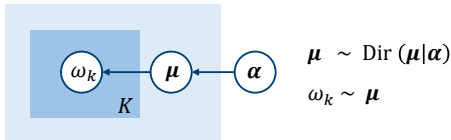


Figure 1: Illustration of the Dirichlet process. K is the number of emotion classes.

of sampling the original one-hot categorical distributions from their relevant utterance-specific Dirichlet priors.

Alternatively, from a frequentist perspective, “soft” labels can be obtained by averaging the one-hot categorical distributions relevant to the emotion class labels, which can be viewed as maximum likelihood estimate (MLE) of the label for each

Wen Wu is supported by a Cambridge International Scholarship from the Cambridge Trust. This work has been performed using resources provided by the Cambridge Tier-2 system operated by the University of Cambridge Research Computing Service (www.hpc.cam.ac.uk) funded by EPSRC Tier-2 capital grant EP/T022159/1.

utterance. An AER model can be trained by minimising the Kullback-Leibler (KL) divergence between the soft labels and its predicted emotion distributions. The DPN and KL training losses can be combined by linear interpolation.

Classification accuracy is no longer an appropriate evaluation metric when considering uncertainty in emotion labelling. Instead, we propose evaluating the model performance in uncertainty estimation in terms of the area under the precision-recall curve (AUPR) when detecting utterances without majority agreed labels at test-time. Two threshold measures were used: the probability of the predicted class or max probability (Max.P) and the entropy of the predictive distribution (Ent.). Four systems were tested: (i) a simple emotion classification system “hard”; (ii) a soft label system “soft” trained by minimising KL divergence; (iii) a DPN system “dpn”; (iv) a system “dpn-kl” trained by combining DPN and KL losses. Experimental results on the widely used IEMOCAP dataset [3] show that the combined loss not only has more stable training performance but also results in improved uncertainty estimates.

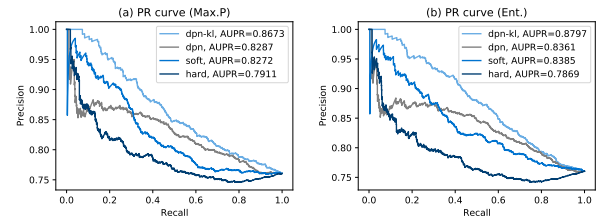


Figure 2: Precision-Recall (PR) curves for the four systems using (a) MaxP and (b) Ent. as the uncertainty measures.

In conclusion, the proposed system preserves uncertainty in emotion labelling by maximising the likelihood of sampling all one-hot labels with inconsistent emotion classes from an utterance-specific Dirichlet distribution. Beyond emotion recognition, label uncertainty is a common issue in many human perception and understanding tasks, since golden reference is usually not well-defined due to the subjective evaluation of annotators. The proposed method could be applicable to other such tasks to handle the uncertainty in labels.

1. References

- [1] L. Devillers, L. Vidrascu, and L. Lamel, “Challenges in real-life emotion annotation and machine learning based detection,” *Neural Networks*, vol. 18, no. 4, pp. 407–422, 2005.
- [2] A. Malinin and M. Gales, “Predictive uncertainty estimation via prior networks,” in *Proc. NeurIPS*, Montréal, 2018.
- [3] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Provost, S. Kim, J. Chang, S. Lee, and S. Narayanan, “IEMOCAP: Interactive emotional dyadic motion capture database,” *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.

PSE-Net: Real-time Personalized Sound Enhancement

Abhinav Mehrotra^{*1}, Alberto Gil C. P. Ramos^{*1}, Nicholas D. Lane¹², Sourav Bhattacharya¹
¹Samsung AI Center, Cambridge, UK ²University of Cambridge, UK

In spite of the progress made in sound extraction methodologies, there exists a considerable gap in the performance of real-time streaming sound enhancement solutions, compared to their non-causal alternatives. To bridge this gap, in this work we propose Personalized Sound Enhancement Network (PSE-Net) that uses causal Time Depth Separable (TDS) convolution blocks with unidirectional RNNs to form an encoder-decoder architecture. The proposed model (a) operates in streaming mode, (b) shows superior Signal-to-Distortion Ratio (SDR), (c) achieves lower word-error-rate (WER) when applied to a downstream ASR task, and (d) demonstrates transferability across different unseen languages. An overview of PSE-Net is illustrated in Figure 1, which follows an encoder-decoder style network with intermediate filter blocks. The final output mask is generated using a fully-connected layer with sigmoid activations.

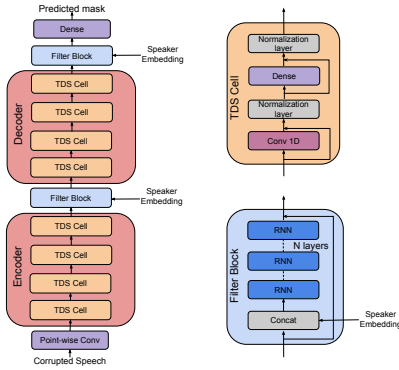


Figure 1: Model architecture for PSE-Net.

Evaluation Settings

Integration with ASR. We use publicly available pre-trained enterprise-grade ASR models, which are available for English, German and Spanish languages [2].

Dataset. We construct the training dataset from LibriSpeech [3] 100h and 360h training splits, which contain clean speech. To evaluate our system with ASR integration, we use test-clean split of LibriSpeech for English ASR model. For Spanish and German ASR models, we use test-split of VoxForge dataset [1].

Baselines. We compare our proposed causal frame-in frame-out PSE-Net model against the state-of-the-art (SOTA) causal frame-in frame-out VoiceFilter-Lite model [4]. In addition, we present results of SOTA in non-causal model, namely the VoiceFilter (VF) model [5] to compare our real-time results with a strong non-realtime baseline.

Results

As shown in Table 1 (under English column), all three top performing PSE-Net models outperform SOTA VoiceFilter-Lite model [4] for enhancing speech corrupted with both noise types. Specifically, the best PSE-

Table 1: SDRi results for PSE-Net and baseline models.

Model	English		Spanish	
	Babble	Ambient	Babble	Ambient
PSE-Net	9.355	12.604	5.930	8.645
VFLite	6.543	12.132	-0.084	8.634
VF	10.164	13.275	6.760	10.329

Net model (highlighted with bold) achieves 9.36 and 12.60 SDRi for suppressing babble and ambient noises respectively. Additionally, we show that the performance of PSE-Net is close to VoiceFilter model [5]. Moreover, we observe that the SDRi is overall better (i.e., higher) for denoising ambient compared to babble noise. Note that in Table 1 we present only top three PSE-Net model configurations due to limited space. Overall, by searching for different parameters, we observe that increasing model complexity does not necessarily lead to better model performance.

Impact on ASR Quality. To quantify the effectiveness of PSE-Net on downstream ASR tasks, we measure the improvement in WER when passing the enhanced speech, as opposed to the corrupted speech, through an ASR model. We use the open source Silero ASR models for three languages. The results show that all PSE-Net variants can effectively suppress both babble and ambient noise, and outperform SOTA VoiceFilter-Lite model [4] with a significant margin.

Language Shift. We investigate how SDRi and WER performance of speech enhancement models trained on English transfers when applied to other languages, namely Spanish and German. The results show that the PSE-Net trained on English performs well when used for enhancing Spanish and German speech. This is encouraging from a practical point of view, since it highlights the potential of applying speech enhancement in a language-agnostic manner.

References

- [1] VoxForge Corpus. <http://www.voxforge.org>, 2006.
- [2] Silero models. <https://github.com/snakers4/silero-models>, 2021.
- [3] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an ASR corpus based on public domain audio books. In *ICASSP*, 2015.
- [4] Quan Wang, Ignacio Lopez Moreno, Mert Saglam, Kevin Wilson, Alan Chiao, Renjie Liu, Yanzhang He, Wei Li, Jason Pelecanos, Marily Nika, and Alexander Gruenstein. VoiceFilter-Lite: Streaming Targeted Voice Separation for On-Device Speech Recognition. In *Interspeech*, 2020.
- [5] Quan Wang, Hannah Muckenhirn, Kevin Wilson, Prashant Sridhar, Zelin Wu, John R. Hershey, Rif A. Saurous, Ron J. Weiss, Ye Jia, and Ignacio Lopez Moreno. VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking. In *Interspeech*, 2019.

Conversational Speech vs. Sustained Phonation for Diagnosis of Parkinson's Disease

Steve W. Beet, Phill Restall, Ladan Baghai-Ravary

AI R&D Team, Aculab plc, Milton Keynes, UK
{steve.beet, phill.restall, ladan.ravary} @ aculab.com

Abstract

In this paper we describe a method to aid in the diagnosis and monitoring of conditions such as Parkinson's Disease (PD), using only informal recordings of spontaneous speech, such as are commonly found on social media.

During system development, cross-validation was used to define gender and age-dependent baselines for those with, and those without, a clinical diagnosis of Parkinson's Disease. The system was then applied to retrospective longitudinal studies of several public figures, and patterns in the results were observed, suggesting how these kinds of measurements could be used both for screening or diagnosis, and for subsequent monitoring of disease progression.

The subjects ultimately diagnosed with Parkinson's Disease showed PD detection rates consistently above the baseline for PD diagnosed subjects. Often this was observable even while they were in their late twenties/early thirties (prior to diagnosis). The detection rate typically increased steadily over the following decade or more. By the time their health prevented them working, their detection rates were usually in the region of 95%.

It was also found that signs of Parkinsonism are often observed several years before any conventional diagnosis, showing the importance of performing longitudinal studies, especially in cases where individual anomalous results could be misinterpreted without the context of historical trends. The availability of retrospective recordings, such as may be provided by social media or recorded telephone conversations, is most valuable in cases such as these.

This work has identified some key factors that affect a subject's speech, as assessed by an automatic PD detection system. In order, the most significant factors were: (1) the presence of Parkinson's Disease, (2) external factors including general health, (3) the subject's age, and (4) the subject's gender.

The PD detection system was based on pre-existing speech technology: X-Vectors and MFCCs [1, 2]. As such it is primarily sensitive to features derived from the overall shape and short-term variation of the speech signal's power spectrum. This contrasts with previous work in this area [3-9], that concentrated on the effects of vocal tremor on the production of voiced speech.

A comparison between spontaneous speech and prepared (read) speeches has shown that this PD detection system is much more sensitive when spontaneous (conversational) speech is analysed. We hypothesise that this might be due to the increased cognitive load required to choose vocabulary

and create grammatical constructs, and the inability to rely on simple 'muscle memory'.

We have demonstrated that the PD detection rate from spontaneous speech can indeed be highly informative, primarily because it can give access to a history of voice recordings, with calibration via age and gender-dependent baselines having been performed during the detection system's development. This allows underlying trends to be identified despite (for example) voice changes due to transient medical conditions.

Index Terms: Parkinson's Disease, speech analysis, X-Vectors, gender-dependency, age-dependency, e-Health

1. References

- [1] Snyder D, Garcia-Romero D, Sell G, Povey D, Khudanpur S (2018) "X-vectors: Robust DNN Embeddings for Speaker Recognition" in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2018
- [2] SRE16 Xvector Model, downloaded from <https://kaldi-asr.org/models/m3>
- [3] Henríquez P, Alonso JB, Ferrer MA, Travieso CM, Godino-Llorente JJ, Díaz-de-María F (2009) "Characterization of healthy and pathological voice through measures based on nonlinear dynamics." in IEEE Trans Audio, Speech, and Lang Process 17(6):1186-1195, 2009.
- [4] Majdinasab F, Karkheiran S, Soltani M, Moradi N, Shahidi G (2016). Relationship Between Voice and Motor Disabilities of Parkinson's Disease. Journal of Voice: official journal of the Voice Foundation, 30(6), 768.e17-768.e22. <https://doi.org/10.1016/j.jvoice.2015.10.022>.
- [5] Baghai-Ravary L, Beet SW (2017) "VoiScan: Telephone Voice Analysis for Health and Biometric Applications" in SPECOM 2017, LNAI 10458, pp. 799-808, doi: 10.1007/978-3-319-66429-3_80, 2017.
- [6] Orozco-Arroyave JR, et al. (2015) "Characterization Methods for the Detection of Multiple Voice Disorders: Neurological, Functional, and Laryngeal Diseases", IEEE Journal of Biomedical and Health Informatics
- [7] Arora S, Tsanas A (2021) "Assessing Parkinson's Disease at Scale Using Telephone-Recorded Speech: Insights from the Parkinson's Voice Initiative", Diagnostics, 2021; 11(10):1892. <https://doi.org/10.3390/diagnostics11101892>
- [8] Alonso JB, de León J, Alonso I, Ferrer MA (2001) "Automatic detection of pathologies in the voice by HOS based parameters." in EURASIP J Appl Signal Process 2001:275-284.
- [9] Little M (2012) "A test for Parkinson's with a phone call" presented at TED Global 2012, video at https://www.ted.com/talks/max_little_a_test_for_parkinson_s_w_ith_a_phone_call

Tree-Constrained Pointer Generator for End-to-end Contextual ASR

Guangzhi Sun, Chao Zhang, Philip C. Woodland

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

{gs534, cz277, pcw}@eng.cam.ac.uk

Contextual knowledge is important for real-world ASR applications. Contextual knowledge is often represented by a list (*biasing list*) of words or phrases (*biasing words*) that are likely to appear in an utterance in a given context. Examples of resources to find biasing lists include a user’s contact book, sides in a presentation, and the ontology of a dialogue system *etc.* These words are mostly rare content words that are important to downstream tasks and are thus highly valuable.

A tree-constrained pointer generator (TCPGen) [1] component is proposed for contextual knowledge integration in a neural-symbolic way. TCPGen structures the biasing words into an efficient prefix tree to serve as its symbolic input and creates a neural shortcut between the tree and the final ASR output distribution to facilitate recognising biasing words. Specifically, a

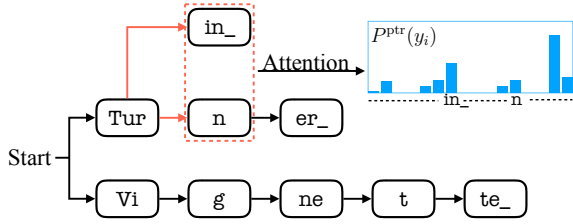


Figure 1: Example prefix tree. With previous output *Tur*, *in_* and *n* are two valid word pieces on which attention will be performed. $P^{ptr}(y_i)$ is the TCPGen distribution.

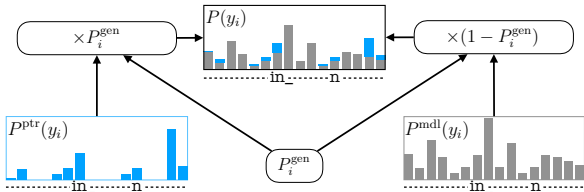


Figure 2: Illustration of interpolation in TCPGen. $P^{ptr}(y_i)$ is the TCPGen distribution. $P^{mdl}(y_i)$ is the distribution from a standard end-to-end model. $P(y_i)$ is the final output distribution. P_i^{gen} is the generation probability.

set of valid wordpieces is obtained by searching the prefix-tree with a given history output sequence (Fig. 1). Then, a scaled dot-product attention is performed to find the TCPGen distribution, $P^{ptr}(y_i)$, where the query combines the history and acoustic information and the keys are the encodings of nodes on the tree. TCPGen also estimates a generation probability, P_i^{gen} , to indicate how much contextual knowledge is needed at the current step based on ASR output hidden states and biasing list

Guangzhi Sun is funded by Cambridge Trust.

This work has been performed using resources provided by the Cambridge Tier-2 system operated by the University of Cambridge Research Computing Service (www.hpc.cam.ac.uk) funded by EPSRC Tier-2 capital grant EP/T022159/1

Table 1: WER and R-WER on Librispeech test-clean and test-other sets using RNN-T. Baseline refers to the standard RNN-T.

System	Test-clean (%)		Test-other (%)	
	WER	R-WER	WER	R-WER
Baseline	4.0	14.1	10.1	33.1
Deep Biasing [2]	3.6	10.4	9.5	25.0
TCPGen	3.4	8.9	8.8	22.2
+ GNN enc.	3.1	7.6	8.2	18.8

contents. The final output distribution is the interpolation between the TCPGen distribution and the original AED or RNN-T output distribution, weighted by the generation probability.

Furthermore, the prefix-tree can be also encoded using a graph neural network (GNN). With GNN encodings, lookahead for future wordpieces in end-to-end ASR decoding is achieved at each tree node by incorporating information about all wordpieces on the tree branches rooted from it, which allows a more accurate prediction of the generation probability of the biasing words. Specifically, the Tree-RNN model is studied here.

Experiments were performed on Librispeech dataset using RNN-T. Biasing lists were arranged following [2]: the full rare word list containing 200k distinct words was obtained by removing the most common 5k words from the Librispeech LM vocabulary. Rare words were defined as words belonging to this list. Biasing lists were then organised by finding words that belong to the full rare word list from the reference of each utterance and adding 1000 distractors. The rare word error rate (R-WER) was used to measure the error rate of rare words. Key results are summarised in Table 1.

In conclusion, TCPGen achieved 35% relative R-WER reduction compared to the baseline, which was further increased to 45% using GNN encodings on LibriSpeech test sets. There are experiments performed with more realistic setups. In [3], contextual knowledge was extracted from slides of presentations in a meeting corpus. In [4], the ontology of a dialogue system was used as contextual knowledge for the dialogue state tracking challenge (DSTC) corpora.

1. References

- [1] G. Sun, C. Zhang & P. C. Woodland “Tree-constrained pointer generator for end-to-end contextual speech recognition”, *Proc. ASRU*, Cartagena, 2021.
- [2] D. Le, M. Jain, G. Keren, S. Kim, Y. Shi, J. Mahadeokar, J. Chan, Y. Shanguan, C. Fuegen, O. Kalinli, Y. Saraf & M. L. Seltzer “Contextualized streaming end-to-end speech recognition with trie-based deep biasing and shallow fusion”, *Proc Interspeech*, 2021.
- [3] G. Sun, C. Zhang & P. C. Woodland “Tree-constrained pointer generator with graph neural network encodings for contextual speech recognition”, *Proc. Interspeech*, Incheon, 2022.
- [4] G. Sun, C. Zhang & P. C. Woodland “Minimising biasing word errors for contextual ASR with the tree-constrained pointer generator”, arXiv: 2205.09058, 2022.

Canonical-Correlated Graph Neural Network for Multimodal Energy-Efficient Speech Enhancement

Leandro Aparecido Passos*, Ahmed Khubaib*, Mohsin Raza*, Amir Hussain†, Ahsan Adeel*‡

*CMI Lab, School of Engineering and Informatics, University of Wolverhampton, England, United Kingdom.

†School of Computing, Edinburgh Napier University, Edinburgh, Scotland, United Kingdom.

‡deepCI.org, Edinburgh, Scotland, United Kingdom

Email: ahsan.adeel@deepci.org.

This abstract presents a multimodal self-supervised architecture for energy-efficient audio-visual (AV) speech enhancement that integrates Graph Neural Networks with canonical correlation analysis (CCA-GNN) [1], [2]. The proposed approach lays its foundations on a state-of-the-art CCA-GNN that learns representative embeddings by maximizing the correlation between pairs of augmented views of the same input while decorrelating disconnected features. The key idea of the conventional CCA-GNN [3] involves discarding augmentation-variant information and preserving augmentation-invariant information while preventing capturing of redundant information. Our AV CCA-GNN model deals with multimodal representation learning context. Our model improves contextual AV speech processing by maximizing canonical correlation from augmented views of the same channel and canonical correlation from audio and visual embeddings. In addition, it proposes a positional node encoding that considers a prior-frame sequence distance instead of a feature-space representation when computing the node's nearest neighbors, introducing temporal information in the embeddings through the neighborhood's connectivity, as depicted in Figure 1. Further, Figure 2 depicts the whole pipeline.

