

UKIS 2024

Cambridge

Abstract Book

UK Speech

UNIVERSITY OF CAMBRIDGE

General information

MAPS & DIRECTIONS

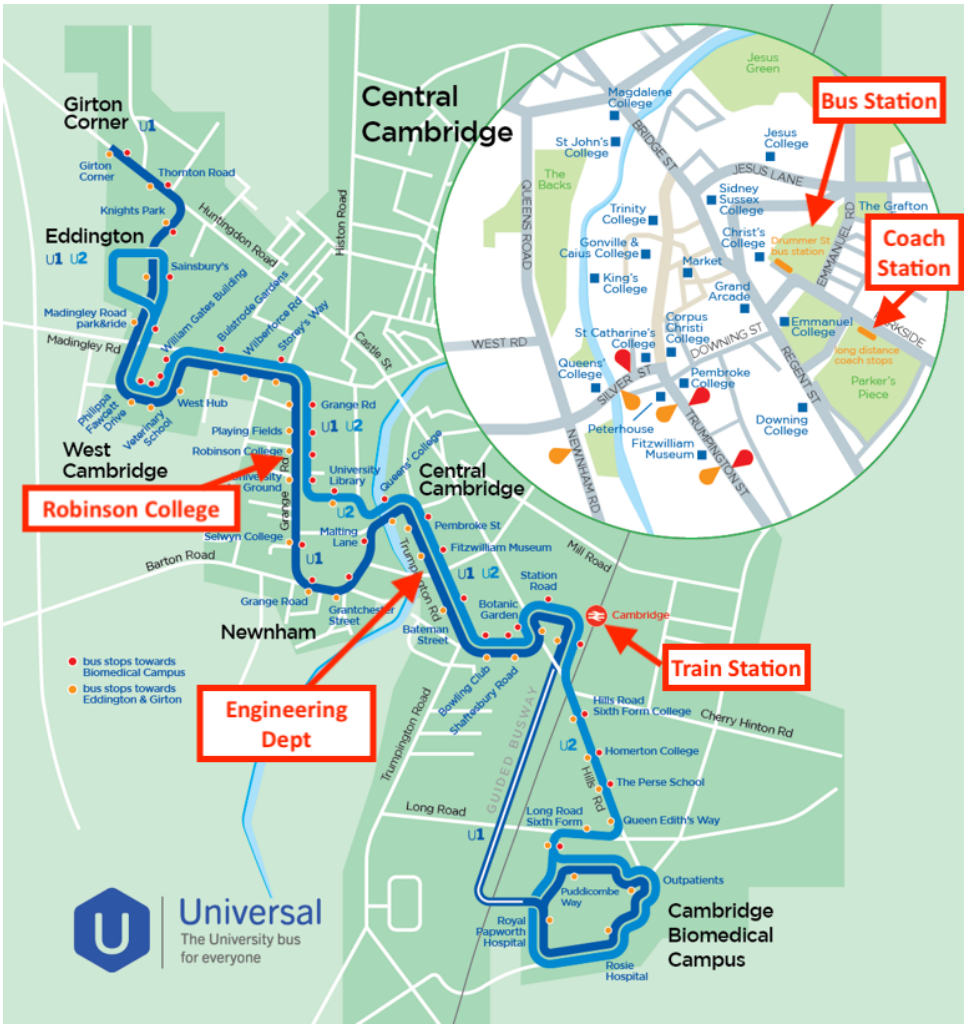
Conference venue: University of Cambridge Engineering Department, Trumpington Street, Cambridge CB2 1PZ



Social event venue: Robinson College, Grange Road, Cambridge CB3 9AN



MAPS & DIRECTIONS



There is no car parking available at the Department but there is lots of bicycle parking.

Time to Department by foot:

- Train Station: 20-25 minutes
- Robinson College: 25-30 minutes
- National Express Coach Station: 15-20 minutes

Taxis:

- Panther taxis: bookable on veezu app or call 01223 715715
- Uber

Voi bikes and e-scooters are available to hire at various points around town

SOCIAL EVENT

On Monday 1st, there will be a drinks reception, dinner and music at Robinson College.



Robinson is the University of Cambridge's newest College. It was founded with a donation by the entrepreneur and philanthropist Sir David Robinson and formally opened in 1981. It now has 100 Fellows, and around 380 undergraduate and 300 graduate students. Its distinctive and Grade 2* listed redbrick building offers a modern environment for living and working.



- 18:30-19:30 **Drinks Reception**
- 19:30 **Dinner sponsored by Google**
- ~21:00-22:30 **Music from A Twist of Melon, (optional) dancing, cash (card) bar**
- 23:00 **Finish**

Dress code: informal conference wear

WIFI

eduroam – available for academic attendees

UniOfCam-Guest wifi

- get free wireless connectivity where the University Wireless Service is available
- sign in using email, Amazon, Facebook, GitHub or LinkedIn
- connect in outdoor public spaces, such as Jesus Green, Parkers Piece and The Mill Pond
- connect in University of Cambridge museums, including Fitzwilliam Museum, Kettles Yard and Scott Polar Research Institute
- connect up to 3 devices each

Connect to UniOfCam-Guest

1. Open the wifi settings on your device. A list of wifi networks will appear.
2. Select the 'UniOfCam-Guest' wifi network from your list of available networks. You'll see the login screen appear on your device.

Try opening a web browser if you do not see a login screen. QR code for support page https://bit.ly/45PZjzN

Organisers



General Chair: Kate Knill
Technical Committee: Keqi Deng, Brian Sun
Keynotes and Talks: Erfan Loweimi, Simon McKnight, Hari Krishna Vydana
Sponsorship: Charles McGhee, Mengjie Qian
Social event and lunches: Rao Ma, Siyuan Tang
Website and communications: Vyas Raina
Local organisation: Adian Liusie, Vatsal Raina, Wen Wu
Administrative support: Kimberley Cole, Toby Elles
Photographer: Howard Mei
General help: Mark Gales, Jonathan Jiang, Xiaodong Wu, Tony Zheng



Simone Graetzer, University of Salford
Catherine Lai, University of Edinburgh
Sébastien Le Maguer, University of Helsinki
Thomas Merritt, Spotify

Gold Sponsor

Google

Silver Sponsor

keen research

Bronze Sponsors

Meta

Other Sponsors

CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

Captioning Partner

SPEECHMATICS

Short Programme

Monday, 1st July 2024

Time Slot	Session Name
12h00-13h00	Registration & Lunch
13h00-13h20	Welcome Message (Kate Knill)
13h20-14h20	**Keynote A (Dr. Catherine Lai)** - Chair: Mengjie Qian **Across the prosodic dimension: Exploring spoken communication beyond text**
14h30-15h30	Poster Session A
15h30-16h00	Coffee Break
16h00-17h00	Oral Session A - Chair: Erfan Loweimi
18h30-22h00	Social Event including Banquet Dinner in Association with Google

Tuesday, 2nd July 2024

Time Slot	Session Name
08h30-09h00	Registration
09h00-10h00	**Keynote B (Prof. Elizabeth Stokoe)** - Chair: Stefano Bannò **How 'conversational' are conversational products and technologies?**
10h00-11h00	Poster Session B
11h00-11h30	Coffee Break
11h30-12h30	Oral Session B - Chair: Brian Sun
12h30-13h30	Lunch
13h30-14h30	Poster Session C
14:30-15h30	**Keynote C (Prof. Jon Barker)** - Chair: Simon McKnight **Using Machine Learning to Improve Hearing Aid Signal Processing: The Clarity and Cadenza Challenges**
15h30-16h15	UKIS 2024 Cambridge: Future Plans & Farewell

Long Programme

Monday, 1st July 2024

12h00-13h00 - Registration & Lunch

13h00-13h20 - Welcome Message (Kate Knill)

13h20-14h20 - Keynote A (chair: Mengjie Qian)

Across the prosodic dimension: Exploring spoken communication beyond text
Dr. Catherine Lai

14h30-15h30 - Poster Session A

Label-Synchronous Neural Transducer for End-to-End ASR
Keqi Deng; Phil Woodland

Impact of Investigator Speech and its removal for Alzheimer's Dementia Detection
Marek Sviderski; Basel Barakat; Becky Allen

Imbalanced Multimodal Learning in Video Conversations
Samantha Kotey; Naomi Harte

Utilising Unsupervised Text-to-Speech Synthesis for Data Augmentation to Improve Accented Speech Recognition
Cong-Thanh Do; Shuhei Imai; Rama Doddipatla; Thomas Hain

The effect of feeding and non-nutritive sucking on speech sound development at ages 2 to 5 years
Sam Burr

Exploring speech representations for proficiency assessment in language learning
Elaf Islam; Chanho Park; Thomas Hain

Affinity: Dialogs to build human-robot rapport
Alina Larson; Matthew Aylett

Phonetics and Phonology inside the Black Box
Iona Gessinger; Erfan A. Shams; Julie Carson-Berndsen

Towards explainable speaker recognition neural networks
Yanze Xu

Can political science help us improve speech synthesis evaluation?
Sébastien Le Maguer; Juraj Šimko

SCORE: Self-supervised Correspondence Fine-tuning for Improved Content Representations
Amit Meghanani; Thomas Hain

Exploring Dominant Paths in CTC-like ASR Models: Unraveling the Effectiveness of Viterbi Decoding
Zeyu Zhao; Peter Bell; Ondřej Klejch

Multilingual Integration in Lyrics Transcription: Data, Language Conditioning, and Transliteration Augmentation
Jiawen Huang; Emmanouil Benetos

Real-Time Spoken Language Processing in Conversational AI: Challenges and Future Directions
Arash Ashrafzadeh; Arash Eshghi; Matthew Aylett

Self-supervised models for dysarthric speech: Understanding representations through visual analysis and probing
Ariadna Sanchez; Simon King

How Much Context Does My Attention-Based ASR System Need?
Robert Flynn; Anton Ragni

Focused Discriminative Training For Streaming Automatic Speech Recognition
Adnan Haider

Speech Interactions Designed With Minoritised Language Speakers
Emily Nielsen; Electra Wallington; Ondřej Klejch; Dani Kalarikalayil Raju; Nina Markl; Gavin Bailey; Thomas Reitmaier; Jennifer Pearson; Matt Jones; Peter Bell; Simon Robinson

MyChat: a "written speech" corpus rich in conversational features
Mai Hoang Dao; Catherine Lai; Peter Bell

Discovery of Grammatical Matrix Language Markers in Code-Switched Text
Olga Iakovenko; Thomas Hain

Behavioral evidence for higher articulation rate convergence following natural than artificial time altered speech
Jérémy Giroud; Kirsty Phillips; Jessica Lei; Matthew H. Davis

Dynamic Time Warping as an Alignment Technique for Vocal Puppetry
Ella Crocker

Bias in Sparse Subnetworks for Multilingual Automatic Speech Recognition
Edward Storey; Naomi Harte; Peter Bell

SALMONN: Towards Generic Hearing Abilities for Large Language Models
Guangzhi Sun

Speech Synthesis for Sarcasm in Low-Resource Scenarios
Zhu Li; Yuqing Zhang

Zero-shot Audio Topic Reranking using Large Language Models
Mengjie Qian; Rao Ma; Adian Liusie; Erfan Loweimi; Kate M. Knill; Mark J.F. Gales

Exploring Category Specific Disorder in Chinese GAD, BD, and MDD using the Verbal Fluency Task and Functional Near-Infrared Spectroscopy
Yufei Ren; Gang Cui; Pozi Liu; Graham Murray

Investigating Listener's Perception of Conversational Speaking Style using an Interview-Based Approach
Adaeze Adigwe; Simon King; Sarenne Wallbridge

Towards End-to-End Spoken Grammatical Error Correction
Stefano Bannò; Rao Ma; Mengjie Qian; Kate M. Knill; Mark J.F. Gales

Hider-Finder-Combiner: Voice Conversion and Voice Privacy through Adversarial Information Hiding
Jacob J Webber; Oliver Watts; Gustav Eje Henter; Jennifer Williams; Simon King

A Comparison of Synthesis Method Impact on Listener Perception of Play-Acted Speech
Emily Lau; Brechtje Post; Kate Knill

A framework for flexible model combination of speech tasks
Shreyas Ramoji; Thomas Hain

Putting Expression in the Irish Synthetic Voice
Anna Maria Giovannini; Zihan Wang; Andrew Murphy; Maria O'Reilly; Ailbhe Ní Chasaide; Christer Gobl

A New Standardized and Reproducible Benchmarking framework for Automatic Personality Perception Experiments over the SSPNet Speaker Personality Corpus
Nesreen Alareef; Evangelia Fringi; Tanaya Guha; Alessandro Vinciarelli

15h30-16h00 - Coffee Break

16h00-17h00 - Oral Session A (chair: Erfan Loweimi)

Speech Modifications for Improved Listening Experience in ADHD Adults
Lucy Valls-Reed; Jennifer Williams

Variability of speech timing features across repeated recordings of non-pathological speech samples
Judith Dineley; Ewan Carr; Lauren L. White ; Catriona Lucas; Zahia Rahman; Tian Pan; Faith Matcham; Johnny Downs; Richard Dobson; Thomas F. Quatieri; Nicholas Cummins

18h30-22h00 - Social Event including Banquet Dinner in Association with Google

Tuesday, 2nd July 2024

08h30-09h00 - Registration

09h00-10h00 - Keynote B (chair: Stefano Bannò)

How 'conversational' are conversational products and technologies?
Prof. Elizabeth Stokoe

10h00-11h00 - Poster Session B

Automatic Speech Recognition System-Independent Word Error Rate Estimation
Chanho Park; Mingjie Chen; Thomas Hain

Frameworks for Assessing Privacy Risks in Affective Speech-Based Systems
Basmah Alsenani; Tanaya Guha; Alessandro Vinciarelli

Just Because We Camp, Doesn't Mean We Should: The Ethics of Modelling Queer Voices.
Atli Sigurgeirsson; Eddie L. Ungless

Portability of Text- v Audio-based Auto-markers for Dialogic Assessment
Simon W. McKnight; Stefano Bannò; Mark J. F. Gales; Siyuan Tang; Kate M. Knill

NOMAD: Unsupervised Learning of Perceptual Embeddings for Speech Enhancement and Non-Matching Reference Audio Quality Assessment
Alessandro Ragano; Jan Skoglund; Andrew Hines

Enhancing ASR performance for dysarthric speech: Deep Learning approaches and dataset expansion.
Leon Turner; Eugenio Donati

Investigating the Role of Visual Information in Audio-Visual Speech Recognition
Zhaofeng Lin; Naomi Harte

Has Artificial Intelligence Rendered Language Teaching Obsolete?
Zoe Handley

Effective Context in Self-Supervised Speech Models
Yen Meng; Hao Tang

NLP and Speech Processing to study Mental Health Recovery Narratives from NEON (Narrative Experiences Online collection)
Shrankhla Pandey; Sarah Morgan

On Robustness of Speaker Retrieval in the Wild: A Comparative Study of x-vector and ECAPA-TDNN Models
Erfan Loweimi; Mengjie Qian; Kate Knill; Mark Gales

Wearable Audio-Visual AI for Use in Hearing Aids
Poppy Welch; Jennifer Williams

Towards Multimodal Turn-taking for Naturalistic Human-Robot Interaction
Sam O'Connor Russell; Naomi Harte

Parameter Efficient Finetuning for Speech Emotion Recognition
Nineli Lashkarashvili; Wen Wu; Guangzhi Sun; Phil Woodland

Tracking Articulatory Feature Transition Phenomena in Speech Embeddings
Patrick Cormac English; Erfan A. Shams; John D. Kelleher; Julie Carson-Berndsen

3rd COG-MHEAR Audio-Visual Speech Enhancement Challenge (AVSEC-3)
Andrea Lorena Aldana Blanco,; Ondřej Klejch; Cassia Valentini-Botinhao; Peter Bell

Research plan: Neural encoding of AI-generated speech prosody by L1 and L2 speakers
Linda Bakkouche

Exploring individual speaker characteristics within a forensic automatic speaker recognition system
Chenzi Xu; Vincent Hughes; Paul Foulkes; Philip Harrison; Poppy Welch; Jessica Wormald; Finnian Kelly; David van der Vloed

An Audio-Based Depression Tracking Model Using Machine Learning
Andrea Vitullo; Eugenio Donati

Characterizing Code-switching: Applying Linguistic Principles for Metric Assessment and Development
Jie Chi; Electra Wallington; Peter Bell

Leveraging Language Affinities in Wav2Vec Finetuning for Low Resourced Languages
Jeffrey Michael; Oscar Saz

Bootstrapping Spoken Information Retrieval for Unwritten Languages
Ondřej Klejch; Electra Wallington; Thomas Reitmaier; Emily Nielsen; Dani Kalarikalayil Raju; Nina Markl; Gavin Bailey; Jennifer Pearson; Matt Jones; Simon Robinson; Peter Bell

Investigation of PEVD-based speech enhancement in wireless acoustic sensor networks
Emilie d'Olne; Patrick A. Naylor

Topical Emphasis synthesis with Voice Puppetry
Emelie Van De Vreken; Catherine Lai; Korin Richmond

SOT Triggered Neural Clustering for Speaker Attributed ASR
Xianrui Zheng; Guangzhi Sun; Chao Zhang; Philip C. Woodland

Fast and high-quality open-source speech synthesis with Matcha-TTS
Shivam Mehta; Ruibo Tu; Jonas Beskow; Eva Szekely; Gustav Eje Henter

A Study of Continual Test-Time Adaptation for Speech Recognition
Xinying Wei; Mingjie Chen; Thomas Hain

Prosodic cues and processing of Bulgarian-English code-switched speech
Yoana I. Dancheva; Margreet Vogelzang; Ianthi M. Tsimpli

On the Optimization Aspects of Variational Appoaches to Speech Representation Learning
Sung-Lin Yeh; Hao Tang

Factors influencing vowel categorisation flexibility
Stephanie Cooper; Brechtje Post

Effects of Voice Similarity on the Quality of Experience of Voice Assistants
Crisron Rudolf Lucas; Andrew Hines

HAFFORMER: A Hierarchical Attention-Free Framework for Alzheimer's Disease Detection From Spontaneous Speech
Zhongren Dong; Zixing Zhang; Weixiang Xu; Jing Han; Jianjun Ou; Björn W. Schuller

Liquid+sonorant epenthesis in Connemara Irish English: bilinguals vs. monolinguals
Kate Tallon

Investigating the Emergent Audio Classification Ability of ASR Foundation Models
Rao Ma; Adian Liusie; Mark J. F. Gales; Kate M. Knill

11h00-11h30 - Coffee Break

11h30-12h30 - Oral Session B (chair: Brian Sun)

SummaryMixing: A Linear-Complexity Alternative to Self-Attention for Speech Recognition and Understanding
Titouan Parcollet; Rogier van Dalen; Shucong Zhang; Sourav Bhattacharya