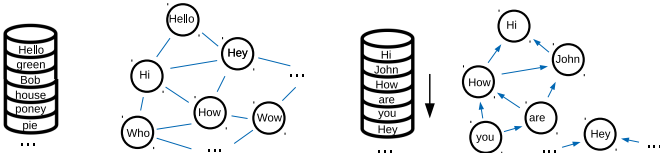


Fig. 1: Node neighborhood modeling considering the standard feature-space (left) and the proposed k prior frames (right).

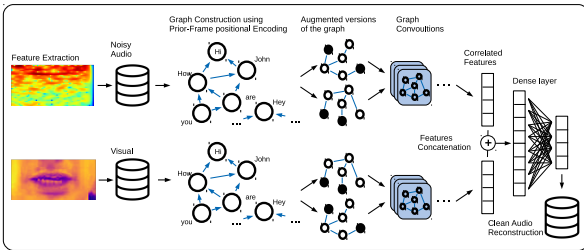


Fig. 2: Proposed pipeline.

Experiments conducted on the benchmark ChiME3 dataset show that the prior frame-based AV CCA-GNN ensures a better feature learning in the temporal context, leading to more energy-efficient speech reconstruction. Figure 3 depicts an example of signal reconstruction, as well as the neuron activation rate, i.e., the energy efficiency.

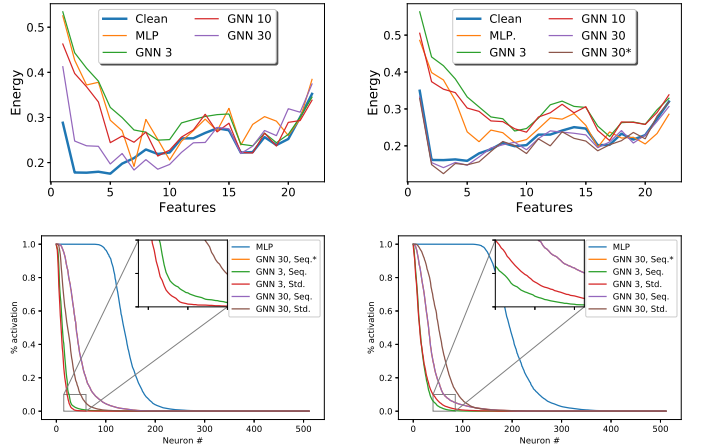


Fig. 3: Top row: Signal reconstruction for the standard (left) and the prior frame (right). Bottom row: Neuron activation rate considering audio (left) and visual (right) channels.

ACKNOWLEDGMENTS

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/T021063/1].

REFERENCES

- [1] L. A. Passos, J. a. P. Papa, J. Del Ser, A. Hussain, and A. Adeel, "Multimodal audio-visual information fusion using canonical-correlated graph neural network for energy-efficient speech enhancement," 2022. [Online]. Available: <https://arxiv.org/abs/2202.04528>
- [2] L. A. Passos, J. P. Papa, and A. Adeel, "Canonical cortical graph neural networks and its application for speech enhancement in future audio-visual hearing aids," 2022. [Online]. Available: <https://arxiv.org/abs/2206.02671>
- [3] H. Zhang, Q. Wu, J. Yan, D. Wipf, and S. Y. Philip, "From canonical correlation analysis to self-supervised graph neural networks," in *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

CognoSpeak: a Cognitive Health Assessment Tool (CcHAT)

Nathan Pevy¹ Dr Bahman Mirhedari¹ Dr Ronan O'Malley² Dr Simon Bell² Lise Sproson³ Swapnil Gadgil⁴
 Rebecca Bright⁴ Ismail Yussuf⁵ Sahra Abdi⁵ Muse Jam⁵ Dr Traci Walker⁶ Professor Markus Reuber²
 Dr Dan Blackburn² Professor Heidi Christensen¹

¹ Department of Computer Science - The University of Sheffield

² Department of Neuroscience - The University of Sheffield

³ NIHR Devices for Dignity Healthcare Technology Cooperative

⁴ Therapy Box

⁵ ISRAAC Somali Community Association

⁶ Division of Human Communication Sciences - The University of Sheffield

CognoSpeak™ is an application for detecting early signs of dementia. The application allows individuals who are reporting early signs of cognitive decline to partake in an interaction with a virtual agent, either independently or with an accompanying other. The application conducts an assessment to differentiate between people with early signs of dementia, mild cognitive impairment, functional memory impairment, and healthy controls. The assessment uses automatic speech recognition-based speech processing and machine learning to predict the most likely outcome based upon interactional, lexical, semantic, and acoustic group differences. The CognoSpeak™ system has been tested on over 220 people and shown to be a reliable and accurate method to assess people with memory concerns. The objective of the application is to provide clinical assessments to medical professionals to guide triaging and the decision making process.

The project has recently secured over one million pounds of funding from the National Institute of Health Research (NIHR) to further test and develop the application. This project (CcHAT: A Cognitive Health Assessment Tool) aims to demonstrate the feasibility of stratifying cognitive decline in primary care using CognoSpeak™. It will work with end users to co-design a new version of the application that will be tested in GP clinics and peoples' own homes to ensure that the application is accessible and accurate for everyone. Key to the success of the project will be demonstrating the feasibility of CognoSpeak™ with underrepresented groups, developing AI-algorithms that are agnostic to diverse language backgrounds, and the delivery of a trial platform to demonstrate health economic effectiveness of CognoSpeak™. The project will generate a final product that could be utilised in the NHS with the necessary regulatory approvals.

The poster that we present at the UK Speech conference will provide an introduction and overview of CognoSpeak™. We will showcase the initial design that has been co-designed and developed in collaboration with end-users and clinicians and provide an overview of the previous research conducted by the research group. Finally, we will outline our future research objectives and plans. We hope to engage and make connections with the speech technology community that are conducting research in related fields.

Attention Forcing for Speech Synthesis

Qingyun Dou & Mark J. F. Gales

University of Cambridge

{qd212,mjfg100}@cam.ac.uk

Abstract

Auto-regressive sequence-to-sequence models with attention mechanisms have achieved state-of-the-art performance in various tasks including speech synthesis. Training these models can be difficult. The standard approach guides a model with the reference output history during training. However during synthesis the generated output history must be used. This mismatch can impact performance. Several approaches have been proposed to handle this, normally by selectively using the generated output history. To make training stable, these approaches often require a heuristic schedule or an auxiliary classifier. This paper introduces attention forcing, which guides the model with the generated output history and reference attention. This approach reduces the training-evaluation mismatch without the need for a schedule or a classifier. Additionally, for standard training approaches, the frame rate is often reduced to prevent models from copying the output history. As attention forcing does not feed the reference output history to the model, it allows using a higher frame rate, which improves the speech quality. Finally, attention forcing allows the model to generate output sequences aligned with the references, which is important for some down-stream tasks such as training neural vocoders. Experiments show that attention forcing allows doubling the frame rate, and yields significant gain in speech quality.

Index Terms: sequence-to-sequence model, attention mechanism, training, speech synthesis, exposure bias

1. Attention-based sequence-to-sequence generation

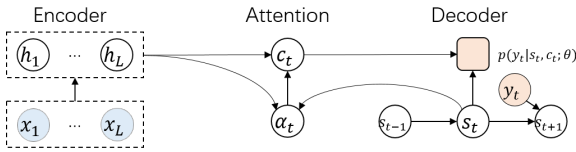


Figure 1: Illustration of an attention-based encoder-decoder

2. Attention forcing

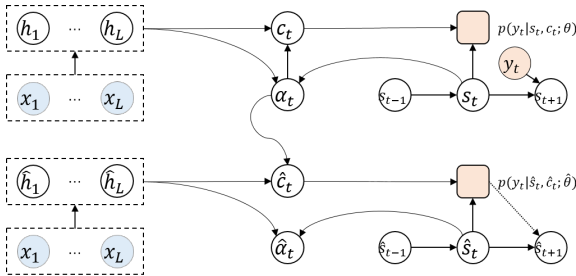


Figure 2: Illustration of attention forcing

3. Experiments

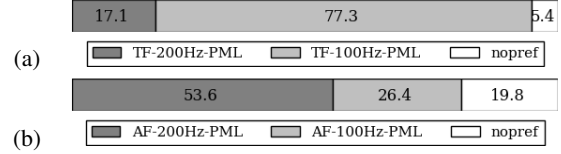


Figure 3: Listening tests comparing 100Hz and 200Hz models; frame-level models trained with (a) teacher forcing, (b) attention forcing; waveform generated using a PML vocoder

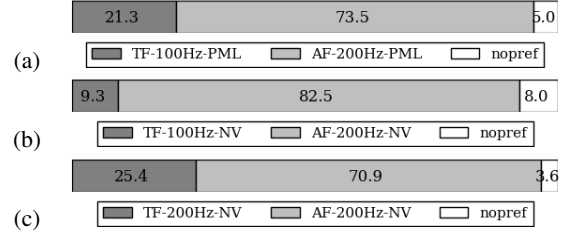


Figure 4: Listening tests comparing teacher forcing and attention forcing; waveform generated using (a) PML vocoder (b,c) neural vocoder

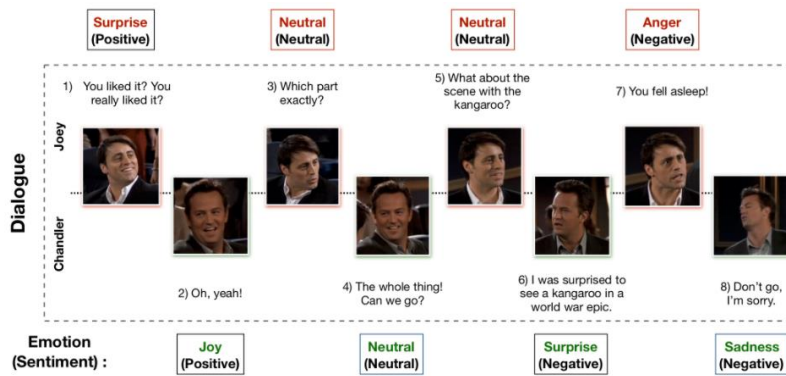
Table 1: Global variance of vocoder features generated by different models, computed over the test set, averaged over all sequences and dimensions

Training	Global variance	
	200Hz	100Hz
Teacher forcing	0.39	0.54
Attention forcing	0.71	0.70

Multimodal Emotion Recognition in Conversations

Abstract

Humans perceive the world by concurrently processing and fusing a number of different and complementary cues such as audio and language. Various emotional states in conversations between multi-speaker reflect multiple cues across modalities. The availability of an enormous quantity of multimodal data and its widespread applications, automatic sentiment analysis and emotion detections in the dialogue has become a hot research topic among the research community. For multi-modal emotion recognition tasks, designing a unified model for modality fusion is challenging due to main factors: 1) variations in learning dynamics between modalities; 2) long-range dependencies between elements across modalities; 3) inherent data non-alignment. Most of the existing fusion techniques rely on recurrent networks or conventional attention mechanisms that do not effectively leverage the complementary nature of audio-language modalities. In the future work, we focus on cross-attention fusion techniques over deep learning models for emotion recognition in conversations from spoken audio and corresponding transcriptions. In particular, it utilizes cross-attention weights to capture the more contributive features across individual modalities, and combine self-attention weights to focus on intra modalities. On the other hand, the interlocutor state, context state between the neighboring utterances and multimodal fusion play an important role in multimodal emotion recognition. Multiple recurrent neural network blocks are used to capture the interlocutor state and contextual state between the utterances. The experiments are conducted on three standard datasets such as MELD, IEMOCAP, and CMU-MOSI.



[Poria et al. 2019]

Addressing user concerns about multi-modal hearing technology

Dorothy A. Hardy¹, Michael A Akeroyd², Adeel Hussain¹, Peter Bell²,
Amir Hussain¹

¹Edinburgh Napier University, UK ²University of Nottingham, UK

³The University of Edinburgh, UK

D.Hardy@napier.ac.uk, A.Hussain@napier.ac.uk

Despite decades of development, many hearing aids are not as good as their users would wish at focusing on one speaker's voice from surrounding noise rather than amplifying all nearby sound. Cognitively-inspired, 5G and Internet-of-Things enabled, multi-modal hearing aids are one potential solution. This paper addresses key concerns around these revolutionary developments in hearing aid design, revealed during a recent workshop.

Index Terms: hearing aid perceptions, audio-visual speech enhancement, speech-in-noise, radio-frequency sensing

1. Audio-visual hearing technology

Listeners naturally follow movements of a speaker's face and mouth, and a sophisticated, multi-stage process uses this information to separate speech from noise and fill in any gaps. One aim of the COG-MHEAR research programme [1], [2] is to create multi-modal hearing aids that act in much the same way by using low-resolution cameras to video the lip region of the target speaker, along with novel algorithms to combine audio and visual information, in order to improve speech quality and intelligibility in real-world noisy environments. However, such technology comes with potential problems in privacy and safety. Our workshop explored these in depth.

2. Obstacles and solutions

Six specialists in hearing technology, most of whom were hearing aid users, were consulted about the new hearing aid design at a workshop in July 2022. They included clinicians and audiologists, plus an industry representative. The discussion raised the following queries and potential remedies:

Q1. Would the sensors be able to select the speech of someone beside or behind the hearing aid user?

The environmental context of a conversation could be assessed through use of omnidirectional microphones and visual sensors.

Q2. How can privacy be preserved whilst using cameras?

Low resolution cameras can be designed to isolate only the lip region of a speaker [3] rather than videoing the entire face. Radio frequency (RF) sensors could be used as alternatives to cameras [4].

Q3. Would the technology work when speakers are wearing masks?

Perhaps surprisingly, yes: our initial work with optical and radio frequency (RF) sensors shows that this is indeed possible.

Q4. Will there be a need for wearers to explain the technology repeatedly in order to allay privacy concerns?

Perhaps initially, but as public awareness (and the discreteness of devices) increases this will reduce. We expect that many areas of technology development will spark a public debate

about cameras and sensors in many applications [5], including in the development of hearing technology.

Q5. Will the data requirements of an audio-visual device be excessive?

No, we doubt it; processing on prototype devices can be carried out in the cloud to minimise latency. Longer term development of software and hardware will lessen data requirements [6].

Q6. Are the new devices safe, especially if internet and radar connections are required beside the ear?

This is critical to achieve for public acceptance of any new product. Initial estimates suggest that safety can be achieved, but it will be critical to adhere to (and ideally do far better than) all radio-frequency exposure regulations [7], [8].

Q7. What will the prototype devices look like?

Most likely any initial prototypes will look similar to "smart glasses", but inventive new designs could improve public acceptance.

3. Conclusions

The COG-MHEAR research programme is addressing the need for multi-modal hearing aids that mimic normal hearing by combining visual and auditory cues in order to improve speech quality and intelligibility in real-world noisy environments. Ongoing consultations with experts in hearing technology have highlighted queries, in particular about privacy and safety when using the proposed devices. These can be addressed through use of privacy-preserving and internet-independent software and hardware along with radio frequency technology.

4. Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/T021063/1].

The research teams are grateful to the participants in the COG-MHEAR workshop on 20th July 2022 for giving their insights into the developing hearing technology.

5. References

- [1] https://cogmhear.org/assets/2021_22COG_MHEAR_ShortenedAnnualReport.pdf
- [2] <https://cogmhear.org/>
- [3] <https://demo.cogmhear.org/>
- [4] <http://eprints.gla.ac.uk/276607/>
- [5] <https://www.bbc.co.uk/news/magazine-16681630>
- [6] doi: 10.48550/arxiv.2206.02671.
- [7] <https://www.comsoc.org/publications/ctn/truth-out-there-examining-science-around-5g-paranoia>
- [8] <https://www.gov.uk/government/publications/radio-waves-reducing-exposure/radio-waves-reducing-exposure-from-mobile-phones>

Model for Assessor Bias in Automatic Pronunciation Assessment

Jose Antonio Lopez, Thomas Hain

SPEECH AND HEARING RESEARCH DEPARTMENT OF COMPUTER SCIENCE,
UNIVERSITY OF SHEFFIELD, UK

In pronunciation assessment, the assessor's perception is influenced by a particular pronunciation template. This assessor may hold a bias towards certain variations in pronunciation which do not necessarily impact communication, yet they may be penalized during the assessment. A model for the assessor bias will benefit pronunciation assessment for the sake of a fair evaluation of a speaker. This work proposes a model for pronunciation assessment as the combination of an assessor independent (**A**) and an assessor specific (**B**) component. The latter is referred to as the bias and acts as an additive offset term over **A**. The resulting assessment function was implemented as a dual model trained to detect mispronounced speech segments. The models incorporate Long-Short Memory and saliency region selection using attention. An experiment was performed using recordings from young Dutch learners of English as second language, which were annotated for mispronunciation by three trained phoneticians (a1, a2, a3). Both models **A** and **B** observed the same acoustic features. Only model **B** was made sensitive to the assessor identity η . The models combined were able to detect mispronunciations given the assessor identity achieving F1 scores of 0.77, 0.68 and 0.86 for a1, a2 and a3 respectively on the Train set and 0.66, 0.53 and 0.81 on the Test set. Model **B** was proven to be sensible to η . Model **A**, although being independent of η , was not similar to a consolidated annotation reference obtained via MAXVote. The attention weights of **A** and **B** were able to illustrate disagreements between assessors related to their own bias.

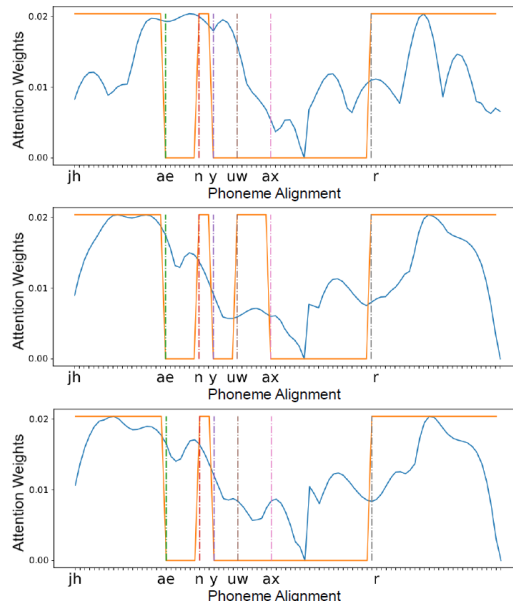


Figure 1 The normalised attention curves (blue) for both outputs focused differently on the same acoustic events. The curves correspond to **A** (top), **Ba2** (mid) and **Ba3** (bottom).