Can We Trust Explainable AI Methods on ASR? An Evaluation on Phoneme Recognition
Xiaoliang Wu; Peter Bell; Ajitha Rajan

12h30-13h30 - Lunch

13h30-14h30 - Poster Session C

Linear-Complexity Self-Supervised Learning for Speech Processing
Shucong Zhang; Titouan Parcollet; Rogier van Dalen; Sourav Bhattacharya

Linear Complexity Unified Streaming and Non-streaming Conformers for Transducer-based Speech Recognition
Titouan Parcollet; Rogier van Dalen; Shucong Zhang; Sourav Bhattacharya

Fine-tuning of Self-supervised Models Jointly with Autoencoders
Wenjie Peng; Thomas Hain

Analysis of self-supervised speech representations for tone languages
Opeyemi Osakuade; Simon King

Investigation of Spanish-accented English Pronunciation Patterns in ASR
Margot Masson; Julie Carson-Berndsen

CognoSpeak: an automatic, remote assessment of early cognitive decline in real-world conversational speech
Madhurananda Pahar; Fuxiang Tao; Bahman Mirheidari; Nathan Pevy; Rebecca Bright; Swapnil Gadgil; Lise Sproson; Dorota Braun; Caitlin Illingworth; Daniel Blackburn; Heidi Christensen

video-SALMONN: Speech-Enhanced Audio-Visual Large Language Models
Guangzhi Sun

Using Structured Conversational Prompts in the Diagnosis of Dementia
Fritz Peters; Heidi Christensen

Crossmodal ASR Error Correction with Discrete Speech Units
Yuanchao Li; Pinzhen Chen; Peter Bell; Catherine Lai

Exploring Acoustic Features for Challenging Voice Anonymisation Speakers
Henry Card; Jennifer Williams

English for Academic Purposes (EAP) Tutors' Knowledge, Beliefs and Practices in Relation to the Use of Artificial Intelligence (AI)
Zoe Handley

Authentic Speaker Recognition
Qiang Huang

Learn and Don't Forget: Adding a New Language to ASR Foundation Models
Mengjie Qian; Siyuan Tang; Rao Ma; Kate M. Knill; Mark J. F. Gales

Phonation based control system using laryngeal bio-impedance and machine learning.
Eugenio Donati; Christos Chousidis

The Effect of Decoding Strategies on the Quality and Diversity of Speech Generated by Large Speech Models
Adaeze Adigwe; Zehai Tu; Simon King

Efficiency and (lack of) flexibility in an LSTM network model of spoken word recognition
Máté Aller; Matthew H. Davis

Designing Conversational Assistants to Discourage Abusive Language
Tanvi Dinkar; Gavin Abercrombie; Chris Pidcock; Benedict Jones; Matthew Aylett

Multichannel Binaural Speech Enhancement using Complex Convolutional Transformer Networks
Vikas Tokala; Emilie d'Olne; Mike Brookes; Simon Doclo; Jesper Jensen; Patrick A. Naylor

Low-dimensional Style Token Control for Hyperarticulated Speech Synthesis
Dan Wells; Miku Nishihara; Korin Richmond; Aidan Pine

Spoken English Learner Corpus Transcription Validation Process
Mateus Miranda

Spontaneous and Scripted Speech Classification for Multilingual Audio
Shahar Elisha; Mariano Beguerisse-Díaz; Emmanouil Benetos

Exploring Gender Disparities in Automatic Speech Recognition Technology
Hend ElGhazaly; Bahman Mirheidari; Nafise Sadat Moosavi; Heidi Christensen

Word Time Prediction for Attention-Based Encoder-Decoder Models in Automatic Speech Recognition
Dongcheng Jiang; Phil Woodland

Semantic Map-based Generation of Navigation Instructions
Svetlana Stoyanchev; Chengzu Li; Chao Zhang; Simone Teufel; Rama Sanand Doddipatla

Leveraging data sources for foundation models in dialectal Arabic TTS
Simon Shelley; Omar Elsherief; Oscar Saz

Using speech graph analysis to differentiate people at clinical high risk of psychosis from healthy controls
Xinyi Liang

Improving Retrieval-Augmented Response Generation in Goal-Oriented Dialogue Question Answering
Norbert Braunschweiler; Abigail Sticha; Rama Doddipatla; Kate Knill

Acoustic Knowledge and Impact of Explainable Artificial Intelligence (XAI) on Audio Models Explainability
Marjan Golmaryami; Jennifer Williams

Accounting for vocal mismatch within an automatic speaker recognition system
Tallulah Buckley; Vincent Hughes; Paul Foulkes; Philip Harrison; Jessica Wormald; Finnian Kelly; David van der Vloed

Evaluating the Effectiveness of the Conformer ASR Model in Handling Spanish Morphology
Mourbi Basak

Muting Whisper: A Universal Acoustic Adversarial Attack on Speech Foundation Models
Vyas Raina; Rao Ma; Charles McGhee; Kate Knill; Mark Gales

Zipfian Properties of English and Finnish Documents
Martin Tunnicliffe; Gordon Hunter

Low-resource Speech Recognition and Dialect Identification of Irish in a Multi-task Framework
Mengjie Qian; Liam Lonergan; Neasa Ní Chiaráin; Christer Gobl; Ailbhe Ní Chasaide

Technical challenges and opportunities in building an Irish-language augmentative and alternative communication (AAC) system
Rían Errity; Emily Barnes; Julie Cummins; Oisín Morrin; Harald Berthelsen; Christoph Wendler; Andy Murphy; Christer Gobl; Neasa Ní Chiaráin; Ailbhe Ní Chasaide

14:30-15h30 - Keynote C (chair: Simon McKnight)

Using Machine Learning to Improve Hearing Aid Signal Processing: The Clarity and Cadenza Challenges
Prof. Jon Barker

15h30-16h15 - UKIS 2024 Cambridge: Future Plans & Farewell

Keynotes

Keynote A - Dr. Catherine Lai

Across the prosodic dimension: Exploring spoken communication beyond text



Abstract

Recent advances in machine learning have made an undeniable impact on the field of speech technology as we've long known it. These advances have also led so some rather bold claims: e.g., Speech-to-text (aka Automatic Speech Recognition) and Text-To-Speech synthesis are solved! What these sorts of claims often miss is that the traditional objectives of speech technologies neglect important aspects of spoken communication beyond text. For example, most machine learning oriented work on spoken language understanding is still focuses on text-based methods, ignoring the fact that how we speak can change how our words are interpreted. Nevertheless, previous work has shown that speech prosody (e.g. pitch, energy and timing characteristics of speech) can be used to signal speaker intent and affect, as well as to infer and project dialogue structure. We also know that prosody can be highly contextually variable. So, to make use of prosody in speech technology, we need to be able to model this variability and to understand what it actually does in spoken communication. In this talk, I will discuss recent work exploring prosodic variation in (English) spoken dialogue, using representation learning methods developed for speech generation and recognition. I argue that there are lot of benefits to be had from self-supervised methods for representation learning on speech and text datasets, but we still need linguistic knowledge to actually make use of the true richness of speech.

Biography

Catherine Lai is a Lecturer in Speech and Language Technology, based in Linguistics and English Language and the Centre for Speech Technology Research at the University of Edinburgh. Her main interest is speech prosody, e.g., intonation and rhythmic properties of speech, how it contributes to spoken dialogue understanding, drawing on both a speech technology/machine learning and linguistic/social science perspectives. She was previously a post-doctoral researcher in the School of Informatics at the University of Edinburgh, where she was the Principal Investigator on a Toyota funded grant on spoken language processing for robot companions. She gained her PhD in linguistics from the University of Pennsylvania, after MSc and BSc (Hons) degrees in computer science and mathematics from the University of Melbourne. Amongst other things, she's also worked on emotion and stance recognition, speech processing for multimedia archives, human evaluation and interaction with speech technologies, the role prosody in theories of semantics/pragmatics, and (in a past life) linguistic tree query languages.

Keynote B - Prof. Elizabeth Stokoe

How ‘conversational’ are conversational products and technologies?



Abstract

Conversational products and technologies are in the headlines more than ever. But how ‘conversational’ are they? And what does ‘conversational’ actually mean? Many products leverage ‘conversation’, from communication training to assessment tools; from to scripted interaction to role-play, and from chatbots to voice assistants. But do they do so in ways that strengthen or do damage in their domains of use? Six decades of research in conversation analysis have identified and described the constitutive practices of human social interaction across the widest range of ordinary and institutional settings. In this talk, I will address the questions of what, when, and how conversational products could and should leverage from conversation analysis.

Biography

Elizabeth Stokoe is a professor in the Department of Psychological and Behavioural Science at The London School of Economics and Political Science. She conducts conversation analytic research to understand how talk works – from first dates to medical communication and from sales encounters to crisis negotiation. She has worked as an industry fellow at technology companies Typeform and at Deployed. In addition to academic publishing, she is passionate about science communication, and has given talks at TED, Google, Microsoft, and The Royal Institution, and performed at Latitude and Cheltenham Science Festivals. Her books include Talk: The Science of Conversation (Little, Brown, 2018) and Crisis Talk (Routledge, 2022, co-authored with Rein Ove Sikveland and Heidi Kevoe-Feldman). Her research and biography were featured on BBC Radio 4’s The Life Scientific. During the Covid-19 pandemic she participated in a behavioural science sub-group of the UK Government’s Scientific Advisory Group for Emergencies (SAGE) and is a member of Independent SAGE behaviour group. She is a Wired Innovation Fellow and in 2021 was awarded Honorary Fellowship of the British Psychological Society.

Keynote C - Prof. Jon Barker

Using Machine Learning to Improve Hearing Aid Signal Processing: The Clarity and Cadenza Challenges



Abstract

At least 1.5 billion people are currently living with hearing loss, and this number will increase as the global population ages. Many of these people would benefit from hearing aids, yet only a fraction of them have them and, further, many who have them do not use their devices often enough. A major reason for the low uptake is that hearing aids do not perform well enough in many everyday situations. Among users' biggest complaints is that speech often remains poorly intelligible when listening in noisy situations, and that hearing aids often do not cope well with music. However, recent advances in machine learning have the potential to directly address these problems and transform the experience of hearing aid users. In this talk, I will be presenting two large EPSRC projects Clarity (speech) and Cadenza (music) that are collaborations between the Universities of Sheffield, Salford, Cardiff, Leeds and Nottingham. These projects have been designed to investigate the potential for hearing aid machine learning, and also to grow the community of researchers working in this area. The projects are using a series of open challenges to achieve these goals. Clarity has been considering speech intelligibility enhancement and speech intelligibility prediction, while the more recent project, Cadenza, has been considering music enhancement through a process of source separation and hearing impairment-aware remixing. The talk will explain some of the difficulties inherent in hearing aid signal processing, how recent advances from the speech community are being applied, and new approaches that are emerging from the latest Clarity/Cadenza challenges. The talk will also present the current challenges (the 3rd Clarity Enhancement Challenge, and the 2nd Cadenza Challenge) which will be ongoing at the time of the UKIS meeting with plenty of opportunity for those who are interested in getting involved.

Biography

Jon Barker is a Professor of Speech Processing at the University of Sheffield. He has been working in the area of speech and audio processing for over 30 years. His earlier work was on modelling human speech processing with a focus on computational models of speech in noise perception. He also has interests in distant microphone speech processing and was a founder of the CHiME series of challenges and workshops for distant microphone ASR and diarisation. More recently he has been trying to bring these interests together in the area of hearing aid processing with research funded via the EPSRC Clarity and Cadenza projects and students funded by Sheffield's UKRI CDT in Speech and Language Technology.

Poster Session A

Label-Synchronous Neural Transducer for End-to-End ASR

Keqi Deng & Phil Woodland

Dept. of Engineering, University of Cambridge, Trumpington St., Cambridge, CB2 1PZ, U.K.
{kd502,pw117}@cam.ac.uk

The neural transducer (NT) is a widely used E2E ASR structure with good streaming properties. However, it still suffers from domain shifts. Domain adaptation is more challenging for E2E ASR than for the HMM-based systems which use a separate language model (LM) that can easily employ text-only data. Although the prediction network in NT models is analogous to the LM in terms of structure, it doesn't perform solely as an LM as it needs to coordinate with the acoustic encoder to generate both blank and non-blank tokens. The prediction of blank tokens poses a challenge when adapting the standard NT using text-only data due to its inconsistency with the LM task.

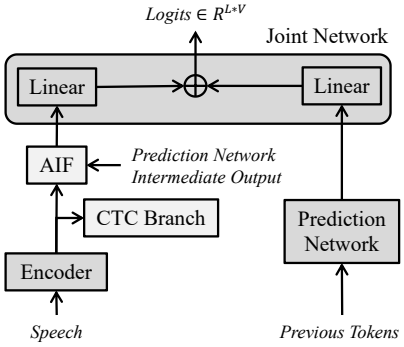


Figure 1: *Illustration of the proposed LS-Transducer. Linear denotes linear classifier. The output logits is a label-level two-dimensional matrix, where L and V are the length and vocabulary size. ⊕ denotes addition.*

This paper proposes a label-synchronous neural transducer (LS-Transducer) (Fig. 1), which extracts a label-level representation from the acoustic encoder output before combining it with the prediction network output, thus obviating the need for blank tokens. This makes the prediction network operate as an explicit LM, which is adaptable with text-only data. To generate this label-level encoder representation, an Autoregressive Integrate-and-Fire (AIF) mechanism (see Fig. 2) is proposed, which is extended from the Continuous Integrate-and-Fire (CIF) approach but with improved efficiency and better robustness to inaccurate unit boundaries. A streaming joint decoding method is also proposed to yield better accuracy.

ASR models were trained and tested on LibriSpeech data. The test/dev-clean/other sets were used for intra-domain evaluation. The source-domain text data included training set transcripts and LibriSpeech LM training text. The TED-LIUM2 dev/test sets and AESRC2020 were used as out-of-domain test corpora. For target-domain adaptation text, training set transcripts were used and TED-LIUM2 also included TED-LIUM2 LM training text. BPE-derived word-piece labels were used as the modelling units. Key results are given in Tables 1 and 2. The proposed LS-Transducer had superior ASR performance and effective domain adaptation capabilities, exceeding standard neural transducers with 12.9% and 24.6% relative WER reductions in intra-domain and cross-domain scenarios respectively.

Keqi Deng is funded by the Cambridge Trust. This work used computing resources provided by the 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.

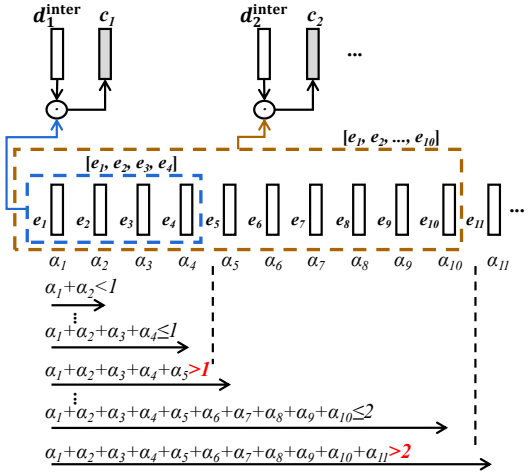


Figure 2: *Illustration of the proposed AIF. ⊙ denotes dot-product attention, whose query $\boldsymbol{d}_j^{\text{inter}}$ is the intermediate output of the prediction network.*

Table 1: *Intra-domain WER on LibriSpeech dev/test sets for online transducer models trained from LibriSpeech-960h.*

Online ASR Models	Test		Dev	
	clean	other	clean	other
Transformer-Pred T-T	3.2	8.0	3.0	7.8
+LM Shallow Fusion	3.1	7.7	2.9	7.5
Proposed LS-Transducer	3.0	7.2	2.9	7.4
+LM Shallow Fusion	**2.7**	**6.8**	**2.6**	**6.7**

Table 2: *Cross-domain WER results on TED-LIUM 2 (Ted2) and AESRC2020 (AESRC) for online transducer models trained from LibriSpeech-960h (LS960). SF denoted shallow fusion.*

Online ASR Models	LS960⇒Ted2		LS960⇒AESRC	
	Test	Dev	Dev	Test
Transformer-Pred T-T	12.7	13.1	19.0	18.7
+Target-domain LM SF	11.9	12.2	16.7	16.2
Proposed LS-Transducer	11.7	12.0	18.2	17.8
+Adapting Prediction Net	10.0	10.3	14.9	14.1
++Target-domain LM SF	**9.1**	**9.6**	**13.6**	**12.6**

Impact of Investigator Speech and its removal for Alzheimer's Dementia Detection

Marek Sviderski, Basel Barakat, Becky Allen

School of Computer Science, Faculty of Technology, University of Sunderland, Sunderland, UK
bh37sy@student.sunderland.ac.uk,{basel.barakat,becky.allen}@sunderland.ac.uk

Abstract

This study investigates the potential of acoustic analysis for early Alzheimer's disease (AD) detection using the Pitt corpus. This corpus contains patient and investigator speech alongside background noise. We explore innovative preprocessing techniques aimed at enhancing diagnostic accuracy. Specifically, we focus on the dual challenge of denoising and the selective removal of non-patient speech to improve the clarity of audio recordings for analysis. Our comparative investigation seeks to find an optimal balance between removing audio noise and preserving essential diagnostic speech markers, crucial for detecting subtle linguistic and acoustic indicators of AD [1, 2].

Methodology

This study utilise the Pitt corpus [1], a significant asset in dementia research, consisting of 1287 audio recordings. These recordings are categorised into healthy control (HC) and Alzheimer's Disease (AD) groups for binary classification tasks, or HC, AD, and Mild Cognitive Impairment (MCI) for multiclass classification. This dataset contains acoustic and textual data and demographic details including: the patient's sex, age, and education level.

Our hypothesis investigates the influence of audio modification techniques, particularly investigator speech removal and denoising, on the performance of AD detection algorithms. To evaluate our hypotheses, we establish an experimental framework (shown in Fig. 1 utilising four distinct datasets:

- **Baseline Version (BL)**: The original, unmodified collection of recordings.
- **Baseline Cut Version (BLC)**: Audio was carefully synchronised with the corresponding transcriptions, and sections of audio containing the investigator's speech were removed to focus solely on the participants' responses.
- **Denoised Version (DN)** : Applies the Facebook's Denoiser [3] model to the BL to reduce background noise.
- **Denoised Cut Version (DNC)**: Implements both the denoising process and the removal of investigator speech.

Utilising the OpenSmile library [4], we extracted features from the eGeMAPS, ComParE, and EmoBase datasets though not initially designed for AD detection, their application in this context has shown promising outcomes. Following Luz's methodology [5] in his Interspeech 2020 paper, these feature sets serve as a foundation for our analysis. We refined these features by removing those highly correlated with the recording length (Pearson's correlation coefficient greater than 0.2), resulting in final sets of 88 (eGeMAPS), 4,997 (ComParE), and 697 (EmoBase) features. For both classification (using models such as Decision Trees, Naive Bayes, and XGBoost among others), we applied Stratified K-Fold to ensure robust performance evaluation. The comparison is then observed through the average performance across all the machine learning models (and their folds).