A siamese RNN architecture to detect deliberate imitation and phonetic convergence in L2-speech

Byron Z. Yuan^{1,2}, Aldo Pastore^{1,2}, Dorina De Jong^{1,2}, Hao Xu³, Luciano Fadiga^{1,2}, Alessandro D'Ausilio^{1,2}

¹Istituto Italiano di Tecnologia, Center for Translational Neurophysiology of Speech and Communication, (CTNSC), Ferrara, Italy

²Università di Ferrara, Dipartimento di Neuroscienze e Riabilitazione, Ferrara, Italy

³ University of California San Diego, Department of Computer Science and Engineering, CA, USA

zheng.yuan@iit.it

Abstract

Phonetic convergence describes the automatic and unconscious speech adaptation of two interlocutors in a conversation. We implemented a Siamese recurrent neural architecture to measure sentence-level phonetic convergence on a scripted speech dataset. Our Alternating Reading Task (ART) dataset was developed to study speech imitation and phonetic convergence, featuring three experimental conditions — solo, interaction, and deliberate imitation. The dataset consists of L2-English speakers - 9 native Italian dyads and 10 native French dyads. The Siamese RNN was trained and tested on the ART dataset for speaker verification, using the cosine similarity as an unbiased measure of convergence and imitation. We found that the Siamese RNN model achieved largely consistent results in modelling the dynamics of phonetic convergence and the speaker's imitation ability. Furthermore, this speaker-independent model has proved scalability and potential for coping with L1-induced speaker variability.

Index Terms: phonetic convergence, recurrent neural network (RNN), speech imitation, alternating reading task

1. Alternating reading task

We proposed a siamese recurrent neural architecture[1] to measure phonetic convergence and tested it on a scripted dialogue speech dataset, which we developed as a version of the alternating reading task[2]. There were three conditions in the task: speakers read each sentence from a story individually; speakers formed dyads and took turns to read the story; each speaker of the dyad was prompted to imitate one another while reading the sentences.

2. Siamese RNN Model

We used a two-network structure with each RNN network processing one audio input. The structure has tied weights. The model was trained for binary speaker verification, i.e., to predict whether the same speaker produced the two input speech utterances. In this way, the neural nets would learn a representation of a speaker's voice characterized by the underlying acoustic-prosodic features. By computing the cosine similarity of the speaker embedding vectors, we extracted a straightforward and unbiased measurement of phonetic convergence.

3. Results

Overall results of three groups of tests with the models in the first group testing solely on the French data, the second on Italian data and the third on French & Italian data. The metric is the binary accuracy of positive predictions, negative predictions and their mean value.

Model	Acc Mean	Acc Pos	Acc Neg
FR	0.9014	0.8716	0.9311
FR+ITA	0.8738	0.9097	0.8379
VCTK+FR+ITA	0.9075	0.8398	0.9751
ITA	0.7932	0.7554	0.8311
FR+ITA	0.7900	0.8994	0.6807
VCTK+FR+ITA	0.8152	0.7915	0.8389
FR+ITA	0.8406	0.9066	0.7745
VCTK+FR+ITA	0.8689	0.8239	0.9139

4. Imitation and convergence detection

We selected the VCTK+FR+ITA model for imitation and convergence detection as it outperformed other models in the speaker verification task involving both L1 backgrounds. We computed the average similarity scores of all speakers across 3 conditions. Results show that either the similarity score of the imitation condition (FR: 0.093 ± 0.168 ; ITA: 0.200 ± 0.269) or that of the interaction condition (FR: 0.090 ± 0.157 ; ITA: 0.150 ± 0.227) is comparatively higher than in the solo condition (FR: 0.035 ± 0.081 ; ITA: 0.038 ± 0.085). Besides, the imitation condition has a higher similarity score than the interaction condition. Considering the L1 background of the speakers, the Italian group shows higher similarity scores across all conditions. Looking into the 4 rounds of alternated reading in the interaction condition, the similarity scores present consistency across L1 French speakers (0.081 ± 0.143 ; 0.108 ± 0.182 ; 0.081 ± 0.149 ; 0.088 ± 0.149) and L1 Italian speakers (0.158 ± 0.239 ; 0.145 ± 0.224 ; 0.148 ± 0.215 ; 0.150 ± 0.231).

5. References

- [1] J. Mueller and A. Thyagarajan, "Siamese recurrent architectures for learning sentence similarity," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.
- [2] V. Aubanel and N. Nguyen, "Speaking to a common tune: Between-speaker convergence in voice fundamental frequency in a joint speech production task," *PLOS ONE*, vol. 15, 5 2020.

Using conversational data to improve prosody in Text-to-Speech synthesis

Johannah O'Mahony, Catherine Lai, Simon King

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

johannah.o'mahony@ed.ac.uk

For isolated utterances, speech synthesis quality has improved immensely thanks to the use of sequence-to-sequence models. However, these models are generally trained on read speech and fail to generalise to unseen speaking styles. Recently, more research is focused on the synthesis of expressive and conversational speech. Conversational speech contains many prosodic phenomena that are not present in read speech style, for example function words are more likely to be accented in conversational speech [1]. We would like to learn these prosodic patterns from data, but unfortunately, many large conversational corpora are unsuitable for speech synthesis due to low audio quality. We first investigate whether a data mixing strategy can improve conversational prosody for a target voice based on monologue data from audiobooks by adding real conversational data from podcasts. We filter the podcast data [2] to create a set of 26k question and answer pairs. We evaluate two FastPitch models: one trained on 20 hours of monologue speech from a single speaker, and another trained on 5 hours of monologue speech from that speaker plus 15 hours of questions and answers spoken by nearly 15k speakers. The results from our listening tests show that the second model generates more preferred question prosody. Results are summarised in figure 1.

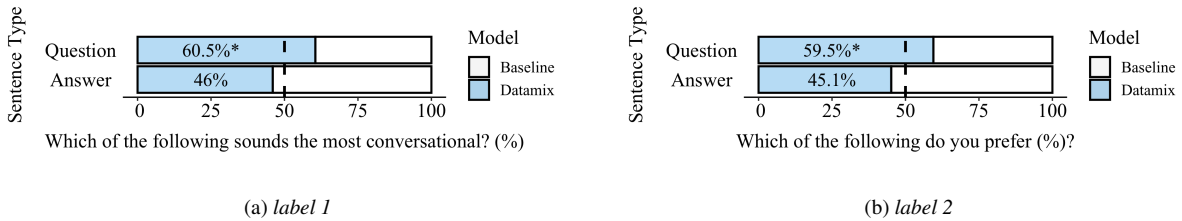


Figure 1: Preference Tests

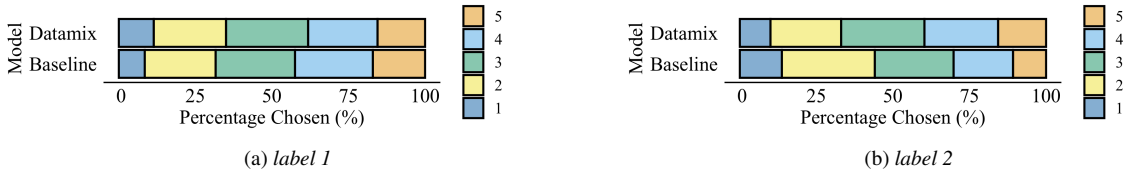


Figure 2: MOS Tests

In a second study, we use our corpus of question-answer pairs to fine-tune a BERT-based linear classifier on a prosodic labelling task using Continuous Wavelet Transform labels. Specifically, we are interested in evaluating whether the addition of the question context increases performance in prosodic labelling of the answer. We explore different task settings, for example comparing a three-way and two-way prominence distinction, as well as adding prosodic information from the question context. Mirroring recent findings [3], we find that the addition of context does not improve model performance in both task settings. We find, however, that the addition of a weighted loss to tackle class imbalance does help. We discuss appropriate baselines for prominence prediction, the need for subjective evaluation in balancing precision and recall, as well as better methods for incorporating context.

1. References

- [1] J. Yuan, J. M. Brenier, and D. Jurafsky, "Pitch Accent Prediction: Effects of Genre and Speaker," in *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology*, 2005.
- [2] A. Clifton, S. Reddy, Y. Yu, A. Pappu, R. Rezapour, H. Bonab, M. Eskevich, G. Jones, J. Karlgren, B. Carterette, and R. Jones, "100,000 Podcasts: A Spoken English Document Corpus," in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, 2021, pp. 5903–5917.
- [3] B. Stephenson, L. Besacier, L. Girin, and T. Hueber, "BERT, can HE predict contrastive focus? Predicting and controlling prominence in neural TTS using a language model," 7 2022. [Online]. Available: <http://arxiv.org/abs/2207.01718>

Non-Linear Pairwise Language Mappings for Low-Resource Multilingual Acoustic Model Fusion

Muhammad Umar Farooq, Darshan Adiga Haniya Narayana, Thomas Hain

Speech and Hearing Research Group, University of Sheffield, UK.

mufarooq1@sheffield.ac.uk

Abstract

Multilingual speech recognition has drawn significant attention as an effective way to compensate data scarcity for low-resource languages. End-to-end (e2e) modelling is preferred over conventional hybrid systems, mainly because of no lexicon requirement. However, hybrid DNN-HMMs still outperform e2e models in limited data scenarios. Furthermore, the problem of manual lexicon creation has been alleviated by publicly available trained models of grapheme-to-phoneme (G2P) and text to IPA transliteration for a lot of languages. In this paper, a novel approach of hybrid DNN-HMM acoustic models fusion is proposed in a multilingual setup for the low-resource languages. Posterior distributions from different monolingual acoustic models, against a target language speech signal, are fused together. A separate regression neural network is trained for each source-target language pair to transform posteriors from source acoustic model to the target language. These networks require very limited data as compared to the ASR training. Posterior fusion yields a relative gain of 14.65% and 6.5% when compared with multilingual and monolingual baselines respectively. Cross-lingual model fusion shows that the comparable results can be achieved without using posteriors from the language dependent ASR.

Index Terms: automatic speech recognition, low-resource, model fusion, multilingual, cross-lingual

This work was partly supported by LivePerson Inc. at the LivePerson Research Centre.

RoomReader: A Multimodal Corpus of Online Multiparty Conversational Interactions

Justine Reverdy¹, Sam O'Connor Russell¹, Louise Duquenne¹,
Diego Garaialde², Benjamin Cowan², Naomi Harte¹

¹Sigma Group, ADAPT Centre, School of Engineering, Trinity College Dublin

²ADAPT Centre, School of Information and Communication Studies, University College Dublin

¹{reverdyj, russelsa, nharte}@tcd.ie, ²{diego.garaialde, benjamin.cowan}@ucd.ie

Abstract

We present RoomReader, a corpus of multimodal, multiparty conversational interactions in which participants followed a collaborative student-tutor scenario designed to elicit spontaneous speech. The corpus was developed within the wider RoomReader Project to explore multimodal cues of conversational engagement and behavioural aspects of collaborative interaction in online environments. However, the corpus can be used to study a wide range of phenomena in online multimodal interaction. The publicly-shared corpus consists of over 8 hours of video and audio recordings from 118 participants in 30 gender-balanced sessions, in the “in-the-wild” online environment of Zoom. The recordings have been edited, synchronised, and fully transcribed. Student participants have been continuously annotated for engagement with a novel continuous scale. We provide questionnaires measuring engagement and group cohesion collected from the annotators, tutors and participants themselves. We also make a range of accompanying data available such as personality tests and behavioural assessments. The dataset and accompanying psychometrics present a rich resource enabling the exploration of a range of downstream tasks across diverse fields including linguistics and artificial intelligence. This could include the automatic detection of student engagement, analysis of group interaction and collaboration in online conversation, and the analysis of conversational behaviours in an online setting.

Index Terms: multimodal corpus, multiparty interaction, online interaction, video-conferencing, engagement detection

1. Problem

The multiple lock-downs that were enforced in many countries to slow the spread of the COVID-19 pandemic forced workers, teachers and students to rapidly adapt to an online environment to continue their professional, educational and social activities. This situation highlighted the multiple issues accompanying video-conferencing interactions (e.g., fatigue, time latency leading to missed or distorted turn-taking cues, or difficulties to perceive disengagement cues from interlocutors), and the lack of available datasets.

2. Conversational Engagement in Educational Context

We define *conversational engagement* by the degree of involvement of students in a topic being discussed and their willingness to continue the interaction. It can be analysed along three dimensions: from visual cues, from linguistic cues, and from elements of the dialogue structure relevant to group cohesion.

3. Main Outcomes

A labelled dataset of 30 online tutorial-style sessions comprising of 118 participants, consisting of 8h44m of multimodal recordings. A collaborative student-tutor scenario aiming to elicit spontaneous speech, [1] with full transcriptions with utterance, word and phoneme level boundaries, along with engagement annotations and associated metrics. [2]

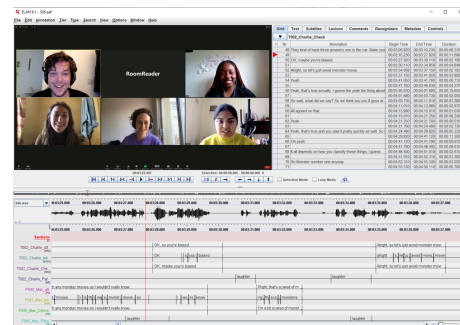


Figure 1: Example of a session in ELAN with participants transcription tiers (ASR and manually corrected).

Expert annotators were recruited to individually annotated the participants along a scale adapted from [3] to fit conversational online settings. The corpus mainly aims to be used for automatic engagement detection, online conversation analysis and the comparison of online vs face-to-face multiparty dialogues.

4. Acknowledgements

This research was supported by Science Foundation Ireland under Grant Agreement No. 20/COV/8537 and 13/RC/2106.P2 at the ADAPT SFI Research Centre at Trinity College Dublin and University College Dublin.

5. References

- [1] M. Koutsombogera and C. Vogel, “Modeling collaborative multimodal behavior in group dialogues: The MULTISIMO corpus,” in *Proceedings of LREC*, 2018, pp. 2945–2951.
- [2] J. Reverdy, S. O’Connor Russell, L. Duquenne, D. Garaialde, B. R. Cowan, and N. Harte, “Roomreader: A multimodal corpus of online multiparty conversational interactions,” in *Proceedings of LREC*, 2022, pp. 2517–2527.
- [3] P. Goldberg, Ö. Sümer, K. Stürmer, W. Wagner, R. Göllner, P. Gerjets, E. Kasneci, and U. Trautwein, “Attentive or not? toward a machine learning approach to assessing students’ visible engagement in classroom instruction,” *Educational Psychology Review*, pp. 1–23, 2019.

Alternative Evaluation Methods of Latent Representations of Speech Audio

Eimear Stanley, Yumnah Mohamied and Peter Bell

University of Edinburgh, UK

Email: s2239604@ed.ac.uk

Self-supervised models have recently shown promising results in the field of speech recognition, by generating latent representations of speech audio, which are then fine-tuned for downstream tasks. Linear classification has become the standard method for evaluating such latent representations. However, it performs poorly for any representations that are not linearly separable, and does not enable a direct analysis of the representations due to the use of gradient-based learning. This work proposes two simple algorithms as alternative evaluation methods: Gaussian classification and k-nearest neighbours (KNN), with a stronger focus on the latter. Using the latent representations generated by a pre-trained self-supervised model, phone classification tasks were carried out using linear, KNN and Gaussian classification. This showed that the latent representations learnt phonetic knowledge, the extent of which varied across layers of the model. The results provided an initial understanding of the phonetic information learnt by the latent representations. The similarity in performance between the KNN and linear classifiers suggests that the KNN is a suitable alternative evaluation method. We believe this will enable fairer comparisons across representations with different underlying forms, while using significantly less data.

Poster Session C

Autovocoder: Vocoding Without Spectrograms

Jacob J Webber, Simon King

The Centre for Speech Technology Research, University of Edinburgh

j.j.webber@ed.ac.uk, Simon.King@ed.ac.uk

Table 1: Comparing encoding and decoding of proposed system with current state-of-the-art (SOTA)

System	Acoustic feature representation (encoder)	Waveform generation (decoder)
SOTA	fast DSP	slow neural network
Autovocoder	learned	fast DDSP + simple deep network

Abstract

Many text-to-speech systems separate waveform generation from the rest of the synthesis process. Using mel-spectrograms as an intermediate representation introduces a dependency on neural vocoding. The autocorrelated nature of waveforms and lossiness of mel-spectrograms can require the use of autoregressive, generative vocoders, which are slow during training and inference.

Neural Source-Filter and LPCNet generate periodic elements of the signal explicitly and refine these with learned filters. These still require hand-crafted signal-processing derived input features. The use of hand-crafted representations can yield redundancy and the omission of key information that is difficult to recover, such as phase.

State-of-the-art vocoders use signal-processing derived features and machine-learned systems to invert them. We reverse this arrangement.

In autovocoder, the final signal is represented by frame-based, learned features of similar dimensionality to standard mel-spectrograms. The proposed system resembles an *autoencoder*, a deep learning system which encodes a signal, and in the process derives a compressed representation of the original input. This representation is inverted using deep networks, but final waveform generation is done explicitly by a differentiable implementation of the iSTFT. Figure 1 shows the full architecture of the proposed system.

As shown in figure 2, autovocoder produces high quality speech in the copy synthesis task, with audio output at similar quality to Griffin-Lim, which is known to perform well when inverting ground-truth spectrograms.

The computational performance of autovocoder by far exceeds comparable systems, including Griffin-Lim. Fast, differentiable waveform generation could facilitate end-to-end speech synthesis systems. Such systems have previously been inhibited by the need to train slow, waveform generating components alongside sequence models.

As shown in Table 2, *autovocoder* generates a waveform more than $4 \times$ faster than Griffin-Lim and many times faster than the neural waveform generators. Performance improvements of *autovocoder* over autoregressive systems should be even larger on hyper-parallel architectures such as GPUs.

Index Terms: speech synthesis, neural vocoder, differentiable DSP

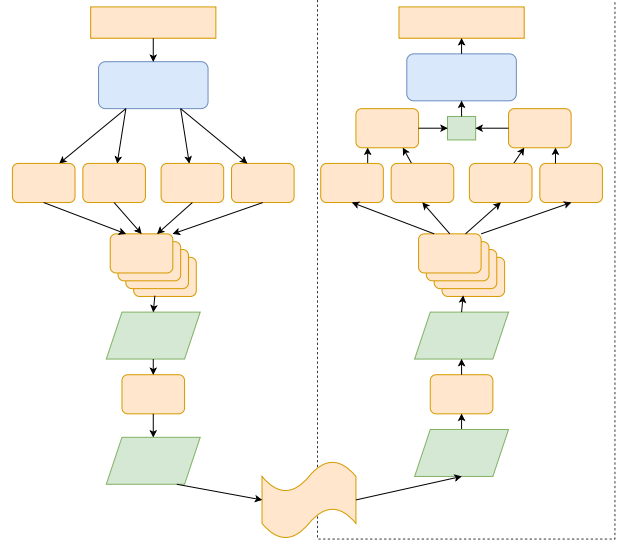


Figure 1: Autovocoder architecture. Dimensions, where given, show what is stored for 1 s of waveform

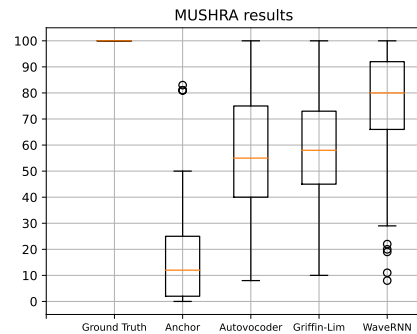


Figure 2: Results of the MUSHRA evaluation. All pairs apart from Griffin-Lim and Autovocoder are significantly different at $p > 0.01$

Table 2: Speed of each system against real time. Higher values indicate faster generation; a value below 1 indicates slower than real time.