Figure 1: *Framework overview*

Results and Discussion

Our study highlights a significant disparity in speech durations between investigators and Alzheimer's Disease (AD) patients, indicating the potential for methodological biases with investigator speech contributing on average up to an additional 228 seconds per session. Analysis of audio processing impacts on AD detection models reveals that denoising and investigator speech removal markedly enhance model accuracy. Specifically, denoising improved accuracy from 0.802 to 0.837 in the ComParE dataset, and removing investigator speech increased eGeMAPS accuracy from 0.768 to 0.804. These improvements underline the efficacy of preprocessing techniques in refining diagnostic models.

These findings stress the crucial role of sophisticated audio preprocessing in AD diagnostics, demonstrating that removing background noise and non-patient speech can significantly refine the analysis of patient speech patterns. Despite the complexity of acoustic signal analysis highlighted by variable model performance across different processing techniques and datasets, these results advocate for the continued development of targeted preprocessing methods. Our aim is to enhance AD detection capabilities, ensuring computational models are precisely tuned to identify cognitive impairments through speech analysis.

1. References

[1] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The Natural History of Alzheimer's Disease," *Archives of Neurology*, vol. 51, no. 6, Jun. 1994.

[2] Z. Nasreddine, V. Garibotto, S. Kyaga, and A. Padovani, "The Early Diagnosis of Alzheimer's Disease: A Patient-Centred Conversation with the Care Team."

[3] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *Interspeech*, 2020.

[4] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proc. ACM Multimedia (MM)*. Florence, Italy: ACM, 2010.

[5] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge," in *Proc. Interspeech*, 2020.

Imbalanced Multimodal Learning in Video Conversations

Samantha Kotey[1] *& Naomi Harte*[1]

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

koteys@tcd.ie, nharte@tcd.ie

1. Abstract

Virtual meetings have become essential tools for maintaining business, social and academic relationships. Video has the ability to connect people visually, allowing them to talk and observe each other's reactions and responses in real-time. Video conferencing platforms comprise of a variety of features, including the ability to record virtual meetings that can be replayed at a later date. Applications that effectively harness these recordings would help to significantly enhance productivity across organisations. One such application is video summarization, which enables users to access important information without having to view the full recording. However, developing applications that can summarise engaging and meaningful conversations is a challenging task, due to the complexity of accurately capturing unstructured information from multiple modality sources [1].

Video summarization can be defined as a binary classification problem, where there is a decision to include the segment in the summary, or not. When there are additional features, multimodal summarization can then be interpreted as a multi-class classification problem for heterogeneous data types. Virtual meetings encompass diverse types of audio, visual and linguistic data, that can incorporate a blend of binary, categorical and feature driven labels. The heterogeneous nature of multiple modality types leads to complex issues when designing models with multifaceted variables. Class imbalances arise when dominant modalities lead the model to disregard other valuable features, or when categories associated within these modalities are underrepresented [2].

Therefore, to study this imbalance problem and address the task of virtual meeting summarization, we extend and adapt the Candor dataset [3] by collecting dense annotations at the utterance level. The multimodal dataset is enriched with human annotations and machine generated labels, to create ground truth summaries of video conversations. Additionally, we propose a modular multimodal transformer architecture. The model, capable of handling up to six modalities, incorporates gated fusion and cost sensitive learning mechanisms.

Our objective is to study the relationships and dependencies between modalities, and systems capable of exploiting these characteristics. We evaluate our model through ROUGE scores, comparing the associated text for each ground truth clip against the corresponding text of the model output. To validate the efficacy of the scores, we conduct a human evaluation study on the generated video summaries. Encouraging results were achieved when combining a fused representation of all modalities, with individual aggregations that utilised class imbalance loss functions. Specifically, our findings highlight the fundamental role of both gesture and pitch in capturing salient information in summaries. This underscores the importance of incorporating non-verbal cues and emotional intonations when capturing the essence of conversational speech.

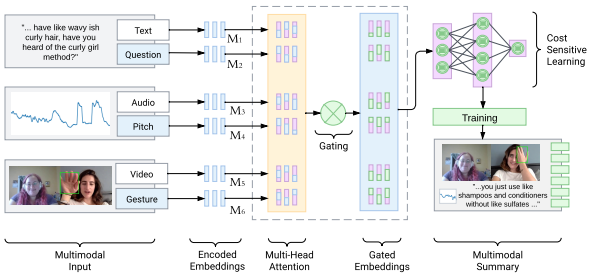


Figure 1: *Overview of our multimodal gated fusion model with multi-headed attention and cost sensitive learning for multimodal summarization. Independent modules are used to enhance the model's ability to predict the most salient utterances in a video chat conversation. There are three sets of features and sub features, totalling six modalities as input to the encoder. Each modality is transformed into a sequence of embeddings that represents the original raw data. The embeddings are projected into a shared embedding space, where the modular system manipulates the features to optimize the model training.*

Index Terms: multimodal summarization, virtual meetings, class imbalances, cost sensitive learning, gated fusion.

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.

2. References

[1] A. Jangra, S. Mukherjee, A. Jatowt, S. Saha, and M. Hasanuzzaman, "A Survey on Multi-modal Summarization," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–36, 2023.

[2] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," *Information Fusion*, vol. 91, pp. 424–444, 2023.

[3] A. Reece, G. Cooney, P. Bull, C. Chung, B. Dawson, C. Fitzpatrick, T. Glazer, D. Knox, A. Liebscher, and S. Marin, "The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation," *Science Advances*, vol. 9, no. 13, p. eadf3197, 2023.

Utilising Unsupervised Text-to-Speech Synthesis for Data Augmentation to Improve Accented Speech Recognition

Cong-Thanh Do[1], *Shuhei Imai*[2], *Rama Doddipatla*[1], *and Thomas Hain*[3]

[1]Cambridge Research Laboratory, Toshiba Europe Ltd., Cambridge, UK
[2]Graduate School of Engineering, Tohoku University, Sendai, Japan
[3]The University of Sheffield, Sheffield, UK

E-mails: cong-thanh.do@toshiba.eu, shuhei.imai.q4@dc.tohoku.ac.jp,
rama.doddipatla@toshiba.eu, t.hain@sheffield.ac.uk

1. Summary

This paper explores the use of unsupervised text-to-speech synthesis (TTS) as a data augmentation method to improve accented speech recognition. Instead of relying on manual transcriptions, TTS systems are trained using a limited amount of accented speech data and their corresponding pseudo-labels. This unsupervised approach allows us to leverage accented speech without the need for manual transcriptions to train TTS systems which can then generate synthetic accented speech data from text prompts. These synthetic samples are then combined with existing non-accented speech data to train automatic speech recognition (ASR) systems. Our ASR experiments are carried out in a self-supervised learning (SSL) framework, utilising a pre-trained Wav2vec2.0 model which is trained on a large amount of unsupervised accented speech data. For training the unsupervised TTS, we use read speech from the L2-ARCTIC and British Isles corpora as our accented speech data. For evaluation, we employ spontaneous conversational speech from the Edinburgh international accents of English corpus. Experimental results demonstrate that fine-tuning Wav2vec2.0 models with synthetic accented speech data, generated by the unsupervised TTS, leads to up to 6.1% relative word error rate (WER) reductions compared to a Wav2vec2.0 baseline model fine-tuned with non-accented speech data from Librispeech.

2. Data Augmentation for Accented Speech Recognition based on Unsupervised TTS

Accented speech recognition is an important research topic of ASR. Because of its importance, this research topic has been receiving attention and being addressed with various research approaches. In general, these approaches can be classified as accent-agnostic approaches, in which the modeling of accents inside the ASR systems is not made specific, and accent-aware approaches in which additional information about the accents of the input speech are used. TTS is an useful technology which can be used to improve ASR in a number of ways, for instance to improve the pre-training of SSL models or to improve the recognition of out-of-vocabulary words in end-to-end ASR. TTS was also used as a data augmentation method to improve speech recognition for Librispeech task and low-resource speech recognition. While having been widely used in various ASR tasks, TTS, especially unsupervised TTS which is trained with unsupervised audio data, has not been extensively studied as a data augmentation method in accented speech recognition.

In this work, we investigate the use of unsupervised TTS as a data augmentation method to improve accented speech recognition. In our approach, we make use of a small amount of accented speech data who do not have manual transcriptions to train TTS systems. This approach enables the use of accented speech data without manual transcriptions to perform data augmentation for accented speech recognition. Indeed, from a small amount of unsupervised accented speech data used to train the TTS systems, we can generate larger amount of synthetic accented speech data once the TTS systems are trained. In this paper, from 58 hours of accented speech data, selected from two speech corpora of read speech: L2-ARCTIC [1] and British Isles [2], we train unsupervised TTS and generate 250 hours more of synthetic accented speech data which help to achieve better gains on the evaluation data of spontaneous conversational speech from the Edinburgh international accents of English corpus (EdAcc) [3].