System	Number of times faster than real-time
Griffin-Lim	$18.43 \times$
Autovocoder	$81.68 \times$
WaveRNN	$0.47 \times$
LPCNet	$1.50 \times$

Exploring Prosody Transfer in Speech Synthesis

Atli Thor Sigurgeirsson, Simon King

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

s2063518@ed.ac.uk, Simon.King@ed.ac.uk

Some recent models for Text-to-Speech synthesis learn to transfer the prosody of a reference utterance to the generated target synthetic speech [1, 2, for example]. This is achieved by using a learned embedding of the reference utterance spectrogram, which is used to condition speech generation. During model training, the reference utterance is identical to the target utterance. Yet, during synthesis, these models are used to transfer prosody from a reference that differs from the text or speaker being synthesized. We address this inconsistency by using

Table 1: MOS results for both real and synthesized samples. Target-speakers are chosen at random.

Model	MOS	
Ground Truth	4.2 ± 0.1	
Ground Truth + HiFi-GAN	3.7 ± 0.1	
	Same-text	Different-text
shuffle	2.8 ± 0.2	2.9 ± 0.2
text-based	2.6 ± 0.2	2.4 ± 0.2
F_0-based	2.9 ± 0.2	2.6 ± 0.2
Daft-Exprt	3.2 ± 0.2	2.4 ± 0.2

Table 2: PT MUSHRA-like scores and target-speaker classification accuracy. Target-speakers are randomly sampled.

Model	MUSHRA-like		Speaker classif.
	Same-text	Different-text	
shuffle	39.0 ± 4.2	25.5 ± 3.1	91.5%
text-based	38.7 ± 4.4	30.4 ± 3.2	88.4%
F_0-based	42.9 ± 4.5	28.4 ± 3.1	91.0%
Daft-Exprt	61.5 ± 4.8	49.3 ± 4.1	46.1%

a different, but prosodically-related, utterance as the reference during training too.

We experiment with two ways of selecting prosodically relevant utterances. First we consider the case where the reference utterance is of the same text as the target utterance but read by a different speaker than the target-speaker. As this generates pairs of the same text there are likely no text-based prosodic-similarity methods that would yield more prosodically-similar utterance pairs. We use a parallel speech corpus [3] to create such (reference, target) pairs. We refer to the model trained on these pairs as **text-based**. To compare to the above text-based selection of the reference, we devised a method based on F_0 similarity because F_0 is the principal acoustic correlate of prosody. We follow [4] in determining prosodic similarity based on F_0 . The model trained with (reference, target) pairs selected in this way is referred to as **F_0 -based**.

We evaluated naturalness, preservation of target-speaker identity, and prosody transfer (PT) for the two training schemes and two additional schemes: our baseline model, referred to as **Daft-Exprt** and is based on [2], trained in the typical manner where the reference and target are identical during training; **shuffle** where the reference is a randomly-selected utterance, which is unlikely to be informative about the target utterance prosody. To evaluate target-speaker identity preservation, we used a discriminative AXY test which asked listeners to indicate whether synthesized sample A sounds more like a ground-truth sample from target-speaker X or a ground-truth sample from the reference speaker Y. We then followed [2] and used a MUSHRA-like test for evaluating prosody transfer, using **shuffle** as the anchor. Each screen was rated by at least 8 different listeners recruited via Prolific¹.

The results indicate that the proposed training schemes, **text-based** and **F_0 -based**, have a detrimental effect on the quality of perceived prosody transfer and are instead comparable to **shuffle**. **Daft-Exprt** is rated better than other models in terms of PT but different-text PT has a significant impact on perceived naturalness to the point where **Daft-Exprt** is actually rated worse than **shuffle**. Furthermore, both subjective and objective results confirm that **Daft-Exprt** does not manage to disentangle prosody and speaker identity in different-speaker PT whilst the other models do. Prosody is text- and speaker-dependent. Therefore, a transferable representation of prosody has to be invariant to the reference speaker and reference text so that it can be applied to any target text and speaker. Since the representation modeled by **Daft-Exprt** is dependent on both the reference speaker and reference text we conclude that the prosodic representation modeled by **Daft-Exprt** is not transferable.

1. References

- [1] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *international conference on machine learning*. PMLR, 2018, pp. 4693–4702.
- [2] J. Zaidi, H. Seuté, B. van Niekerk, and M.-A. Carboneau, "Daft-exprt: Robust prosody transfer across speakers for expressive speech synthesis," *arXiv preprint arXiv:2108.02271*, 2021.
- [3] M. S. Ribeiro *et al.*, "Parallel audiobook corpus," 2018.
- [4] A. Rilliard, A. Allauzen, and P. Boula-de-Mareuil, "Using dynamic time warping to compute prosodic similarity measures," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

¹<https://www.prolific.com/>

Code-switched Text Generation on Parallel Data

Jie Chi¹, Brian Lu²

¹CSTR, University of Edinburgh

²CLSP, Johns Hopkins University

jie.chi@ed.ac.uk, zlu39@jhu.edu

Abstract

Code-switching refers to the alternation of languages within a conversation, a phenomenon that is of increasing importance considering the rapid rise in the number of bilingual speakers in the world. It is particularly challenging for ASR owing to the relative scarcity of code-switching speech and text data, even when the individual languages are themselves well-resourced. This work proposed to overcome this challenge by generating realistic code-switching sentences from their translations into L1 and L2. The generated code-switching text is then used to improve language modelling in ASR. We investigated two encode-decoder based approaches, and conducted experiments on SEAME [1], a Mandarin-English code-switched speech corpus. Firstly, transcripts of code-switched utterances in the training set (see [2] for data split) were translated to both English and Mandarin with Google Translator. Then, inspired by [3], we trained a pointer-generator network on them with a different setting. Two monolingual translations are concatenated as input and the code-switched sentences are used as output. Next, to further alleviate the dependence on code-switched data, we trained a transformer encoder-decoder model on parallel text only. We trained the model to translate both to L1 and L2 given either L1 or L2. Given a sentence in L1 or L2, we generated its code-switched variants via constrained beam search where we set a target number of switching points. To evaluate the usefulness of generated text for ASR, we trained separate trigram LMs on each type of generated text. In addition to the perplexity of LM on devset, an LFMMI-TDNN acoustic model train on Kaldi was also used to evaluate LM in ASR. As shown in Table 1, we found that regarding overall PPL and WER, both pointer-generator and finetuned models outperform the baseline, which is the trigram LM trained on only monolingual data and translations, and the improvement becomes more significant if only evaluating on CS subset. We also further showed that the improvement actually comes from a better modelling on CS boundaries by illustrating the cross-entropy breakdown on boundaries as shown in Figure 1.

	Baseline	PointGen	Pretrain	Finetune
PPL_{all}	112.34	106.29	151.41	110.64
PPL_{cs}	122.53	106.27	149.73	108.49
PPL_{mono}	91.44	106.36	155.48	116.12
WER_{all}	29.25/34.50	27.77/33.76	30.18/35.17	29.13/34.72
WER_{cs}	28.79/34.97	26.83/33.13	29.19/35.40	28.06/34.39
WER_{mono}	31.08/33.84	31.36/34.62	33.94/36.13	33.23/35.17

Table 1: PPL and WER on SEAME dev sets, where PPL is evaluated on the combination of dev_{man} and dev_{sge} , and WER is evaluated individually on dev_{man}/dev_{sge} .

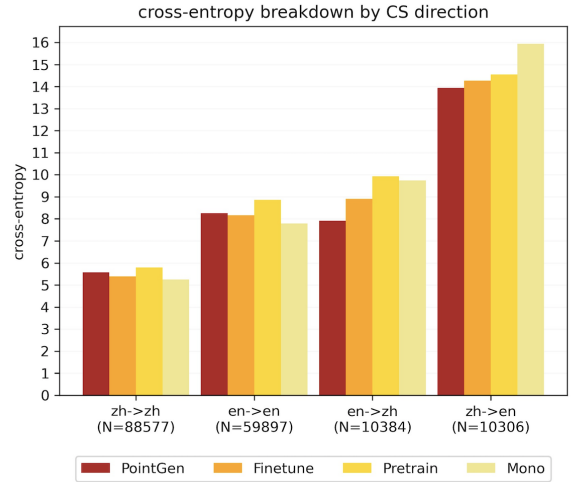


Figure 1: Cross-entropy breakdown by the direction of code-switching at token boundaries.

1. Acknowledgements

This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Sciences.

2. References

- [1] D.-C. Lyu, T.-P. Tan, E.-S. Chng, and H. Li, "Mandarin—english code-switching speech corpus in south-east asia: Seame," *Lang. Resour. Eval.*, vol. 49, no. 3, p. 581–600, sep 2015. [Online]. Available: <https://doi.org/10.1007/s10579-015-9303-x>
- [2] Z. Zeng, Y. Khassanov, V. T. Pham, H. Xu, E. S. Chng, and H. Li, "On the end-to-end solution to mandarin-english code-switching speech recognition," *CoRR*, vol. abs/1811.00241, 2018. [Online]. Available: <http://arxiv.org/abs/1811.00241>
- [3] G. I. Winata, A. Madotto, C. Wu, and P. Fung, "Code-switched language models using neural based synthetic data from parallel sentences," *CoRR*, vol. abs/1909.08582, 2019. [Online]. Available: <http://arxiv.org/abs/1909.08582>

Voice Puppetry for the People: Harnessing Dramatic Performance for Speech Synthesis

Matthew P. Aylett^{1,2}, Skaiste Butkute¹ and Christopher J. Pidcock¹

¹CereProc Ltd., ²Scott-Morgan Foundation

matthewaylett@gmail.com

Abstract

State-of-the-art speech synthesis owes much to modern AI machine learning, with recurrent neural networks becoming the new standard. However, how you say something is just as important as what you say. If we draw inspiration from human dramatic performance, ideas such as artistic direction can help us design interactive speech synthesis systems which can be finely controlled by a human voice. This “voice puppetry” has many possible applications from film dubbing to the pre-creation of prompts for a conversational agent. Previous work in voice puppetry has raised the question of how such a system should work and how we might interact with it. In this paper we describe a prototype vocal puppetry system and discuss a number of use cases for the system.

Index Terms: intonation, speech synthesis, AAC, social-robots

1. Introduction

Copy resynthesis is a technique where the parameterisation of human voice audio is used to directly control a speech synthesis rendition. It is a well established technique used to develop and evaluate speech synthesis systems (e.g. [1, 2]).

In addition, copy resynthesis can be used as a basis for a *voice puppetry* system. Such a system allows natural speech input to control the output speech for a target voice [3, 4]. This contrasts with, but is related to, *voice morphing*, where a source speaker’s voice is converted directly into a target speaker’s voice without the requirement of a speech synthesis system. So called *Puppetry* systems are commonly used to control the rendering of graphics and lip-syncing, where human movements are mapped onto a potentially very different body form. The ability to extend this control to the vocal performance of an artificial character’s voice, for example to recite poetry, is an important area of research. For an example of puppetry see <https://tinyurl.com/yxpy88y4> which demonstrates a female synthetic voice following a source male voice reading Shakespeare Sonnet 18.

Vocal puppetry is ideal for using speech synthesis in a performance context. For example it can transfer the intonation and timing of a joke. In this example you can hear a source voice and then two synthesis voices completing the joke based on the source voice <https://tinyurl.com/2hrb9pf6>.

A second important use case is for disabled persons who use speech synthesis as a voice replacement. Here it can be used to finely control the performance of speeches and pre-recorded interviews. Vocal puppetry was used by Peter Scott-Morgan in an interview with Stephen Fry in the Hay Book Festival and to address the House of Lords of the Disabilities Rights Act in 2021. An actor rendered the speech, it was then converted into speech synthesis mark-up which could then be edited and altered by the user to get exactly the rendition required.

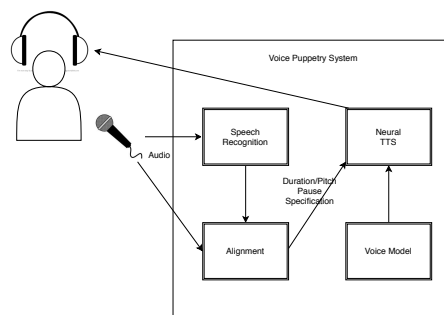


Figure 1: A voice puppetry system depends on speech recognition, speech alignment and neural TTS to reproduce output in a target voice.

```
<break time="1.100" type="4"/><usel dur_tgt="0.188 0.063 0.070"
f0_tgt="170.314 170.314 170.143 170.029 170.057 169.429" nrg_tgt=
"6.79 8.28 8.56"><lex phonemes=" . sh a l l">shall
</lex></usel><usel dur_tgt="0.120" f0_tgt="149.657 143.714"
nrg_tgt="8.09"><lex phonemes=" . a i l">i l l
</lex></usel><usel dur_tgt="0.087 0.030 0.092 0.075 0.237" f0_tgt="141.429 139.829 138.686
137.886 132.971 122.914 125.086 126.114 130.457 125.543" nrg_tgt
="5.02 7.00 6.97 3.68 7.95"><lex phonemes=" . k @ 0 m . p e @ l">
compare</lex></usel><usel dur_tgt="0.109 0.171" f0_tgt="115.829
116.514 123.829 110.114" nrg_tgt="6.13 7.90"><lex phonemes=" . dh
i i l">thee</lex>
```

Figure 2: An example of XML mark-up for vocal puppetry for the first words of Sonnet 18 - “Shall I compare thee”

2. References

- [1] W. J. Holmes, “Copy synthesis of female speech using the JSRU parallel formant synthesiser,” in *First European Conference on Speech Communication and Technology*, 1989, pp. 2513–2516.
- [2] O. Turk and M. Schroder, “Evaluation of expressive speech synthesis with voice conversion and copy resynthesis techniques,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 965–973, 2010.
- [3] M. P. Aylett, D. A. Braude, C. J. Pidcock, and B. Potard, “Voice puppetry: Exploring dramatic performance to develop speech synthesis,” in *Proc. 10th ISCA Speech Synthesis Workshop*, 2019, pp. 117–120.
- [4] Y.-Y. Fan, S. Shin, and V. Samanta, “Evaluating expressiveness of a voice-guided speech re-synthesis system using vocal prosodic parameters,” in *Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion*. ACM, 2019, pp. 67–68.

Improving diagnostic procedures for epilepsy through automated recording and analysis of patients' history

Nathan Pevy¹ Heidi Christensen¹ Traci Walker² Markus Reuber³

¹ Department of Computer Science - The University of Sheffield

² Division of Human Communication Sciences - The University of Sheffield

³ Academic Neurology Unit, Department of Neuroscience, University of Sheffield

The common causes of Transient Loss of Consciousness (TLOC) are syncope, epilepsy and dissociative seizures. Questionnaire-based decision making tools for non-specialists who initially assess patients with TLOC (such as clinicians working in primary or emergency care) reliably differentiate between patients who have experienced syncope and those who have had seizures but are more limited in their ability to differentiate between epileptic and functional (dissociative) seizures (FDS). Previous conversation analysis research has demonstrated that qualitative expert analysis of how people talk to clinicians about their seizures can help distinguish between these two causes of TLOC. The objective of this research is to explore whether an automated analysis of spoken descriptions of TLOC can capture some of the interactional, linguistic, and semantic differences between people with epilepsy and FDS and improve the predictive performance of current questionnaire based methods.

There were 76 people with a diagnosis of syncope (N=16), epilepsy (N=24), or FDS (N=36) who completed a symptom and medical history questionnaire (iPEP) through the online web application, of which 61 continued and also completed the interaction with a virtual agent. All models were trained using Support Vector Machines and nested leave-one-cross validation. A model trained using the iPEP responses provided a baseline for the three-way classification. Predictions of epilepsy or FDS from the iPEP model were passed into the language analysis module that was designed to improve the misdiagnosis of people with epilepsy and FDS. The module contained three SVM models trained using independent feature sets extracted from the responses of people with epilepsy or FDS: features designed to measure formulation effort and uncertainty, 9 semantic categories measured using the LIWC application, and TF-IDF features based upon verbs, adjectives, and adverbs. These features were based around previous conversation analysis research. A meta-learner model was trained to predict the diagnosis for this subset of patients using the predictions from the iPEP and each of the language analysis models. The overall performance metric combined the iPEP predictions for the patients who were not incorporated into the language analysis and the meta-learner predictions for those who were.

The SVM trained using the iPEP responses for all participants demonstrated an accuracy of 65.8%. The accuracy of the formulation effort features, LIWC semantic categories, and TF-IDF features for the differentiation between epilepsy and FDS was 85.7%, 85.7%, and 75.5%, respectively. Combining the iPEP and language analysis improved the detection of epilepsy and FDS and increased the overall accuracy of the iPEP to 85.5%.

The results of this proof of principle study suggest that an automated analysis of language can help the differential diagnosis between epilepsy and FDS, and therefore improve the overall three-way classification of current clinical decision tools.

Deliberation-Based Multi-Pass Speech Synthesis

Qingyun Dou, Mark J. F. Gales

University of Cambridge, United Kingdom

{qd212,mjfg100}@cam.ac.uk

Abstract

Sequence-to-sequence (seq2seq) models have achieved state-of-the-art performance in a wide range of tasks including Neural Machine Translation (NMT) and Text-To-Speech (TTS). These models are usually trained with teacher forcing, where the reference back-history is used to predict the next token. This makes training efficient, but limits performance, because during inference the free-running back-history must be used. To address this problem, deliberation-based multi-pass seq2seq has been used in NMT. Here the output sequence is generated in multiple passes, each one conditioned on the initial input and the free-running output of the previous pass. This paper investigates, and compares, deliberation-based multi-pass seq2seq for TTS and NMT. For NMT the simplest form of multi-pass approaches, where the free-running first-pass output is combined with the initial input, improves performance. However, applying this scheme to TTS is challenging: the multi-pass model tends to converge to the standard single-pass model, ignoring the previous output. To tackle this issue, a guided attention loss is added, enabling the system to make more extensive use of the free-running output. Experimental results confirm the above analysis and demonstrate that the proposed TTS model outperforms a strong baseline.

Index Terms: attention-based sequence-to-sequence, exposure bias, speech synthesis

1. Introduction

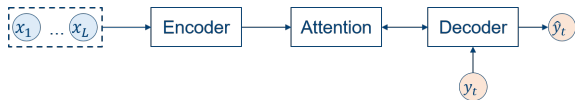


Figure 1: Illustration of an attention-based encoder-decoder

2. Deliberation-based multi-pass seq2seq

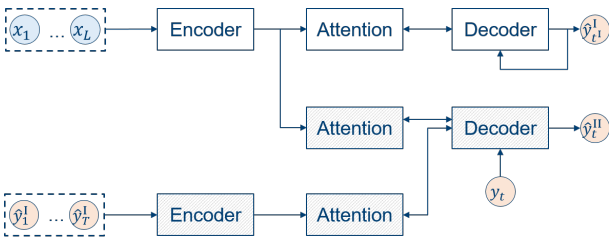


Figure 2: Illustration of deliberation-based multi-pass seq2seq

3. Experiments

3.1. Machine translation

Experimental results:

Table 1: BLEU of various NMT systems

model	BLEU $\uparrow$
TF	31.10 $\pm$ 0.27
SS	31.45 $\pm$ 0.45
AF	31.54 $\pm$ 0.14
FR-TF	31.74$\pm$0.27
TF-TF	31.29 $\pm$ 0.05

3.2. Speech synthesis

Experimental results:

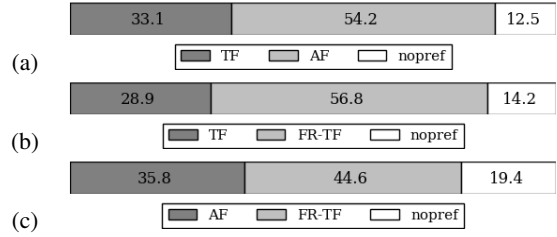


Figure 3: AB preference tests comparing TF, AF and FR-TF

Table 2: MOS, GV and DTW distance of various TTS systems

	MOS $\uparrow$	GV $\uparrow$	DTW $\downarrow$
reference	4.42 $\pm$ 0.09	0.0235	0
TF	3.67 $\pm$ 0.11	0.0171	6.29
SS	3.70 $\pm$ 0.12	0.0167	6.02
AF	3.89 $\pm$ 0.10	0.0219	5.59
FR-TF	4.03$\pm$0.10	0.0223	5.64
TF-TF	—	0.0136	6.73

Exploring Novel Methods for Automatic Speech Recogniser Based Intelligibility Prediction

Zehai Tu, Ning Ma, Jon Barker

University of Sheffield, Department of Computer Science, Sheffield, UK

{ztu3, n.ma, j.p.barker}@sheffield.ac.uk

1. Introduction

An accurate objective speech intelligibility prediction algorithm is important for many applications such as speech enhancement for hearing aids. Recently, automatic speech recognition (ASR) models are gaining more interest for intelligibility prediction, as they can show similar recognition patterns to human listeners. The majority of ASR-based intelligibility prediction systems leverage the ASR recognition correctness, such as word correctness scores (WCS), as the intelligibility predictor. However, this can fail in some situations, e.g., when the speech is highly distorted but the ASR still makes a correct *guess*.

We propose two novel methods for intelligibility prediction with deep neural network-based ASR that do not use ASR recognition results. Figure 1 shows the two proposed methods. The first method is intrusive, i.e. requires the clean speech signal as reference, and takes advantage of hidden representations of the ASR for intelligibility prediction [1]. Specifically, a DNN-based ASR model takes both a processed speech and its corresponding reference speech as inputs, and extracts their DNN hidden representations at the same layers. The similarity of the two hidden representations is then measured and used to estimate the intelligibility. The second method is non-intrusive, thus does not require the reference signal, and links the sequence-level uncertainty estimated by an ensemble of ASR to speech intelligibility [2]. The estimation is unsupervised [3] therefore does not require any listener intelligibility label. Experiments show they both can make more accurate intelligibility prediction than ASR WCS.

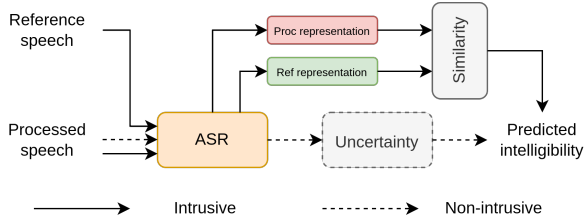


Figure 1: Proposed intrusive and non-intrusive methods for intelligibility prediction with ASR.

2. Experiments and results

The proposed methods are evaluated by the first round Clarity Prediction Challenge (CPC1) [4], which provides a large number of simulated domestic noisy speech signals enhanced by various hearing aid algorithms, and the corresponding recognition responses by hearing impaired listeners. The CPC1 includes two tracks: (1) *closed-set*, that is the listeners and systems in the evaluation set are overlapped with those in the training data; (2) *open-set*, that is the systems or listeners in the

Table 1: Evaluation results on the CPC1.

	RMSE ↓	NCC ↑	KT ↑
Closed-set			
Baseline	0.285	0.621	0.398
ASR WCS	0.250	0.729	0.523
Proposed intrusive method	0.231	0.773	0.498
Proposed non-intrusive method	0.233	0.768	0.499
Open-set			
Baseline	0.365	0.529	0.391
ASR WCS	0.250	0.723	0.534
Proposed intrusive method	0.235	0.763	0.530
Proposed non-intrusive method	0.246	0.734	0.512

evaluation set are not included in the training data. In order to simulate the hearing impairment, we use the Cambridge MSBG hearing loss model [5–7] to process the speech signals. The CPC1 baseline consists of the MSBG model and MBSTOI [8].

The results given in Table 1 indicate that the two proposed methods can outperform not only the CPC1 baseline, but also ASR WCS in terms of root-mean-square error (RMSE) and normalised cross-correlation (NCC). The poster will explain the proposed approaches in full and will provide a more detailed analysis of results.

3. References

- [1] Z. Tu, N. Ma, and J. Barker, “Exploiting hidden representations from a dnn-based speech recogniser for speech intelligibility prediction in hearing-impaired listeners,” in *INTERSPEECH*, 2022.
- [2] —, “Unsupervised uncertainty measures of automatic speech recognition for non-intrusive speech intelligibility prediction,” in *INTERSPEECH*, 2022.
- [3] A. Malinin and M. Gales, “Uncertainty estimation in autoregressive structured prediction,” in *ICLR*, 2020.
- [4] J. Barker, M. Akeroyd, T. J. Cox, J. F. Culling, J. Firth, S. Graetzer, H. Griffiths, L. Harris, G. Naylor, Z. Podwinska, E. Porter, R. Viveros Munoz *et al.*, “The 1st clarity prediction challenge: A machine learning challenge for hearing aid intelligibility prediction,” in *INTERSPEECH*, 2022.
- [5] T. Baer and B. C. J. Moore, “Effects of spectral smearing on the intelligibility of sentences in noise,” *JASA*, vol. 94, no. 3, pp. 1229–1241, 1993.
- [6] B. C. J. Moore and B. R. Glasberg, “Simulation of the effects of loudness recruitment and threshold elevation on the intelligibility of speech in quiet and in a background of speech,” *JASA*, vol. 94, no. 4, pp. 2050–2062, 1993.
- [7] M. A. Stone and B. C. Moore, “Tolerable hearing aid delays. i. estimation of limits imposed by the auditory path alone using simulated hearing losses,” *Ear and Hearing*, vol. 20, no. 3, pp. 182–192, 1999.
- [8] A. H. Andersen *et al.*, “Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions,” *Speech Communication*, vol. 102, pp. 1–13, 2018.

View-Specific Assessment of L2 Spoken English

S. Bannò^{1,2,3}, B. Balusu³, M.J.F. Gales³, K.M. Knill³, K. Kyriakopoulos³

¹Fondazione Bruno Kessler, Trento, Italy

²University of Trento, Trento, Italy

³Cambridge University Engineering Department, Cambridge, UK.

sbanno@fbk.eu, {mjfg100,kmk1001}@cam.ac.uk

Abstract

The growing demand for learning English as a second language has increased interest in automatic approaches for assessing and improving spoken language proficiency. A significant challenge in this field is to provide interpretable scores and informative feedback to learners through individual viewpoints of learners’ proficiency, as opposed to holistic scores. Thus far, holistic scoring remains commonly applied in large-scale commercial tests. As a result, an issue with more detailed evaluation is that human graders are generally trained to provide holistic scores. This paper investigates whether view-specific systems can be trained when only holistic scores are available. To enable this process, view-specific networks are defined where both their inputs and structure are adapted to focus on specific facets of proficiency. It is shown that it is possible to train such systems on holistic scores, such that they provide view-specific scores at evaluation time. View-specific networks are designed in this way for pronunciation, rhythm, text, use of parts of speech and grammatical accuracy. The relationships between the predictions of each system are investigated on the spoken part of the Linguaskill proficiency test. It is shown that the view-specific predictions are complementary in nature and capture different information about proficiency.

Only holistic scores are available for most spoken language assessment training data sets. Thus, the training data set comprises $\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}$ where $\mathbf{x}^{(i)}$ is the set of features, or sequence of features, extracted from the audio and ASR system, and $y^{(i)}$ the associate reference score. To train view-specific models the assessment process can be split into two distinct stages, where initially the features $\mathbf{x}$ are mapped to view-specific features $\mathbf{v}$, and then fed into the score-prediction network. Thus, for a particular view

$$\hat{y}_v^{(i)} = \mathcal{F}_v(\mathbf{x}^{(i)}) = f_v(\mathbf{g}_v(\mathbf{x}^{(i)})) = f_v(\mathbf{v}^{(i)}) \quad (1)$$

where the desired training data comprises $\mathcal{D}_v = \{\mathbf{x}^{(i)}, y_v^{(i)}\}$. Unfortunately, there are no view-specific reference grades, $y_v^{(i)}$, associated with each of the training observations, $\mathbf{x}^{(i)}$, just overall holistic grades, $y^{(i)}$. To address this problem, the form of the feature extractor $\mathbf{g}_v(\mathbf{x}^{(i)})$ is constrained so that only information about a specific view is contained within $\mathbf{v}^{(i)}$. For example, if only information about the text spoken is in $\mathbf{v}^{(i)}$, irrespective of the pronunciation of the words, then the same feature vector $\mathbf{v}$ can be obtained from the different values of $\mathbf{x}$.

Table 1 shows the performance on a multi-level free speaking English test of 5 single-view graders and their combination in contrast to the baseline holistic grader in terms of RMSE, considering both the individual models and the ensembles. As can be seen, the ensemble approach gives a significant improvement on all the graders, including the baseline. Performance varies across views with the text view yielding the best performance, and the rhythm and grammatical error correction (es) and part-of-speech (pos) graders about 0.1 RMSE worse. Combination of the graders yields an improvement over the baseline showing they are complementary for the task of predicting holistic grades. Figure 2 shows that differences can be observed in the RMSE against proficiency level depending on the view being assessed. This should enable the holistic score to be more interpretable, providing feedback to learners on specific aspects of their speaking skills to address.

Model	Indiv.	Ens.
baseline	0.522±0.009	0.406
pron ($\mathbf{p}_r$)	0.454±0.003	0.451
rhythm ($\mathbf{r}_y$)	0.551±0.037	0.490
text ($\mathbf{t}_x$)	0.409±0.007	0.409
es ($\mathbf{e}_s$)	0.497±0.001	0.495
pos ($\mathbf{p}_s$)	0.499±0.003	0.497
$\mathbf{p}_r \mathbf{r}_y \mathbf{t}_x \mathbf{e}_s \mathbf{p}_s$	-	0.386

Table 1: Performance of the single-view graders and baseline in terms of RMSE. Individual models vs ensembles.

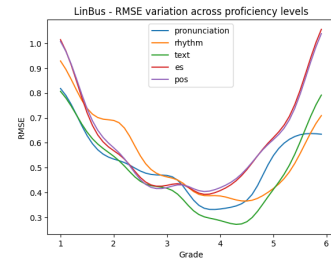


Table 2: RMSE variation across proficiency levels for pronunciation, rhythm, text, grammatical error correction (es) and part of speech (pos) grader views.

This paper reports on research partially supported by Cambridge University Press and Assessment. The authors would like to thank Dr. Linlin Wang for providing the ASR transcription, Vyas Raina for the text model, Yiting Lu for the GEC model, and the ALTA Speech Technology Project Team for general discussions and contributions to the evaluation infrastructure. Full paper to be presented at Interspeech 2022.

Why is My Social Robot so Slow? How a Conversational Listener can Revolutionize Turn-Taking

Matthew P. Aylett¹, Andrea Carmantini^{1,2} and David A. Braude¹

¹CereProc Ltd., ²University of Edinburgh, Informatics

matthewaylett@gmail.com

Abstract

Current machine dialog systems are predominantly implemented using a sequential, utterance based, two-party, speak-wait/speak-wait approach. human-human dialog is 1) not sequential, with overlap, interruption and back channels; 2) processes utterances before they are complete and 3) are often multi-party. The current approach is stifling innovation in social robots where long delays (often several seconds) is the current norm for dialog response time, leading to stilted and unnatural dialog flow. In this paper, by referencing a light weight word spotting speech recognition system - Chatty SDK, we present a practical engineering strategy for developing what we term a *conversational listener* that would allow systems to mimic natural human turn-taking in dialogue.

Index Terms: dialog, conversational listener, ASR, social-robotics

1. Introduction

“As conversational systems (in various forms) are becoming ubiquitous, it is clear that turn-taking is still not handled very well in those systems. They often tend to interrupt the user or have very long response delays, there is little timely feedback, and the flow of the conversation feels stilted.” Skantze [1] p1.

Most so-called, conversational systems used by social robots are, in reality, two-party, speak-wait/speak-wait systems. Human conversation in contrast, is often multi-party, allows for fluid interruption and back channeling. About 10% of the speech material is overlapped, with speakers often speaking, briefly, at the same time. Furthermore, human participants typically respond within 200ms, whereas current digital systems can spend several seconds processing before saying anything. This results in less fluid interaction and impacts the functionality of social robots in areas such as education - e.g. *“Certain occurrences of social referencing appeared exclusively in the interaction with the robot... after a delay in the dialogue occurred and the robot required too much time to provide an adequate utterance.”* [2], System designed for older users - e.g. *“the turn-taking delays in the dialogue were significant, which hinders the communication.”* [3] and results in systems being regarded as inferior and poor at carrying out their tasks - e.g. *“users not only rate the incremental system as more responsive, but also rate its recommendation performance as higher.”* [4].

There is a body of previous work looking at incremental dialog processing¹ (see [1] for a review) and there are toolkits available to implement incremental processing of dialog for example InproTK. These systems follow a waterfall design pattern

where each module can form hypotheses based on incremental input but allow replanning if these hypotheses are rejected as new data is processed. This approach has a number of severe drawbacks: processing data without end pointing and a right context often produces much inferior results to a system that *waits* for an utterance to conclude; the architecture is complex and difficult to debug and test; the extra processing power required is multiplied across all levels of the system. Thus, despite this work we are unaware of any commercial social robot that makes use of incremental processing to implement human style dialog turn-taking.

In this paper we suggest a hybrid approach to the problem. Rather than retooling the entire architecture of a system to deal with incremental processing we suggest adding a new module, a *conversational listener* to the system which would allow a more flexible approach to implementing human like dialog processing. This approach is informed by four observations:

1. Systems often have a strong expectation of the type of response a user is likely to make in a dialog context.
2. Human’s typically respond very quickly to dialog turns that require simple responses or contain predictable content. Whereas longer inter-turn intervals are typical when a response requires significant processing.
3. Human’s often start speaking before they have decided what they are going to say.
4. Before large scale, open domain, multi-speaker, continuous speech recognition was available legacy system made good use of processor efficient key word spotting.

2. References

- [1] G. Skantze, “Turn-taking in conversational systems and human-robot interaction: a review,” *Computer Speech & Language*, vol. 67, p. 101178, 2021.
- [2] N. F. Tolkendorf, C. E. Crawshaw, and K. J. Rohlfing, “Comparing the effects of a different social partner (social robot vs. human) on children’s social referencing in interaction,” in *Frontiers in Education*, vol. 5. Frontiers Media SA, 2021, p. 569615.
- [3] J. Oliveira, G. S. Martins, A. Jegundo, C. Dantas, C. Wings, L. Santos, J. Dias, and F. Perdigão, “Speaking robots: The challenges of acceptance by the ageing society,” in *RO-MAN 2017S*. IEEE, 2017, pp. 1285–1290.
- [4] V. Tsai, T. Baumann, F. Pecune, and J. Cassell, “Faster responses are better responses: Introducing incrementality into sociable virtual personal assistants,” in *9th International Workshop on Spoken Dialogue System Technology*. Springer, 2019, pp. 111–118.

¹All dialog processing is incremental in some respects because you don’t know what the next utterance will be. However, incremental in this context means processing before you discover the end-point of a dialog partner’s current utterance.

Creating New Voices using Normalizing Flows

*Piotr Biliński, Thomas Merritt, Abdelhamid Ezzer, Kamil Pokora, Sebastian Cygert,
Kayoko Yanagisawa, Roberto Barra-Chicote, Daniel Korzekwa*

Amazon Alexa

{bilipiot, thommer, ezzer, kamipoko, scygert, yakayoko, rchicote, korzekwa}@amazon.com

Abstract

Creating realistic and natural-sounding synthetic speech remains a big challenge for voice identities unseen during training. As there is growing interest in synthesizing voices of new speakers, here we investigate the ability of normalizing flows in text-to-speech (TTS) and voice conversion (VC) modes to extrapolate from speakers observed during training to create unseen speaker identities. Firstly, we create an approach for TTS and VC, and then we comprehensively evaluate our methods and baselines in terms of intelligibility, naturalness, speaker similarity, and ability to create new voices. We use both objective and subjective metrics to benchmark our techniques on 2 evaluation tasks: zero-shot and new voice speech synthesis. The goal of the former task is to measure the precision of the conversion to an unseen voice. The goal of the latter is to measure the ability to create new voices. Extensive evaluations demonstrate that the proposed approach systematically allows to obtain state-of-the-art performance in zero-shot speech synthesis and creates various new voices, unobserved in the training set. We consider this work to be the first attempt to synthesize new voices based on mel-spectrograms and normalizing flows, along with a comprehensive analysis and comparison of the TTS and VC modes.

Index Terms: speech synthesis, new voices, zero-shot, text-to-speech, voice conversion, normalizing flows

Phonetic Analysis of Self-supervised Representations of English Speech

Dan Wells, Hao Tang, Korin Richmond

The Centre for Speech Technology Research, University of Edinburgh

{dan.wells, hao.tang, korin.richmond}@ed.ac.uk

Abstract

Self-supervised speech representation learning aims to produce representations of unlabeled speech audio which are useful for some downstream task, such as automatic speech recognition or speech synthesis. Recent approaches, such as HuBERT [1], incorporate quantisation of continuous representations to discover vocabularies of discrete speech units. The quality of discovered units is often evaluated using metrics based on frame-level alignment with phone transcripts. For example, purity measures indicate the degree to which discrete units are shared across multiple phone labels, which might reveal confusions between individual sounds, or the diversity of units aligned to a single phone label, possibly corresponding to context-dependent or sub-phone level representations. These metrics are typically computed in aggregate across all frames in the test corpus, potentially hiding significant differences between individual phones. With this work, we provide more fine-grained analysis of the phonetic bases of discrete units extracted from English speech using a pre-trained HuBERT model. By aligning phones to a discrete unit vocabulary with restricted capacity, we gain insight into the relative priority of different aspects of phone articulation in HuBERT representations. Though we limit our analysis to English, these insights are made more cross-linguistically relevant by framing them in terms of common physical characteristics of broad articulatory classes rather than individual phones.

We find that discrete speech units from HuBERT (extracted at a fixed framerate of 50 Hz) often correspond to sub-phonetic events, and that fine dynamics such as the distinct closure and release portions of plosives tend to be represented by sequences of units, as seen in Figure 1. Moreover, particular units tend to be shared across similar acoustic regions for different phones (e.g. plosive closures across multiple places of articulation) more often with a limited vocabulary of 50 units, while 100 units offer more capacity to distinguish individual phones (e.g. with place-distinctive plosive release bursts represented by different units for different phones). Other phone classes appear to be less well represented by HuBERT units, but this may actually reflect our choice of phonemic labels in unit-phone alignment. For example, nasals in English are likely to assimilate to the place of articulation of a following plosive (e.g. ‘input’ /ɪmpʊt/ → [ɪmpʊt]), which will be reflected in discrete unit sequences derived from speech audio but not in the corresponding phonemic transcripts. Our work provides a reference for the phonetic properties of discrete units discovered by HuBERT, facilitating analyses of other speech applications based on this model, alongside such considerations for future evaluation of self-supervised speech representations.