Table 1: *WERs on the development and test sets of the Edinburgh international accents of English corpus (EdAcc).*

Fine-tuning data \ Test data	Dev-set	Test-set
LS 960h (baseline)	32.8	35.1
LS 960h + AccD	32.4	34.6
LS 960h + TTS-LS 960h	31.4	33.8
LS 960h + TTS-AccD	31.0	33.2
LS 960h + AccD + TTS-LS 960h + TTS-AccD	30.8	33.2

Experimental results are shown in Table 1 where all the ASR models are fine-tuned from one Wav2vec2.0 pre-trained model with different fine-tuning data. In Table 1, LS 960h denotes 960 hours of Librispeech training data, AccD denotes 58 hours of unsupervised accented speech training data from L2-ARCTIC and British Isles corpora, TTS-LS 960h denotes 250 hours of synthetic non-accented speech data generated by TTS system trained on the LS-960h data, and TTS-AccD denotes 250 hours of synthetic accented speech data generated by TTS system trained on the AccD data. The Wav2vec2.0 models which are fine-tuned with synthetic accented speech data, generated by the unsupervised TTS, yield up to 6.1% relative WER reductions compared to the baseline.

3. References

[1] G. Zhao, S. Sonsaat, A. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, "L2-ARCTIC: a non-native English speech corpus," in *Proc. INTERSPEECH 2018*.

[2] I. Demirsahin, O. Kjartansson, A. Gutkin, and C. Rivera, "Open-source multi-speaker corpora of the English accents in the British Isles," in *Proc. of the 12th Conference on Language Resources and Evaluation*, Marseille, France, May 2020, pp. 6532–6541.

[3] R. Sanabria, N. Bogoychev, N. Markl, A. Carmantini, O. Klejch, and P. Bell, "The Edinburgh international accents of English corpus: towards the democratization of English ASR," in *Proc. 2023 IEEE ICASSP*, Rhodes Island, Greece, June 2023, pp. 1–5.

The effect of feeding and non-nutritive sucking on speech sound development at ages 2-5 years

Dr Sam Burr[1,2,3], *Professor Yvonne Wren*[1,2,4], *Dr Toity Deave*[5]

[1] Bristol Speech and Language Therapy Research Unit, North Bristol NHS Trust, Bristol, UK;
[2] School of Education, University of Bristol, Bristol, UK;
[3] Centre for Speech, Hearing and Communication Research, Cardiff Metropolitan University, Cardiff, UK;
[4] Bristol Dental School, University of Bristol, UK;
[5] Centre for Academic Child Health, University of the West of England, Bristol
sburr@cardiffmet.ac.uk

Background: Children with speech sound disorder (SSD) constitute the largest group of referrals to speech and language therapy (SLT) services in the UK[1-3]. This study aimed to identify potential risk factors for SSD regarding early feeding and sucking behaviours, which could support parents to make informed choices about how they care for their children, as well as informing SLT clinical practice in the UK and internationally. The study also aimed to identify whether different patterns of speech sound development were observed for children with different histories of early feeding and/or non-nutritive sucking (NNS).

Methods: This study used birth cohort data from the Avon Longitudinal Study of Parents and Children (ALSPAC)[4] and data from National Health Service (NHS) clinical SLT caseloads of children diagnosed with SSD. Data were obtained from parent questionnaires and clinical speech assessments.

Results: Different patterns of feeding and NNS were associated with different speech sound outcomes between ages 2- and 5-years-old. Associations between feeding and specific consonant sound errors were observed, specifically alveolar and velar sounds. NNS was not shown to impact speech outcomes at 5-years-old.

Conclusions: This study found that different patterns of speech sound development may be associated with different infant feeding and sucking histories. Further investigation of the influence of early feeding behaviours on specific consonant sounds at ages 2- and 5-years-old is required to inform clinical practice, including the neurodevelopmental mechanisms underlying feeding and sucking skills.

References

[1] J. Broomfield and B. Dodd, "Children with speech and language disability: caseload characteristics", *International Journal of Communication Disorders* [online], vol 39, no. 3, pp.303-324, 2004.

[2] P. Eadie, A. Morgan, O.C. Ukoumunne, K. Ttofari Eecen, M. Wake and S. Reilly, "Speech sound disorder at 4 years: prevalence, comorbidities, and predictors in a community cohort of children", *Developmental Medicine and Child Neurology* [online], vol 57, no. 6, pp.578–584, 2015.

[3] Y. Wren, L.L. Miller, T.J. Peters, A. Emond, and S. Roulstone, "Prevalence and Predictors of Persistent Speech Sound Disorder at Eight Years Old: Findings from a Population Cohort Study", *Journal of Speech Language and Hearing Research* [online], vol 59, no. 4, pp.647-673, 2016.

[4] Avon Longitudinal Study of Parents and Children (ALSPAC) University of Bristol, UK. 1991.

Exploring Speech Representations for Proficiency Assessment in Language Learning

Elaf Islam, Chanho Park*, Thomas Hain*

Dept. of Computer Science, The University of Sheffield, Sheffield, United Kingdom
{ejislam1,cpark12,t.hain}@sheffield.ac.uk

Abstract

Automatic proficiency assessment can be a useful tool in language learning, for self-evaluation of language skills and to enable educators to tailor instruction effectively. Often assessment methods use categorisation approaches. In this paper an exemplar based approach is chosen, and comparisons between utterances are made using different speech encodings. Such an approach has the advantage to avoid formal categorisation of errors by experts. Aside from a standard spectral representation pretrained model embeddings are investigated for the usefulness for this task. Experiments are conducted using speechocean762 database, which provides 3 levels of proficiency. Data was clustered and performance of different representations is assessed in terms of cluster purity as well as categorisation correctness. Cosine distance with Whisper representations yielded better clustering performance.

Index Terms: self-supervised representation, proficiency assessment, language learning

1. Speech Representation

In this section, we discuss three different speech representations employed in our study: Mel frequency cepstral coefficient, Wav2vec 2.0, and Whisper. Each representation offers unique advantages and contributes to the accurate assessment of proficiency in language learning.

2. Distance Measurement

Two different similarity measurements are used in this work based on two types of utterance representation. An utterance is either represented by features extracted at a fixed rate, or in the form of an overall embedding. Similarity between vector representations often uses the cosine distance whereas a common solution to compute distances between speech sounds is the use of Dynamic Time Warping (DTW), an example is Mel cepstral distortion as used for example in voice conversion assessment.

3. Data

In this study, the speechocean762 corpus was used, which is a free, open-source corpus of 5,000 English utterances collected from 250 Mandarin speakers [1]. The training set contains 2,500 utterances, 15,849 words, and 47,390 phones, while the test set contains 2,500 utterances, 15,967 words, and 47,688 phones. It assigns an utterance-level attribute score ranging from 0 to 10 to each speech for accuracy, fluency, completeness, prosody, and overall score.

* Equal contribution.

4. Experimental Setup

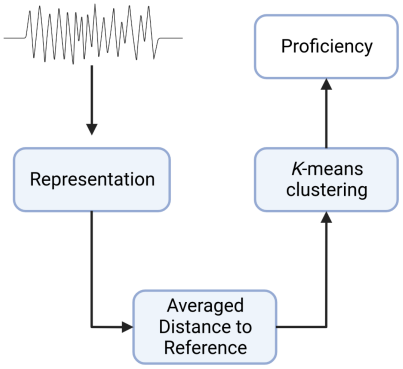


Figure 1: *An overview of the experimental setup employed in this study [2].*

5. Results

5.1. Purity scores and accuracy

Table 1: *Purity score and accuracy of clusters built using K-means in different representation spaces.*

Distance	Repr.	Init. centroids	Purity	Accuracy
3*DTW	MFCC	3*Random	0.427	0.425
	Wav2vec2		0.422	0.422
	Whisper		0.383	0.383
2*Cosine	Wav2vec2	2*Fixed	0.343	0.343
	Whisper		**0.429**	**0.429**

6. References

[1] J. Zhang, Z. Zhang, Y. Wang, Z. Yan, Q. Song, Y. Huang, K. Li, D. Povey, and Y. Wang, "speechocean762: An open-source non-native english speech corpus for pronunciation assessment," *arXiv preprint arXiv:2104.01378*, 2021.

[2] E. Islam, C. Park, and T. Hain, "Exploring speech representations for proficiency assessment in language learning," in *9th Workshop on Speech and Language Technology in Education (SLaTE) Proceedings*. International Speech Communication Association (ISCA), 2023, pp. 151–155.

Affinity: Dialogs to build human-robot rapport

Alina Larson[1], *Ronnie Smith*[1,2], *Jose Cerqueira Fernandes*[1,2], *Coena Das*[1,2], *Matthew Aylett*[1,3]

[1]Heriot-Watt University, UK
[2]The National Robotarium, UK
[3]CereProc, UK

A.Larson@hw.ac.uk, R.Smith@hw.ac.uk, C.Das@hw.ac.uk, J.Cerqueira_Fernandes@hw.ac.uk, M.Aylett@hw.ac.uk

Abstract

As conversations with artificial agents become more common, the long delays and pauses are increasingly met with frustration. Current robot are often speak-wait systems that don't allow for speech overlap, interruptions, and only process utterances once they are complete. It is our goal to address these challenges and implement human-like turn-taking for the social robot Haru. We will create a gold standard by recording a human/human casual semi-scripted dialogue based on a current Haru demo. Transcripts of theses dialogues will then used to train statistical models for human-robot dialogue and the audio recordings to analyse and model human-like turn-taking behavior in this restricted domain. The aim of the project is to create a conversational demo which is fluid, responsive, and can deal with natural turn-taking behavior.