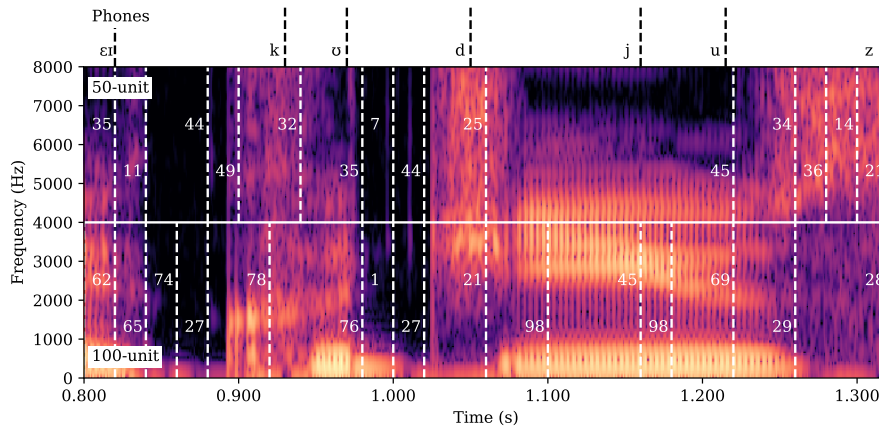


Figure 1: Spectrogram for the phrase ‘could use a boost’ from VCTK utterance p236.132, labeled with phones and units from 50- and 100-unit HuBERT models. Dynamic articulations are represented by sequences of discrete units, for example under the 50-unit model the initial plosive /k/ has distinct closure (unit 44), release burst (49) and aspiration (32) portions.

Acknowledgements: This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Sciences.

1. References

- [1] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

Comparison of Audio-Visual Speech Enhancement Models with Hearing Aid Key Performance Indicators

Jasper Kirton-Wingate¹, Mandar Gogate¹, Tassadaq Hussain¹, Amir Hussain¹

¹Edinburgh Napier University

jasper.kirton-wingate@napier.ac.uk, mandar.gogate@napier.ac.uk, t.hussain@napier.ac.uk,
a.hussain@napier.ac.uk

1. Abstract

Inspired by human cognition, several methods exist for simultaneously utilising audio and visual inputs for de-noising speech from target speakers. Recently Audio-Only (AO) Deep Learning (DL) based approaches have improved performance by learning end-to-end non-linear and highly sequence-time context informed mappings from masked to cleaned speech signals. Additionally, modelling the correlation between audio and visual target speech allows better separation of the speech signal from the audio mixture as the visual signal is (usually) noise free in a face-to-face conversation. These methods are promising for helping Hearing Aid (HA) and Cochlear Implant users understand speech in noisy everyday situations. However, HA's are small, low power devices and the utilisation of visual data bears additional computational cost. The speech signals and listening scenarios that a HA user is subject to will also be highly variant, and thus a Speech Enhancement (SE) algorithm must generalise well to these signals and contexts. Additionally, some users will be more sensitive to degradation of speech quality than others. As a result, the efficacy of a SE model for a HA is based on trade-offs between several Key Performance Indicators (KPIs).

There are several classes of DNN AV SE models and pre-processing methods that will be compared in this respect. Methods can be split into 3 classes: Direct Mapping (DM), Mask Approximation (MA) and Indirect Mapping (IM) [1]. Within these methods, there are different approaches to Multi-Modal Fusion (MMF) between Audio and Visual channels: such as Concatenation based, Addition/Product based, Squeeze-Excitation fusion and Attention based. Furthermore, there are various Objective Functions (OF) used for training, such as MAE, MSE and STOI.

Different models are evaluated in terms of cost of the product of training on a single example E , the size of the training dataset D and the number of hyper-parameter experiments H . For this study, SOTA methods previously evaluated on the GRID and TCD-TIMIT corpus in terms of commonly used metrics like STOI and PESQ [1], are chosen from each of the model classes to review. These models are then evaluated on the TCD-TIMIT dataset which has a gender balanced set of 62 speakers with a phonetically balanced group of sentences, and also includes 30 degree camera orientations. The performance is evaluated at a range of realistic SNRs for the CHIME-3 noise set to evaluate the performance metrics [2].

The following enhancement performance metrics are used in evaluation due to their importance for Hearing Impaired listeners: the Hearing Aid Speech Quality index (HASQI) and the Hearing Aid Speech Perception Index (HASPI). These metrics evaluate the signal after a pathology induced signal degradation inferred from a filter bank model simulating inner and outer hair cell loss, which can then be compared with the enhanced (or clean) signals. The computational complexity in Floating Point Operations (FLOPs) per training and testing example is also given for each architecture, alongside the number of model parameters. Additionally, the latency on a typical GPU is given and discussed in terms of current HA hardware. The complexity in Big O notation for any pre-processing and OF is also discussed. The aim of this study is to review current methods whilst providing a general framework for HA focused model comparison, highlighting trade-offs with model complexity and efficiency in terms of evaluation scores and model generalisation, in order to give researchers and industry the opportunity to review the cost and performance of model deployment, training or fine-tuning.

1.0.1. KPI's and Comparison

Table 1: An example of two HA algorithms with KPIs [3]

Model	SNR (DB)	HASQI	HASPI	FLOPs (M)	Parameters (M)	Latency (ms)
CRNN	—	—	—	3.58	9.51	
TAP – CRNN	—	—	—	5.39	11.03	

Index Terms: Speech Enhancement, Machine Learning, Hearing Aids, Computational Efficiency, Green AI

2. References

- [1] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1368–1396, 2021.
- [2] E. V. S. W. Jon Barker, Ricard Marxer, "The third 'chime' speech separation and recognition challenge: Dataset, task and baselines," *2015 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2015)*, Dec 2015, Scottsdale, AZ, United States. [jfh101211376](#).
- [3] T. Hussain, W.-C. Wang, M. Gogate, K. Dashtipour, Y. Tsao, X. Lu, A. Adeel, and A. Hussain, "A novel temporal attentive-pooling based convolutional recurrent architecture for acoustic signal enhancement," *IEEE Transactions on Artificial Intelligence*, pp. 1–1, 2022.

Simulation of Teacher-Learner Interaction in English Language Pronunciation Learning

Elaf Islam¹, Thomas Hain²

Speech and Hearing Group, Department of Computer Science, University of Sheffield, UK

¹ ejislam1@sheffield.ac.uk, ² t.hain@sheffield.ac.uk

Abstract

Acquiring a second language (L2) is a challenging process, while learners tend to depend on the norms and categories of their L1 while learning to interpret L2 sounds. Also, each learner has a distinct language learning strategy which is particular actions or behaviours that learners use to learn L2. Using a suitable teaching strategy to give corrective feedback improves the learner's language acquisition. Therefore, many researchers study the teacher and learner's interactions in language learning. Simulations could be used to improve learning process, optimise processes, and test learning and teaching theories. To simulate the teacher role, Computer-assisted language learning (CALL) presented. Some other research focuses on identifying the cognitive and perceptual operations that enable the learner mind to act. However, There is a little computational simulation of the learning system, but without joining the teacher system and the learner system, especially in the English language learning field. Here we simulate teacher-learner interaction in English pronunciation learning. Simulation the pronunciation learning process involves modeling the teacher, the learner and their interaction. In the baseline experiment of this project, the teacher is an English native speaker and the learner is a Chinese speaker in the early stage. The teacher model includes modeling for pronunciation assessment, learner behaviour and feedback generation. And the learner model includes modeling for acoustic perception, learner knowledge and response generation. Our initial results shows some implication of the simulation for improving the pronunciation learning process.

Index Terms: second language acquisition, teacher-learner interaction, simulation, language learning.

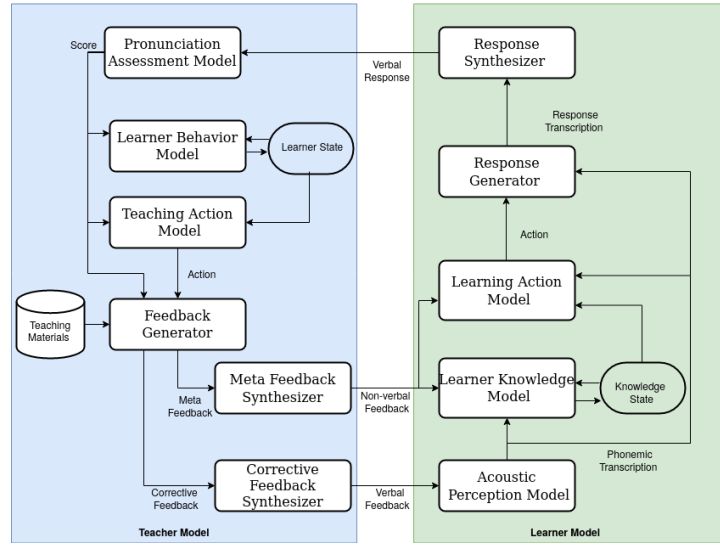


Figure 1: The general proposed teacher-student interaction model.

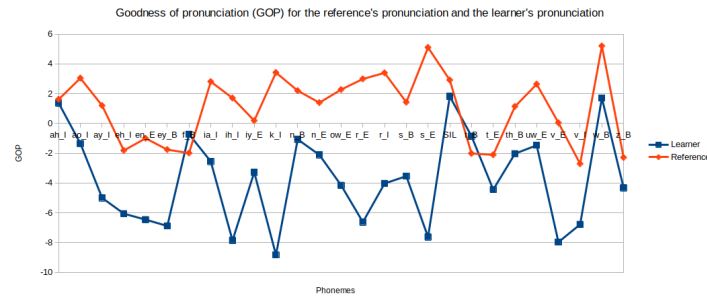


Figure 2: The reference's pronunciation is the average GOP of a native speaker from Audio MNIST dataset. The learner's pronunciation is the average GOP of the pronunciation assessment model.

A New Benchmark Multi-modal Speech Corpus With Two Target Speakers

Adeel Hussain¹, Jasper Kirton-Wingate¹, Kia Dashtipour², Mandar Gogate², Amir Hussain², Peter Derleth³

¹ ²Edinburgh Napier University ³Sonova Research

adeel.hussain, jasper.kirton-wingate, k.dashtipour, m.gogate, a.hussain@napier.ac.uk

1. Abstract

Recent work on personalised speech enhancement aims to learn personal speaker embeddings and differentiate speakers from background noise [1] [2]. A main concern in these studies, as-well as obtaining clean speech signals, is Target Speaker Suppression (TPS), whereby in a conversation the one (or more) target speakers can be suppressed by a SE model especially in a complex acoustic environment. There is significant literature on this in for audio only signals, but work is limited in the Audio-Visual (AV) case [3]. AV cues can provide a huge amount of information that helps us to differentiate target speakers, and foreground from background. However, to the best of our knowledge, a corpus which evaluates Hearing Aid (HA) performance within AV multi-target speaker scenarios from a HA user perspective does not exist.

The majority of existing corpora are either Audio-Only(AO) or AV datasets that are recorded under rigorous conditions with controlled noise levels, reverberation and only addressing situations with a single talker [1]. Therefore, we will collect a novel corpus in conventional challenging circumstances for hearing impaired listeners such as an open-plan office, cocktail party environment, and open-plan dining room. The corpus will be used for the purpose of conducting listening tests to evaluate speech intelligibility and quality in scenarios with two main target speakers, alongside ambient background noise in real-life-like environments known to be problematic for the hearing impaired with current technology, with the intention of improving and advancing the speech enhancement of future multi-modal hearing aids. It is to be noted that, the corpus is appropriate for machine learning as well as various applications in speech and hearing technologies, acoustics, and psychoacoustics.

The sentences chosen to be spoken were taken from British IEEE Harvard sentences, which are phonetically balanced, and include 5 keywords to test recall intelligibility. The data comprises of recordings of 720 sentences. No restrictions were placed on the British English dialect, birthplace, schooling area, or age of the actors who recorded the sentences. Throughout the corpus recording, gender balancing was achieved. Figure 1 illustrates the experimental setup for the recordings. The sentences are employed in a turn-taking scenario and a more challenging sentence overlap scenario. The target speaker can be modulated by the head orientation angle theta, or by other more complicated measures that predict attention modulation such as eye tracking or EEG based measures. These are not provided with the corpus but can be measured in lab via use of a VR headset, for which the corpus is compatible.

We utilise an Ambisonic microphone alongside a HD 360 camera to collect signals for Head Related Transform (HRT) from the HA user position and highly directional shotgun microphones for the 2 actors in order to collect the ground truth clean signals.

Index Terms: Speech enhancement, Machine learning, Intelligibility, British English, Hearing, Audio-Visual

2. References

- [1] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1368–1396, 2021.
- [2] S. E. Eskimez, T. Yoshioka, H. Wang, X. Wang, Z. Chen, and X. Huang, "Personalized speech enhancement: New models and comprehensive evaluation," in *ICASSP*, 2022.
- [3] R. Gu, S.-X. Zhang, Y. Xu, L. Chen, Y. Zou, and D. Yu, "Multi-modal multi-channel target speech separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 530–541, 2020.

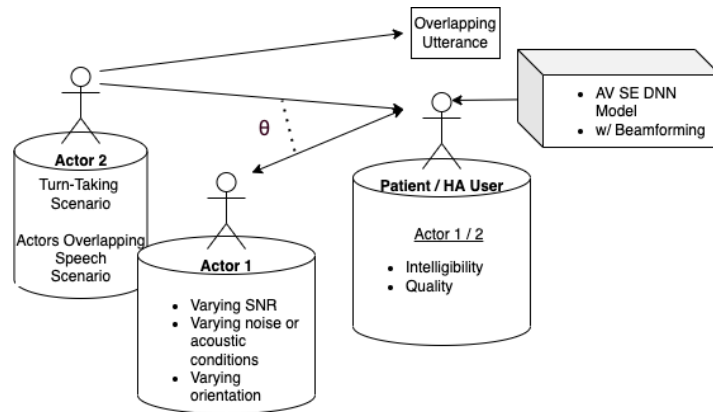


Figure 1: Schematic diagram of multi-modal speech recording

Gender Bias and Universal Substitution Adversarial Attacks on Grammatical Error Correction Systems for Automated Assessment

Vyas Raina, Mark Gales

University of Cambridge

{vr313, mjfg}@cam.ac.uk

Abstract

Grammatical Error Correction (GEC) systems perform a sequence-to-sequence task, where an input word sequence containing grammatical errors, is corrected for these errors by the GEC system to output a grammatically correct word sequence. With the advent of deep learning methods, automated GEC systems have become increasingly popular. For example, GEC systems are often used on speech transcriptions of English learners as a form of assessment and feedback - these powerful GEC systems can be used to automatically measure an aspect of a candidate's fluency. The count of *edits* from a candidate's input sentence (or essay) to a GEC system's grammatically corrected output sentence is indicative of a candidate's language ability, where fewer edits suggest better fluency. The count of edits can thus be viewed as a *fluency score* with zero implying perfect fluency. However, although deep learning based GEC systems are extremely powerful and accurate, they are susceptible to adversarial attacks: an adversary can introduce a small, specific change at the input of a system that causes a large, undesired change at the output. When considering the application of GEC systems to automated language assessment, the aim of an adversary could be to cheat by making a small change to a grammatically incorrect input sentence that conceals the errors from a GEC system, such that no edits are found and the candidate is unjustly awarded a perfect fluency score.

Nevertheless most adversarial attack generation approaches in literature require multiple queries of the target system. However, in the setting of language assessment, a candidate cannot query a GEC system. To overcome this issue, this work uses *universal* adversarial attacks, where the *same* small change has to be made to any input sequence, such that the errors are concealed from the GEC system to obtain a perfect fluency score. As the candidates are non-native speakers of English, it is further required that the form of the attack has to be simple to apply. The simplest such attack is in the form of universal substitutions to exploit potential gender biases in a GEC system. For example, a candidate could replace all *male* pronouns with *female* pronouns, e.g. any occurrence of *he* is replaced with *she*. To determine the extent of threat of this form of adversarial attack, experiments were performed using a popular, publicly available Transformer-based GEC system, the *Gramformer*, when applied to three benchmark GEC datasets (Table 1). The impact

	FCE	BEA	CoNLL
F1 (%)	49.8	45.2	37.1

Table 1: GEC system performance

This work reports on research supported by Cambridge Assessment, University of Cambridge.

of a universal gender pronoun substitution attack is shown in Table 2. For all datasets the GEC system is worryingly biased by the gender, where a candidate can reduce the number of edits made by the GEC system by simply swapping all male gender pronouns with female pronouns (m2f).

Substitution	FCE	BEA	CoNLL	
m2f	-7.2%	-2.8%	-0.5%	↓
f2m	+64.3%	+15.3%	+14.8%	↑

Table 2: Change (%) in Avg. Edits with gender substitution.

The gender pronoun substitution attack can be generalized to a universal substitution attack: a fixed dictionary mapping of word substitutions can be defined for some target words. When a target word appears in an input sequence it is replaced with its corresponding substitution word. For automated assessment with GEC, an adversary can learn and define the optimal dictionary of word mappings that when applied to any input deceives the GEC system into making no edits. The adversary can sell this dictionary to candidates looking to engage in mal-practice - this is a universal substitution attack approach that is agnostic to the original input sequence.

To mimic a realistic setting, the universal substitution dictionary is learnt using only the FCE train set and impact of the adversarial attack is evaluated on other test sets. For computational feasibility, the number of target words has to be limited, as identification of the optimal substitution word demands a greedy search through the English vocabulary. Selection of target words is thus hand-crafted: the most frequent words in the FCE train set, separately for each part of speech (POS), are identified. The universal learnt substituted words are matched in POS with the target words they replace. In this work, target words are restricted to nouns, adjectives or adverbs, e.g. it is found that the target noun *life* should be substituted with the noun *metamorphosis* to reduce number of edits. Table 3 presents the impact of the universal substitution dictionary when applied to the unseen BEA and CoNLL test sets, where the dictionary has only a total of 14 target words (6 nouns, 4 adjectives, 2 adverbs and 3 gender pronouns). Note that results are presented only for the samples that are affected by the substitutions. It is interesting to note that even with such few target words there is a reduction in the number of edits made by the GEC system on unseen test sets.

Data	No Attack	Sub Attack
CoNLL	2.554	2.437
BEA	2.665	2.512

Table 3: Avg. number of GEC edits after Universal attack.

Title

Is there an auditory uncanny valley for synthesised speech?

Authors

Alice Ross, Catherine Lai, Martin Corley

Affiliation

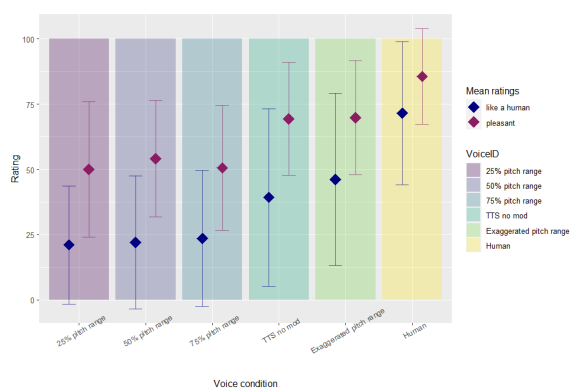
University of Edinburgh

Abstract

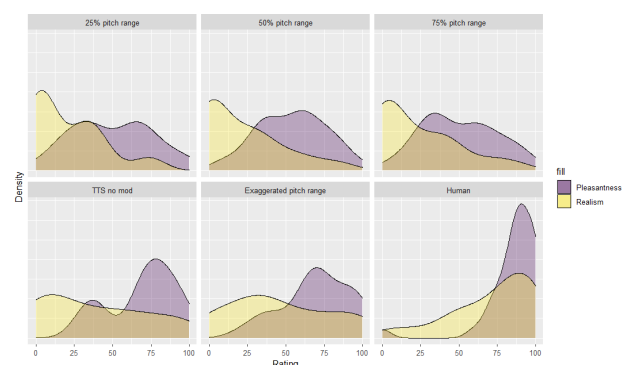
In human-robot interaction, a well-established and productive topic of research has been the exploration of uncanny valley effects (UVE) in visual perception: humans tend to find simplified, unambiguously mechanical robots more appealing than lifelike androids which appear almost, but not quite, human. It is unknown whether uncanny valley effects could be triggered by purely auditory stimuli, for example, realistic text-to-speech (TTS) voices trained on human voice data. As TTS technology has rapidly improved in recent years, such voices are now encountered in a wide range of settings, often with no visible agent representing the 'speaker'.