Index Terms: speech recognition, human-computer interaction, computational linguistics, backchannels

1. Introduction

Although great strides have been made in the field of robotics, fluid conversation and human-like turn-taking remains one of the greatest challenges for social robots today. Conversation between humans involves overlap, interruption, backchannels, and speech can be processed before an utterance is complete. In contrast, most social robots that use speech synthesis and recognition tend to be speak-wait systems, requiring silence to process input before generating output. This results in long pauses after a human user is finished speaking before a robot can generate a response.

It is our goal to implement human-like turn-taking for the robot Haru without long speak-wait delays. Haru, a social robot developed by Honda Research Institute, is equipped with onboard ASR (Chatty SDK, developed by CereProc) that is capable of incremental word-spotting. Additionally, voice activation detection (VAD) enables Haru to detect whether a human interlocutor is still speaking, allowing for faster turn-taking.

Our current work in the Affinity Project is to develop faster turn-taking to allow for the use of backchannels in human-robot dialog. To this end, we have developed a series of semi-scripted getting-to-know-you conversational topics, for use in a human-human dialog. By recording a series of conversations between humans, we can establish a gold standard for conversational turn-taking between humans and robots.

The conversational prompts we use were originally developed for *Tiers of Friendship* [1], with a few modifications designed to encourage longer participant responses. These longer responses will provide ample opportunity for pauses and backchannels, which are of particular interest to us.

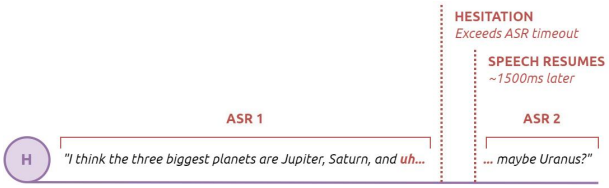


Figure 1: *Illustrative example of how hesitation may result in several discrete ASR results in human-robot dialog, where only one is desired. The circle marked 'H' represents a human speaker.*

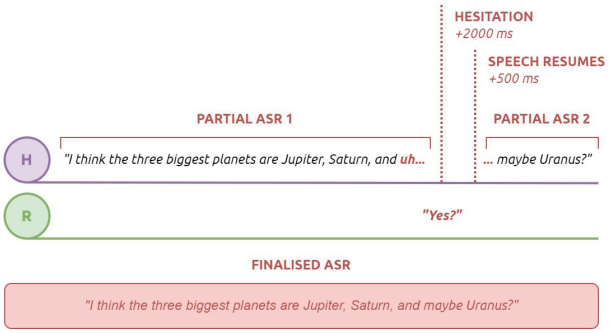


Figure 2: *Illustrative example of how multiple partial ASR results can be combined when the human speaker hesitates. Furthermore, it shows how this is an opportunity for backchannelling from the robot to the human, to encourage the speaker to continue. The circle marked 'H' represents the human speaker, while the circle marked 'R' represents the robot.*

The next phase of the Affinity Project will involve transcribing these human-human dialogs in RASA, and modeling turn-taking and backchanneling for Haru for human-robot dialog, using the same conversational topics.

2. References

[1] E. Nichols, S. R. Siskind, L. Ivanchuk, G. Pérez, W. Kamino, S. Šabanović, and R. Gomez, "Hey haru, let's be friends! Using the Tiers of Friendship to build rapport through small talk with the tabletop robot haru," in *Proc. IEEE/RSJ 2022 International Conference on Intelligent Robots and Systems (IROS)*, Kyoto, Japan, Oct. 2022, pp. 6101–6108.

Phonetics and Phonology inside the Black Box

Iona Gessinger, Erfan A. Shams, Julie Carson-Berndsen

ADAPT Centre, School of Computer Science, University College Dublin, Ireland
iona.gessinger@ucd.ie, erfan.shams@ucd.ie, julie.berndsen@ucd.ie

Automatic speech recognition (ASR) has made tremendous progress, with low average word error rates for high-resource languages [1]. From a linguistic perspective, this raises the question as to whether the underlying deep-learned models use the same phonetic cues for their predictions that linguists have identified as relevant markers for human speech perception, or whether they draw from other sources in the speech signal. However, to date, these models remain largely inaccessible to linguistic research. We present a methodology for investigating phonetic and phonological phenomena in current state-of-the-art ASR models which have been trained on large corpora.

Current ASR models use the transformer network architecture for end-to-end word or phoneme transcriptions. The transformer consists of embedding layers which are usually hidden from the user and often referred to as a black box. Recent research in explainable artificial intelligence (XAI) has led to new ways of investigating what is going on in the hidden representations. Probing the models with a downstream task is a way to discover what information is encoded in different layers of the model. Phone identification is a common probing method [2], which can be further broken down into the detection of individual articulatory features. In the case of consonants, for example, these are *place of articulation* (POA; 11 categories; see classification framework of the International Phonetic Alphabet), *manner of articulation* (MOA; 8 categories), and *voicing* (2 categories). Probing these articulatory features separately provides insight into the specific information embedded in the representations [3].

We present an approach which facilitates the investigation of phonetic and phonological processes by probing transformer-based ASR models, such as *wav2vec 2.0* [4] and *OpenAI Whisper* [1], using a simple machine learning model per articulatory feature. These probing models are trained on the hidden representations of the aforementioned ASR models, which are extracted using the audio data of the TIMIT corpus [5]. TIMIT also provides time-aligned phonetic transcriptions which are mapped to the articulatory features used as targets for training the probing models. In order to visualise the findings, we have developed a toolkit which allows us to follow the emergence of phonetic feature patterns through the transformer layers.

For the purposes of illustration, we identify areas within the TIMIT corpus where coarticulatory phenomena may occur; for example, regressive assimilation of voicing and/or POA. We cluster realisations of such transitions from various speakers in the corpus using agglomerative clustering based on the combined articulatory feature predictions (11+8+2, see above) per time frame. Identifying such clusters allows us to look at the expected variety of transitional patterns, but may also reveal coarticulatory patterns that lie below the threshold of gradient change to the auditory input that would cause human annotators to assign a different phoneme category [6]. We can further analyse the alternative predictions derived from the probe outputs, to discover whether they illustrate a smooth transition between articulatory feature categories (see Figure 1). These findings can lead to a better understanding of the decision-making process of ASR models.

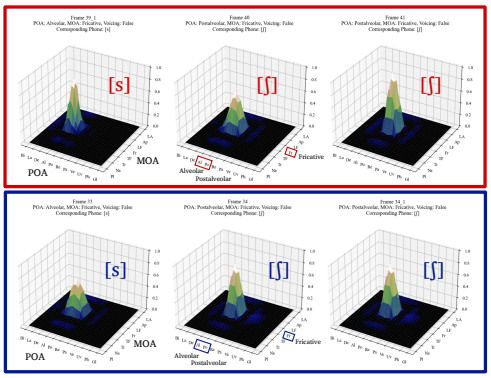


Figure 1: *Articulatory feature transition for patterns with identical phone predictions, but different POA/MOA feature prediction probabilities. Top middle frame leans towards alveolar POA, while bottom first frame leans towards postalveolar POA.*

1. References

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *ICML*, 2023, pp. 28 492–28 518.

[2] J. Shah, Y. K. Singla, C. Chen, and R. R. Shah, "What all do audio transformer models hear? Probing acoustic representations for language delivery and its structure," *arXiv:2101.00387v2*, 2021.

[3] P. C. English, E. A. Shams, J. D. Kelleher, and J. Carson-Berndsen, "Following the embedding: Identifying transition phenomena in wav2vec 2.0 representations of speech audio," in *ICASSP*, 2024, pp. 6685–6689.

[4] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS*, vol. 33, 2020, pp. 12 449–12 460.

[5] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," Linguistic Data Consortium, 1993.

[6] B. McMurray, "The myth of categorical perception," *JASA*, vol. 152, no. 6, pp. 3819–3842, 2022.

This research was conducted with the financial support of Science Foundation Ireland at ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology at University College Dublin [13/RC/2106_P2]. It is accepted for presentation at the 10th Biennial International Conference on the Linguistics of Contemporary English (BICLCE), Alicante, 26–28 September 2024.

Towards explainable speaker recognition neural networks

Yanze Xu[1], *Mark Plumbley*[1], *Wenwu Wang*[1]

[1]Centre for Vision, Speech and Signal Processing, University of Surrey

yanze.xu@outlook.com

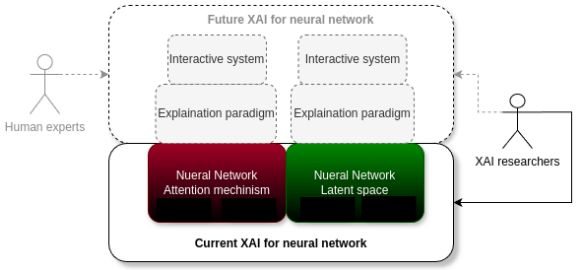


Figure 1: *The overview of the current and future Explainable AI (XAI) research for neural network*

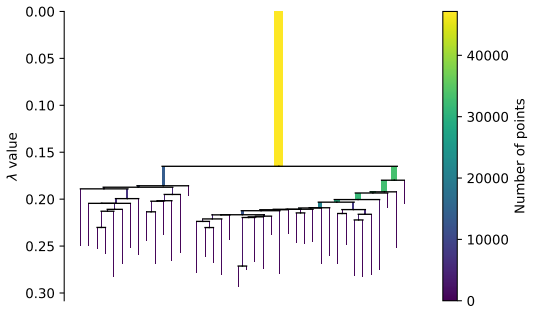


Figure 2: *A demonstration to the hierarchy of the representation space of speaker recognition network*

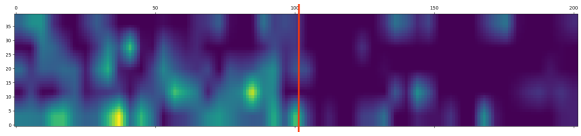


Figure 3: *A demonstration to the class-discriminative ability of GradCAM's attention map on contrastive-learned speaker recognition*

Neural networks trained from large amounts of data can solve different tasks using learned representations instead of hand-crafted features. However, these learned representations are hard to understand. As shown in Fig. 1. "Explainable AI" is an emerging research area to explain the decision-making process of AI systems such as neural networks. In this thesis, we aim to investigate explainable AI research for speaker recognition neural networks, where the neural network is trained to discriminate the speaker identity of voices.