A between-subjects online experiment investigated 205 listeners' evaluations of an array of manipulated TTS voices all trained on data from a single speaker, with the aim of describing the relationship between the voices' perceived human-likeness and pleasantness. The evidence obtained is compatible with a plateau in a generally positive correlation between realism and approval. The TTS voices used proved insufficiently realistic to draw conclusions about auditory uncanny valley effects, i.e. strong negative reactions to voices perceived as very human-like, because all were rated below 50% on average 'human-likeness'; fruitful possibilities are established for future research on this topic.

Key figures



1. Grouped mean scores for 'human-likeness' and pleasantness of six voice conditions



2. Density plots showing the distribution of 'human-likeness' (yellow) and pleasantness (violet) scores for each of the conditions

Keywords

speech synthesis, TTS evaluation, human-computer interaction, prosody, uncanny valley

Exploration of A Self-Supervised Speech Model: A Study on Emotional Corpora

Yuanchao Li, Yumnah Mohamied, Peter Bell, Catherine Lai

Centre for Speech Technology Research, University of Edinburgh

{yuanchao.li, ymohamie, c.lai, peter.bell}@ed.ac.uk

Abstract

Research on self-supervised speech models has grown fast during the past few years and have proven feasible for use in various downstream tasks. Some recent work has started to look at the characteristics of these models [1, 2]. However, what these models are actually learning is still understudied and questions and concerns remain about why and how these models benefit downstream tasks. For example, *Are the generated representations optimal for every task? How to utilize them for different purposes?* With these questions in mind, in this work, we conduct a study on emotional corpora to explore a popular self-supervised model – wav2vec 2.0 (W2V2). On two emotional corpora: IEMOCAP (IEM) and RAVDESS (RAV), the following experiments are conducted.

Probing SER performance. We first implement a layer-wise analysis by using the output of every individual layer within the transformer network to demonstrate how information encoded by W2V2 contributes to Speech Emotion Recognition (SER). Next, as there is no common practice of how to utilize W2V2 representations as input features for downstream tasks, we compare the performance of three commonly used approaches of using W2V2 representations as input features: 1) taking the last layer output; 2) taking the average of all layer outputs; 3) taking the weighted average of all layer outputs (assigning a trainable weight to each layer output). We also propose a fourth approach which excludes the last two layers from averaging as they generally underperform other layers. We evaluate the performance using Unweighted Accuracy (UA). Like most downstream tasks, we use W2V2 models as frozen feature extractors and build a simple downstream model comprising only two dense layers.

Probing paralinguistic information. In this experiment, we measure the similarities between each layer’s output and different types of paralinguistic features to see how W2V2 retains well-known acoustic correlates of speech perception. We evaluate the similarity using Canonical Correlation Analysis (CCA). The selected paralinguistic features are mainly based on eGeMAPS, commonly used as a minimal set of features for SER. We also extract MFCCs as linguistic (phone) features for comparison.

Probing layer correlation. To better understand how different layer outputs are correlated with each other before and after fine-tuning, and how W2V2 encodes information and contributes to SER, we calculate pair-wise CCA similarities of W2V2 representations from every layer and plot the similarities using heat maps to visualize the correlations. We only discuss IEM, as the same patterns are found on RAV.

Probing hierarchical properties. Since self-supervised learning (SSL) enables frames to capture context information, the representations are expected to contain higher-level meanings. To verify this, we prepare the extracted paralinguistic features

at frame, phoneme, and word levels and measure their similarities with W2V2 representations using CCA, respectively. We use all the paralinguistic features provided by eGeMAPS and implement the composition of hierarchical features.

Probing preference for emotions. Different emotions have different paralinguistic patterns. For example, angry and happy emotions usually have high intensity and pitch, while sad and calm emotions have low intensity. Hence, we calculate CCA similarities between paralinguistic features with W2V2 representations of every emotion for discriminative analysis. We also use all the paralinguistic features in eGeMAPS as in the previous task.

From the results of all the probing experiments*, we present the following major findings:

1) Fine-tuning affects W2V2 by transforming it from an acoustic-aware model into a linguistic-aware model. The layers of the first half of the transformer network are responsible for encoding acoustic-level information, as all three models show almost the same patterns. The latter half starts encoding linguistic information as pattern differences occur, but an exception is that the last two layers of pre-trained model reconstruct the input.

2) W2V2 should be used with caution on downstream tasks because it potentially loses important paralinguistic information. As information that is not helpful to automatic speech recognition is discarded with layer depth on fine-tuned models, pre-trained model is a better choice for tasks that are largely paralinguistic-dependent. Moreover, the best layer outperforms layer averaging for SER, while the last layer is generally the worst choice.

3) While W2V2 (and possibly other similar SSL models) is a universal solution for downstream tasks, it is not suitable for all of them. It does not, for example, outperform previous SER works that take raw waveform as input but use less sophisticated end-to-end structures. Besides, as some paralinguistic information (prosody) is largely involved in pragmatics such as turn-taking and backchanneling, W2V2 may not be able to model these dialog-level functions.

1. References

- [1] A. Pasad, J.-C. Chou, and K. Livescu, “Layer-wise analysis of a self-supervised speech representation model,” in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 914–921.
- [2] Y. Li, P. Bell, and C. Lai, “Fusing asr outputs in joint training for speech emotion recognition,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7362–7366.

*Results are skipped in this abstract but will be shown in presentation. Full paper has been submitted to SLT 2022.

Cross lingual wav2vec finetuning in mutually intelligible language pairs

Jeffrey Josanne Michael, Toby Godwin, Oscar Saz

Emotech Ltd, London, UK

jeffrey@emotech.co, toby@emotech.co, oscar@emotech.co

Abstract

Wav2vec and its multi-lingual variant XLSR have been hugely powerful at representing and at making use of unlabelled speech data. The main motivation of the latter was to overcome the scarcity of available unlabelled data for the target language by using a large number of languages at pre-training. However, monolingual pretraining with well-selected data can still outperform the larger multilingual pretrained models. In this work, we evaluate the effectiveness of using a mutually intelligible language for pretraining, comparing it to multilingual pretraining. In particular, we use Urdu as the target language and use Hindi as the closer language for pretraining. We see that our approach reduces the Word Error Rate (WER) significantly by 18% even though the amount of unlabelled data that was used at pretraining was significantly lower than that of the XLSR or the wav2vec2 model.

1. Introduction

Self-supervised learning has revolutionized speech tasks like Automatic Speech Recognition (ASR) for low or ultra low resourced languages. Wav2vec, HuBERT models act as the extractor of feature and context embeddings for speech data. These models are trained in a self supervised manner to learn suitable representations for unlabelled speech. For ASR tasks, a relatively smaller amount of labelled data is then used to finetune to use for the downstream recognition task. However, domain mismatch is a recognized problem with these models. We don't usually see competitive results when the domain used at the downstream finetuning process varies from the domain used at pretraining (telephony vs VOIP audios, conversational vs read speech). As far as the pretraining data is closer to the finetuning data, the models perform significantly better.

1.1. Mutually intelligible languages

Mutual intelligibility is a phenomenon where the speakers of different languages can understand each other without any prior knowledge of the other language. There are three main kinds of mutually intelligibility in languages: mutually intelligible in scripts; mutually intelligible in spoken form; or both script and spoken. This paper focus on languages mutually intelligible in the spoken form. Namely examples of such language pairs are Hindi - Urdu [1], Mandarin - Dungan, and Ukrainian - Belarusian. In this work, we consider Urdu and Hindi as the mutually intelligible pair. In our case, the latter has far more unlabelled/labelled data compared to the former.

2. Experimental setup

In this work, we compare the XLSR model from [2] against the model that was pre-trained on 4200 hours of Hindi data [3].

We use Urdu to finetune both the aforementioned models using 4 hours of CommonVoice Urdu data and 84 hours of the

ARL Urdu Speech Database. For evaluation, we use Common-Voice's Urdu test set.

In this work, we use Urdu as the target language to finetune on the task of speech recognition. Wav2vec architecture provides the feature representation - including the contextual representation. The representation that is produced is very vital for the downstream tasks. The representation is highly dependent on the unlabelled pre-training data, hence the closer the pretraining data is to the finetuning data, the better the performance is. As discussed, Urdu is mutually intelligible with Hindi in its spoken form. We use a Hindi pretrained model, which has significantly higher data resources, to finetune Urdu on. The difference here is to use a CTC layer with Urdu characters instead of Hindi characters since the target language is Urdu.

As a comparison for the proposed approach, we also transliterate (using indic-trans [4]) a similar sized Hindi data to Urdu graphemes. Like the proposed method, we use the same CTC layer with Urdu graphemes to finetune. Unfortunately, this technique did not help in increasing the performance since there are about 10% of Urdu characters that do not have a 1-to-1 mapping to Hindi script.

Pretrained model	Finetuning	WER
XLSR	Urdu	33.2%
XLSR	Hindi transliterated	34.9%
Pretrained on Hindi	Urdu	27.1%

3. Conclusion

As we show, for low resourced, mutually intelligible languages pairs the high resourced language can be used as the pretraining language while the low resourced language of the pair can be used for finetuning to get better performance compared to using a pretrained model with a large variance of languages.

4. References

- [1] Muhammad Suffian Nizami, Tafseer Ahmed and Muhammad Yaseen Khan, "Hindustani or Hindi vs. Urdu: A Computational Approach for the Exploration of Similarities Under Phonetic Aspects" *International Journal of Advanced Computer Science and Applications*, 2020. <http://dx.doi.org/10.14569/IJACSA.2020.0111191>
- [2] Conneau, Alexis and Baevski, Alexei and Collobert, Ronan and Mohamed, Abdelrahman and Auli, Michael, "Unsupervised Cross-lingual Representation Learning for Speech Recognition", *Inter-speech*, 2021
- [3] Harveen Singh Chadha and Anirudh Gupta and Priyanshi Shah and Neeraj Chhimwal and Ankur Dhuriya and Rishabh Gaur and Vivek Raghavan, "Vakyansh: ASR Toolkit for Low Resource Indic languages", arXiv, arXiv:2203.16512, 2022.
- [4] Bhat, Irshad Ahmad and Mujadia, Vandan and Tammewar, Anirudha and Bhat, Riyaz Ahmad and Shrivastava, Manish, "IIIT-H System Submission for FIRE2014 Shared Task on Transliterated Search" *Proceedings of the Forum for Information Retrieval Evaluation*, pp.48-53, ACM, 2015

Segmental evaluation of Text-to-Speech Synthesizers

Ayushi Pandey¹, Sébastien Le Maguer¹, Julie Carson-Berndsen², Naomi Harte¹

¹Sigmedia Lab, ADAPT Centre, School of Engineering, Trinity College Dublin, Ireland

²ADAPT Centre, School of Computer Science, University College Dublin, Ireland

pandeya@tcd.ie, lemagues@tcd.ie, julie.berndsen@ucd.ie, nharte@tcd.ie

Abstract

Segmental properties of Text-To-Speech (TTS) synthesizers have been studied for their influence on various perceived attributes of synthesized speech. However, they have received very limited attention for modern, neural vocoder-based TTS. In this paper, we compare segmental properties of WaveNet vocoder voices with a natural voice, and the best-performing non-neural synthesizers of the 2013 Blizzard Challenge [1]. We extended the 2013 dataset with two new voices generated using a WaveNet vocoder [2]. Acoustic-phonetic features of obstruent consonants and their neighbouring vowels were compared between the natural voice and each of these TTS systems. Statistical analysis was conducted using the Kruskal-Wallis test, and Dunn's test.

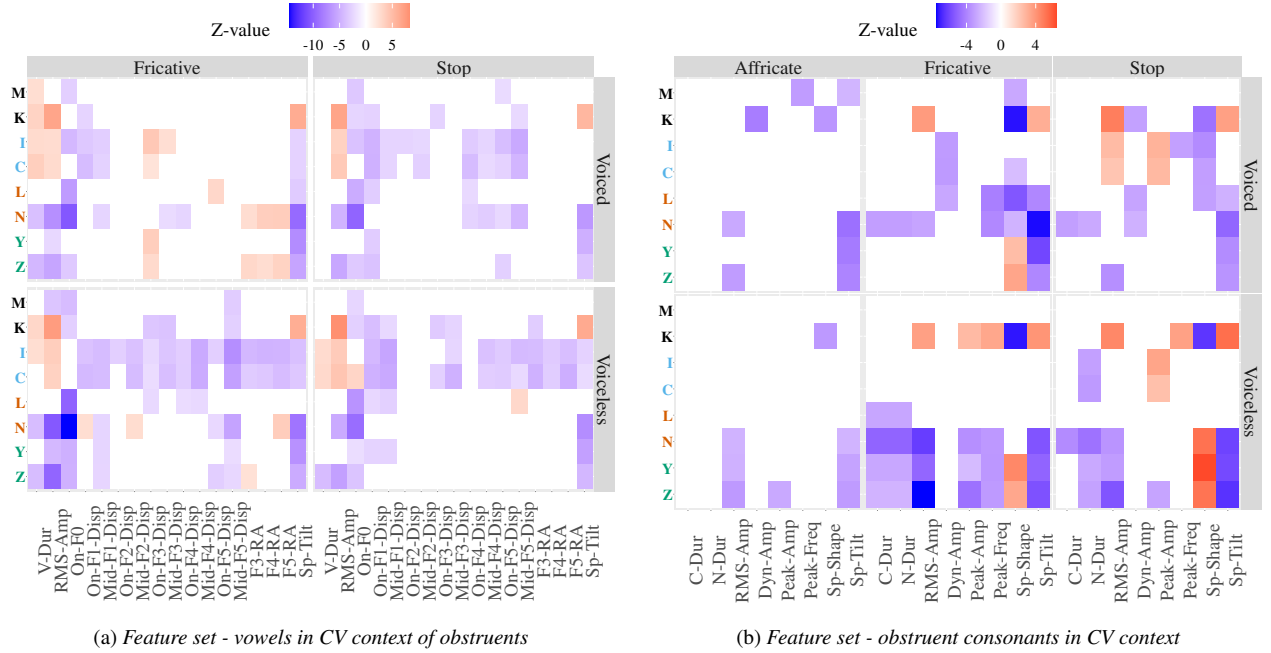


Figure 1: Deviation from natural compared across TTS techniques using Dunn's z-test statistic. White cells indicate no significant difference ($p\text{-val} > 0.05$). TTS Techniques: **M K** - Hybrid, **I C** - HMM, **L N** - Unit-Selection, **Y Z** - WaveNET

Figure 1 shows that, when compared to the reference natural voice, WaveNet vocoder performs very well in modelling vowels, but features like F0 at onset and spectral tilt show significant deviations from the natural voice. Compared to other TTS systems, several features (like vowel dispersions, and consonant duration) which had shown strong deviations from natural, were found to not differ from natural in the WaveNet vocoder systems. Among consonants, neural voices deviate most from natural in the context of voiceless fricatives. This indicates that neural voices can model segments with a periodic structure better than noise regions. Carefully designed listener tests are needed to firmly establish a relationship between these features of obstruents and perceived attributes of synthetic speech. In the long-term, we envisage that the segmental analysis presented in this paper can be used to detect those system-specific weaknesses that may not be diagnosed by subjective evaluations alone. Additionally, using controllable neural TTS architectures like Wavebender [3], specific locations of distortion can be improved.

Index Terms: WaveNet, obstruents, TTS evaluation

1. References

- [1] S. King and V. Karaiskos, "The blizzard challenge 2013," in *The Blizzard Challenge Workshop*, 2013, <http://festvox.org/blizzard/bc2013/summary-Blizzard2013.pdf>.
- [2] S. Le Maguer, S. King, and N. Harte, "Back to the future: Extending the blizzard challenge 2013," in *InterSpeech (To appear)*, 2022.
- [3] G. T. D. Beck, U. Wennberg, Z. Malisz, and G. E. Henter, "Wavebender gan: An architecture for phonetically meaningful speech manipulation," *arXiv preprint arXiv:2202.10973*, 2022.

Phonetically Guided Transfer Learning for Low-Resource Accented English

Edward Storey, Naomi Harte

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

storeyedl@tcd.ie, nharte@tcd.ie

Automatic Speech Recognition (ASR) deep learning models have reported Word Error Rates (WERs) as low as 2.2% in the English language [1]. It is well documented however, that only high-resource accents such as Standard American (SA) achieve error rates this low, while low-resource accents do not [2] [3]. One technique that has proven valuable in improving error rates for low-resource data is transfer learning. Transfer learning is a technique utilising pre-trained networks and further training them on new data. Transfer learning has been used in ASR to fine-tune on accented speech and transfer models to other languages [4].

High-resource ASR datasets bias heavily towards first language (L1) accented English, most commonly SA, this limits deep learning models to phonemes found only in these limited accents. We propose to tackle this issue by training a high performance model in Mandarin to introduce it to non L1 English phonemes. Then, using transfer learning, we will re-purpose the model into English. This experiment will show that using linguistic knowledge of low-resource Chinese accented data can help to improve recognition.

Table 1: WERs for 8 of the countries in the Datatang Accented English Dataset (DAE) [3] when testing the state-of-the-art Quartznet15x5 model [1] trained on the LibriSpeech360 (LB360) [5] and separately the AISHELL-2 (AS2) [6] dataset. LB360 shows results from before and after transfer learning on DAE while AS2 is only post.

	Datasets	Accents								
		US	UK	CHN	IND	JPN	KOR	PT	RU	Avg
WER	LB360	9.6%	12.2%	15.9	13.4%	24.2%	18.5%	11%	13.1%	14.2%
	LB360 & DAE	4.8%	7.2%	9.8%	8.9%	10.2%	9%	6.3%	6.8%	7.7%
	AS2 & DAE	15.5%	25.2%	23.2%	21.83%	17.47%	15.5%	16.9%	15.2%	18.9%

The “LB360” row in Table 1 shows that when trained on the primarily SA accented LibriSpeech360 dataset (LB360) L1 speakers is the US have the lowest WER. Non-L1 English speaking countries such as China and Japan have much higher WERs by comparison. Japan having 24.2% WER compared to the 9.6% WER on speakers from the USA. The difference between the highest and lowest WER has been reduced on the “LB360 & DAE” row. Here transfer learning has been applied with the Datatang accented English dataset (DAE). However, we still see Japan and China having the highest WERs and so transfer learning has not solved this imbalance. For the “AS2 & DAE” row, the model has been trained on the AISHELL-2 dataset (AS2), a Mandarin language dataset. Transfer learning was then applied with DAE. Unlike with LB360, Chinese accents have a lower WER at 23.2% than the L1 British English speakers at 25.2%. Japanese and Korean speakers also have low relative WERs, with Japanese at 17.47% and both US and Korean speakers having an equal WER of 15.5%.

Mandarin has a different phonetic alphabet to English and is more closely related to Japanese and Korean languages. The difference in phonemes between Mandarin and English may explain the overall higher WER of the AS2 trained model. While the relation between phonemes found in Mandarin, Japanese and Korean explains the low relative WER of these accents on the AS2 model. These preliminary results show that having phonetic and linguistic knowledge of an accent can improve performance on low-resource data.

1. References

- [1] S. Kriman, S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, and Y. Zhang, “Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6124–6128.
- [2] C. Huang, T. Chen, S. Li, E. Chang, and J. Zhou, “Analysis of speaker variability,” in *Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, 2001, pp. 1377–1380.
- [3] X. Shi, F. Yu, Y. Lu, Y. Liang, Q. Feng, D. Wang, Y. Qian, and L. Xie, “The accented english speech recognition challenge 2020: Open datasets, tracks, baselines, results and methods,” in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6918–6922.
- [4] J. Luo, J. Wang, N. Cheng, E. Xiao, J. Xiao, G. Kucsko, P. O’Neill, J. Balam, S. Deng, A. Flores, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, and J. Li, “Cross-language transfer learning and domain adaptation for end-to-end automatic speech recognition,” in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 2021, pp. 1–6.
- [5] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An asr corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.
- [6] J. Du, X. Na, X. Liu, and H. Bu, “Aishell-2: Transforming mandarin asr research into industrial scale,” 2018. [Online]. Available: <https://arxiv.org/abs/1808.10583>

This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224 and ADAPT SFI Research Centre under Grant No. 13/RC/2106.P2. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission

Dysarthric Speech Recognition From Raw Waveform with Parametric CNNs

Zhengjun Yue^{1,2,†}, Erfan Loweimi^{1,3,†}, Heidi Christensen², Jon Barker², Zoran Cvetkovic¹

¹ Department of Engineering, King’s College London, UK

² Speech and Hearing Research Group (SPandH), University of Sheffield, UK

³ Centre for Speech Technology Research (CSTR), University of Edinburgh, UK

{zhengjun.yue, erfan.loweimi, zoran.cvetkovic}@kcl.ac.uk,

{heidi.christensen, j.p.barker}@sheffield.ac.uk

Abstract

Raw waveform acoustic modelling has recently received increasing attention. Compared with the task-blind hand-crafted features which may discard useful information, representations directly learned from the raw waveform are task-specific and potentially include all task-relevant information. In the context of automatic dysarthric speech recognition (ADSR), raw waveform acoustic modelling is under-explored owing to data scarcity. Parametric convolutional neural networks (CNNs) can compensate for this problem due to having notably fewer parameters and requiring less training data in comparison with conventional non-parametric CNNs.

In this paper, we explore the usefulness of raw waveform acoustic modelling using various parametric CNNs for ADSR. We investigate the properties of the learned filters and monitor the training dynamics of various models. Furthermore, we study the effectiveness of data augmentation and multi-stream acoustic modelling through combining the non-parametric and parametric CNNs fed by hand-crafted and raw waveform features. Experimental results on the TORGO dysarthric database show that the parametric CNNs significantly outperform the non-parametric CNNs, reaching up to 36.2% and 12.6% WERs (up to 3.4% and 1.1% absolute error reduction) for dysarthric and typical speech, respectively. Multi-stream acoustic modelling further improves the performance resulting in up to 33.2% and 10.3% WERs for dysarthric and typical speech, respectively.

Index Terms: dysarthric automatic speech recognition, raw waveform acoustic modelling, parametric CNNs

[†] Equal contribution.

ZY, EL and ZC are supported by EPSRC Project EP/R012180/1 (SpeechWave). ZY, HC and JB are supported by the European Union’s H2020 Marie Skłodowska-Curie programme TAPAS (Grant agreement No. 766287).

Joint Modelling of Automatic Speaker Verification and Spoofing Countermeasure Systems

Poppy Welch, Jennifer Williams

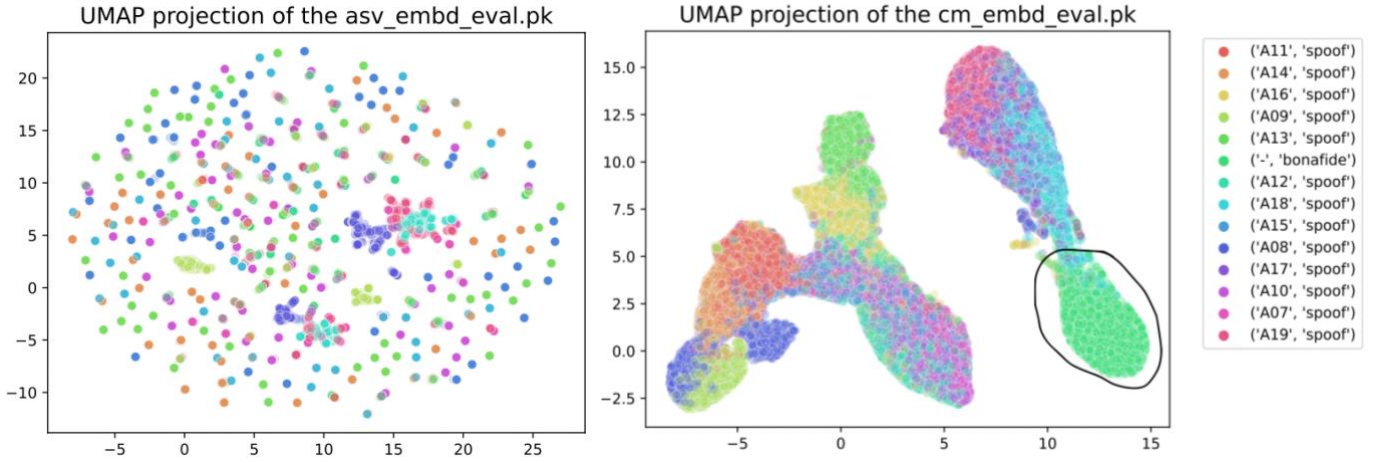


Fig 1. UMAP Visualisation of speaker embeddings extracted from an automatic speaker verification system (left) and a spoofing countermeasures system (right).

Embeddings obtained from evaluation data with spoofed and bona fide data (circled in image)

The use of automatic speaker verification (ASV) systems for the rejection or acceptance of a claimed identity of a speaker has become increasingly widespread in security authentication and access control scenarios. As speech technology has advanced, these systems have become more vulnerable to spoofing attacks, resulting in a critical need for the development of suitable protection. Countermeasures (CM) systems can be implemented for the detection of these spoofing attacks.

The ASV and CM tasks are largely treated as two separate issues, with CM models being developed as standalone systems that operate with a fixed ASV system. It has been argued that optimizing both the ASV and CM systems jointly may improve the reliability of an overall integrated system. This type of model is described as a spoofing-aware speaker verification system (SASV).

Part of the difficulty in constructing a SASV model is that the ASV and CM subsystems each have contradictory objective functions, despite their overall aim to reject any imposters and accept the target speakers. The ASV subsystem is required to ignore variable conditions such as the background environment and channel in order to robustly identify speakers, whereas the CM subsystem identifies spoofed utterances using artefacts from these same conditions.

This proposed work aims to start investigating the training and the output of a SASV model. Firstly, as the ASV system is of primary importance, we aim to investigate what data should be used when training an ASV subsystem in the context of a jointly optimized integrated model. In addition to this, we investigate how the speaker embeddings extracted from a SASV system differ from those extracted from separate ASV and CM systems. We aim to address this by using a SASV model with a weighted joint loss function for simultaneous learning for the ASV and CM tasks.

Oral Session B

The 2nd Clarity Enhancement Challenge: A machine learning challenge for hearing aid speech intelligibility enhancement

Will Bailey¹, Michael A. Akeroyd², Jon Barker¹, Trevor J. Cox³, John F. Culling⁴, Simone Graetzer³, Graham Naylor², Zuzanna Podwińska³, Zehai Tu¹

¹ Department of Computer Science, University of Sheffield, UK

² School of Medicine, University of Nottingham, UK

³ Acoustics Research Centre, University of Salford, UK

⁴ School of Psychology, Cardiff University, UK

claritychallengecontact@gmail.com

1. Outline

This work outlines the design of the second Clarity Enhancement Challenge (CEC2) for enhancing speech signals to improve intelligibility for listeners wearing hearing aids. The challenge was designed to promote the development of new enhancement algorithms for use in hearing aid signal processing chains. The challenge was launched in April 2022, with submission closing on the 1st September 2022. Results are set to be announced in December 2022. CEC2 builds on previous Clarity challenges, namely CEC1 and the Clarity Prediction Challenge (CPC1).[1]

2. Dataset design

Participants are supplied with three datasets, comprising 6000 dynamic training scenes, 2500 dynamic development scenes and a further 1500 dynamic scenes that are held out from participants until the evaluation phase, giving a total of 10,000 unique scenes in the complete dataset. The scenes are spatially rendered within simple shoe-box simulated rooms using higher order ambisonics. These consist of a single target speaker and up to three interfering sources spatially distributed around the virtual room. The sources are drawn from an open source speech corpus of 40 speakers.[2] Whereas in CEC1 scenes were composed of a target speaker and a single interferer that may be either competing speech or domestic noise, interferers in CEC2 are drawn from corpora of domestic noise, speech and music. The scenes are dynamic and emulate a listener looking away from the target speaker and turning to face them (approximately) after the target speaker has begun speaking. A further increase in difficulty was added by including variation in the timing of the onset of the target speakers. The training, development and evaluation datasets were produced with a range of signal-to-noise ratios between -12 dB and 6 dB in order to provide a wide range of difficulty for participants.

3. Challenge rules

Participants are asked to submit processed audio files. There are no constraints on whether the system is intrusive or non-intrusive in design or on computational complexity. However, all systems must be causal, requiring no more than 5ms lookahead. Systems will be evaluated using the HASPI prediction algorithm[3] modified with a ‘better ear’ stage to accommodate stereo or binaural inputs (note that the previous Clarity challenges used MBSTOI[4] to benchmark intelligibility prediction).

4. Outputs, dissemination and further work

CEC2 is part of the second round of Clarity challenges, which are designed to encourage innovation in the enhancement of audio for, and the prediction of, speech intelligibility for hearing aids. The project aims to encourage active research on signal processing, particularly using deep learning approaches. It also aims to provide a growing repository of tools for the creation of datasets and baseline systems and also for testing and evaluating models and algorithms, particularly those with applications in hearing aid development.

5. References

- [1] S. Graetzer, J. Barker, T. J. Cox, M. A. Akeroyd, J. F. Culling, G. Naylor, E. Porter, and R. Viveros-Muñoz, “Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing,” in *Interspeech*, 2021.
- [2] S. Graetzer, M. A. Akeroyd, J. Barker, T. J. Cox, J. F. Culling, G. Naylor, E. Porter, and R. Viveros-Muñoz, “Dataset of British English speech recordings for psychoacoustics and speech processing research: The Clarity speech corpus,” *Data in Brief*, vol. 41, April 2022. [Online]. Available: <https://eprints.whiterose.ac.uk/184415/>
- [3] J. M. Kates and K. H. Arehart, “The hearing-aid speech perception index (HASPI),” *Speech Communication*, vol. 65, pp. 75–93, 2014.
- [4] A. H. Andersen, J. M. de Haan, Z. H. Tan, and J. Jensen, “Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions,” *Speech Communication*, vol. 102, pp. 1–13, 2018.

Back to the Future: Extending the Blizzard Challenge 2013

Sébastien Le Maguer¹, Simon King², Naomi Harte¹

¹Sigmedia Lab, ADAPT Centre, School of Engineering, Trinity College Dublin, Ireland

²The Centre for Speech Technology Research, University of Edinburgh, UK

lemagues@tcd.ie, Simon.King@ed.ac.uk, nharte@tcd.ie

Nowadays, speech synthesis technology is synonymous with the use of Deep Learning. To understand what Deep Learning has solved, we need to keep a connection between past and present technologies. However, we do not have a dataset that allows this comparison on common ground.

In this paper, we propose an extension of the 2013 edition[1] of the Blizzard Challenge in which we compare top-tier systems from the past to modern technologies in a controlled setting. We selected the task 2013-EH2 which is a sentence-level segmented corpus of about 20 h of speech with corresponding transcriptions. The speaker is a female American professional narrator and actor. From this edition, we selected the best representative of each historical synthesis technology: System K representing the Hybrid family, system N representing the Unit-Selection family and system C representing the Parametrical HMM-Based family.

We added four systems representing combinations of modern acoustic models and neural vocoders. For the acoustic models, we used Tacotron[2] and FastPitch[3]. The neural vocoders we selected are WaveNet[4] and Parallel WaveGAN[5].

A large scale subjective evaluation was conducted to evaluate naturalness and the results are presented in Figure 1 and Table 1.

Our results show that, as expected, modern technologies generate more natural synthetic speech. However, these systems are still not perceived to be as natural as the human voice. Crucially, we also observed that the Mean Opinion Score (MOS) of the historical systems dropped a full MOS point from their scores in the original edition. This demonstrates the relative nature of MOS: it should generally not be reported as an absolute value despite its origin as an Absolute Category Rating.

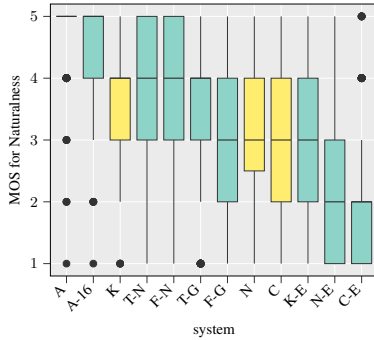


Figure 1: Listening test results. The yellow ones are those obtained in the 2013 challenge and the green ones with the suffix “-E” are the new results from our listening test. The median is represented by a solid bar across a box showing the quartiles; whiskers extend to 1.5 times the inter-quartile range and outliers are represented as circles. The systems are ordered using the mean MOS.

	A-16	T-N	F-N	T-G	F-G	K	N	C
T-N	■	NA						
F-N	■		NA					
T-G	■			NA				
F-G	■	■			NA			
K	■	■	■	■	■	NA		
N	■	■	■	■	■	■	NA	
C	■	■	■	■	■	■		NA

Table 1: Significance results of a pairwise Wilcoxon signed rank test with Bonferroni correction for our listening test. Each cell marked with ■ indicates that the two systems are considered different with a p-value < 0.01 for both novel and news sentences; with ■ only for news sentences; with ■ only for novel sentences. We do not have any samples of news sentences for the natural voice, so the column A-16 has only been computed using novel sentences.

1. References

- [1] S. King and V. Karaiskos, “The blizzard challenge 2013,” in *The Blizzard Challenge Workshop*, 2013, <http://festvox.org/blizzard/bc2013/summary-Blizzard2013.pdf>.
- [2] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, “Tacotron: A fully end-to-end text-to-speech synthesis model,” *CoRR*, 2017. [Online]. Available: <http://arxiv.org/abs/1703.10135>
- [3] A. Łańcucki, “Fastpitch: Parallel Text-to-Speech with Pitch Prediction,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Jun. 2021, pp. 6588–6592.
- [4] A. Van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “WaveNet: A generative model for raw audio,” *CoRR*, vol. abs/1609.03499, 2016. [Online]. Available: <http://arxiv.org/abs/1609.03499>
- [5] R. Yamamoto, E. Song, and J.-M. Kim, “Parallel Wavegan: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, May 2020, pp. 6199–6203.

This work has been accepted at Interspeech 2022

Fine Grained Spoken Document Summarization Through Text Segmentation

Samantha Kotey¹, Rozenn Dahyot² & Naomi Harte¹

¹ADAPT Centre, School of Engineering, Trinity College Dublin, Ireland

²ADAPT Centre, Department of Computer Science, Maynooth University, Ireland

koteys@tcd.ie, Rozenn.Dahyot@mu.ie, nharte@tcd.ie

Abstract

Podcast transcripts are long spoken documents of conversational dialogue. Challenging to summarize, podcasts cover a diverse range of topics, vary in length, and have uniquely different linguistic styles. There is a growing need to develop automated tools for dialogue summarization and information access [1]. Podcast dialogue can be hours long, and summaries help listeners to choose an episode from numerous shows. Previous studies in podcast summarization have generated short, concise dialogue summaries. However, conversations can contain multiple story lines, which are difficult to capture in a short summary. Long summaries provide users with an opportunity to consume more interesting content, which may otherwise be missed.

In this work, we propose a method to generate different types of long fine-grained summaries, which are fluent, coherent and describe sub-topic narratives. Our proposed pipeline approach involves three stages, as illustrated in Figure 1.

In **stage 1**, we leverage a readability formula and curate a selection of podcast transcripts from the Spotify podcasts dataset [2], forming the training data. Through readability score metrics, we analyze the relationship between lexical qualities and description length.

In **stage 2**, we process the evaluation data, by segmenting long text transcripts into smaller segment chunks. Each episode is divided and numbered into segments, and each segment is defined by a change in topic conversation. Through text segmentation, we form a filtered down version of each transcript, with a reduced input size.

Finally, in **stage 3**, we fine-tune a long sequence transformer on the curated dataset and apply the model to the filtered evaluation data.

We show that appropriate filtering creates comparable results on ROUGE and serves as an alternative method to truncation. Experiments show our model outperforms previous studies on the Spotify podcast dataset when tasked with generating longer sequences of text.

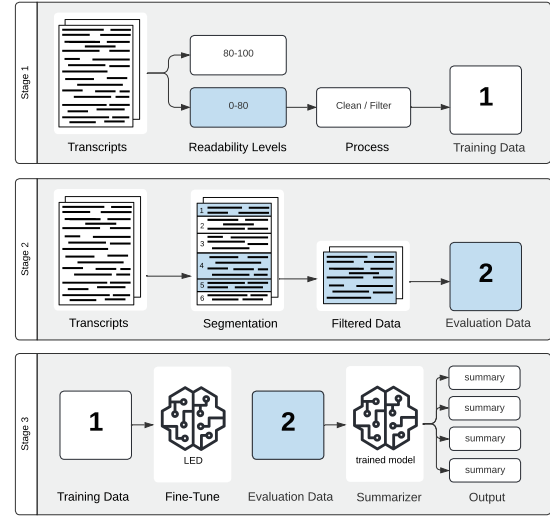


Figure 1: Overview of our approach to generate different summary types through text segmentation.

Index Terms: Spoken document summarization, text segmentation, long sequence transformers, readability formulas.

Acknowledgments: This research was conducted with the financial support of Irish Research Council (IRC) under Grant Agreement GOIPG/2019/2353 and the ADAPT SFI Research Centre under Grant No. 13/RC/2106_P2.

1. References

- [1] R. Jones, H. Zamani, M. Schedl, C.-W. Chen, S. Reddy, A. Clifton, J. Karlgren, H. Hashemi, A. Pappu, Z. Nazari, L. Yang, O. Semerci, H. Bouchard, and B. Carterette, “Current challenges and future directions in podcast information access,” in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021.
- [2] A. Clifton, S. Reddy, Y. Yu, A. Pappu, R. Rezapour, H. Bonab, M. Eskevich, G. Jones, J. Karlgren, B. Carterette, and R. Jones, “100,000 podcasts: A spoken english document corpus,” in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 5903–5917.