One of the explainable AI approaches focuses on researching the learned representation space formed by the neural network. This is mainly achieved by visualizing the latent representation space using flat clustering algorithms like T-SNE (t-Distributed Stochastic Neighbor Embedding). However, flat clustering methods can only group data into independent clusters but cannot reveal the relationship between these clusters. To explain the relationships between clusters in the learned representation space, we employ hierarchical clustering algorithms to extract dendrograms from learned representations of the speaker recognition neural network using SLINK (Single-Linkage Clustering) and HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), as shown in Fig. 2. The extracted dendrograms are then explained with reference to gender, nationality, and identity information, by finding and quantifying which cluster can well represent certain semantic classes: the cluster-class matching task.

Another explainable AI approach investigating the attention mechanism during neural network decision-making. This is mainly achieved by utilizing CAM (Class Activation Map) and its variants to visualize attention maps for convolutional neural networks (CNNs). We examine two CAM variants, Gradient-weighted Class Activation Mapping (GradCAM) and Hierarchical Class Activation Maps for Localization (LayerCAM), to visualize attention maps for the speaker recognition model particularly trained by contrastive loss. Subsequently, a qualitative case study is performed to discuss the class-discriminative ability of GradCAM and LayerCAM, as shown in Fig. 3. While quantitative distribution analysis is performed to reveal the numerical characteristics of GradCAM and LayerCAM's attention maps.

We also investigate attention map evaluation based on Randomized Input Sampling for Explanation-Evaluation (RISE-eval). We find that RISE-eval is prone to a "self-similarity" problem. To overcome this problem, we build a new evaluation method, which we call "Robust RISE-eval". We show that the resulting Robust RISE-eval evaluation method has ability to penalize self-similarity problem without affecting the evaluation of the normal attention maps.

In summary, we investigate explainable AI research for the speaker recognition neural network from three perspectives: explaining the hierarchy of its latent representation space; visualizing and analysing its attention mechanism for the model particularly trained by contrastive loss; and making its attention map evaluation more robust. In the future, we hope to find a systematic way that can explain a lot of neural networks that can discriminate human voices.

Can political science help us improve speech synthesis evaluation?

Sébastien Le Maguer, Juraj Šimko

University of Helsinki, Helsinki, Finland
{sebastien.lemaguer, juraj.simko}@helsinki.fi

1. A bit of context

Speech synthesis evaluation is now at a crossroad. As speech synthesis models have improved during the last decade, the direct consequence is the obsolescence of the current standard protocols. Indeed, the Absolute Category Rating (ACR) [1] (or MOS-test) and its corresponding score MOS have now shown problematic limits [2].

Alternative protocols such as the MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) [3] or the preference tests exists. Yet, MUSHRA is improperly used as its recommendation explicitly requires the introduction of anchors (i.e., artificially constructed, degraded speech signals aiming to stabilise the rating scale). Such anchors are yet to be defined in the context of a speech synthesis evaluation. Researchers and engineers simply overlook them when conducting a MUSHRA in the context of speech synthesis evaluation; violating de-facto the recommendation. It is also well kown that using "out-of-context" (generic) evaluation protocols does not capture well the differences between modern systems [4]. Yet, the metric to be used while considering the context A proposed alternative which have been advocated for [5] is the use of preference tests (or AB(X) test) to conduct the evaluation as it leads to a relative analysis. However, conducting such evaluation remains cumbersome as comparing multiple systems implies multiplying the number of steps.

Relying on the same premise - the analysis needs to be conducted in a relative manner - we propose to rely on a political science staple: the Ranked-Choice Voting (RCV).

2. (Adapted) RCV at the rescue

RCV is not new[1] and has been well studied in computational social sciences [6] for a long time. It relies on following premise: during an election, instead of selecting a preferred candidate, the voter will rank the candidate by decreasing order of preference. The appeal of the RCV is its ability to provide a more diverse and representative outcome as the second choices are also taken into account. The difficulty mainly lies on the selection of the "winner" and multiple strategies to address it have been proposed [6].

Our goal is not to determine the best system but how a given system compares to given anchors. To do so, we first propose to consider standard systems trained using standard datasets and a given configuration as anchors. This is widely available and we should take opportunity of it. Then, as we are not interested in a score but as relative comparison of systems, we do not faces the constraint of MUSHRA.

Then, we propose to adapt the RCV paradigm to fit our needs: 1) we only consider one round; 2) listeners can associate the same rank to multiple systems (equivalent to "no-preference"); 3) listeners can choose to not rank samples which they deem of a too poor quality; 4) the samples are randomized and their corresponding system are hidden to the listener.

For the present submission, no control over the samples (except play and repeat) is given to the listener. The investigation of a more refined control (as proposed for MUSHRA) is left for future work.

[1] it can be traced back at least to the 13th century.

3. Experimental protocol

In order to have a comparison point, we propose to conduct the RCV evaluation based on the material of *Experiment 4* presented in [2]. This experiment consisted on evaluation the naturalness of four speech synthesis systems, obtained by combining two acoustic models (FastPitch/Tacotron) and two neural vocoders (WaveNet/WaveGAN), and the natural voice as the hidden reference. The results of this experiments show that almost all the synthesis systems were deemed equivalent with the only exception being FastPitch/WaveNet being considered significantly better. By using RCV, we want to determine 1) if a clear ranking emerges 2) how different strategies can lead to a different "winner".

Experiment 4 consisted of each listener evaluating all the synthesis combinations for all utterances. This allows us to convert the scores to a system ranking for each utterance and each listener. By doing so, we can now compare the rankings obtained by the RCV procedure to the ones obtained using the ACR.

The results and discussion will be presented during the conference.

4. References

[1] "Methods for subjective determination of transmission quality," ITU-T, Geneva, Tech. Rep. P.800, 1996.

[2] S. Le Maguer, S. King, and N. Harte, "The limits of the mean opinion score for speech synthesis evaluation," *Computer Speech & Language*, vol. 84, p. 101577, 2024.

[3] ITU-T, "Method for the subjective assessment of intermediate sound quality (MUSHRA)," International Telecommunication Union (ITU-R), Tech. Rep. BS.1534-1, 2001.

[4] J. O'Mahony, P. Oplustil-Gallegos, C. Lai, and S. King, "Factors Affecting the Evaluation of Synthetic Speech in Context," in *ISCA Speech Synthesis Workshop*, 2021, pp. 148–153.

[5] S. Shirali-Shahreza and G. Penn, "Mos naturalness and the quest for human-like speech," *IEEE Spoken Language Technology Workshop (SLT)*, Dec 2018.

[6] H. Nurmi and R. P. Palha, "A theoretical examination of the ranked choice voting procedure," in *Transactions on Computational Collective Intelligence XXXVI*. Springer, 2021, pp. 1–16.

SCORE: Self-supervised Correspondence Fine-tuning for Improved Content Representations

Amit Meghanani, Thomas Hain

The University of Sheffield, UK

ameghanani1@sheffield.ac.uk, t.hain@sheffield.ac.uk

Abstract

Self-supervised fine-tuning (SSFT) approaches are becoming popular to obtain task-specific representations. These task-specific representations are used for robust performance on various downstream tasks by fine-tuning on the labelled data. Nevertheless, enhancing Self-supervised learning (SSL)-based speech representations with unlabeled speech for content-related tasks has been noted as challenging and computationally expensive. This work presents a cost-effective SSFT method named **S**elf-supervised **Corre**spondence (SCORE) fine-tuning to adapt the SSL speech representations for content-related tasks. The proposed method uses a correspondence training strategy, aiming to learn similar representations from perturbed speech and original speech. SCORE fine-tuned HuBERT outperforms the vanilla HuBERT on SUPERB benchmark with only a few hours of fine-tuning ($<$ 5 hrs) on a single GPU for automatic speech recognition, phoneme recognition, and query-by-example tasks, with relative improvements of 1.09%, 3.58%, and 12.65%, respectively. SCORE provides competitive results with the recently proposed SSFT method SPIN, using only 1/3 of the processed speech compared to SPIN.

1. Proposed Method

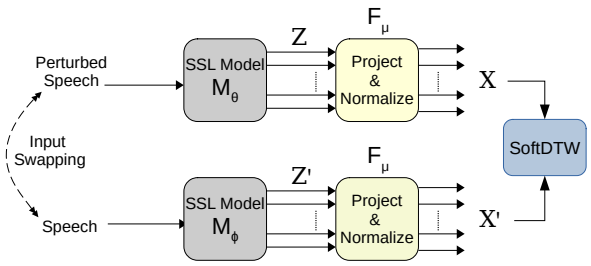


Figure 1: *SCORE processes a pair consisting of the original and perturbed speech inputs. It then compares the output sequence X generated by the learnable model M_θ with the output sequence X' produced by the frozen model M_ϕ using the soft-DTW loss function.*

The SCORE fine-tuning method is illustrated in Fig 1. Pre-trained SSL models (HuBERT or WavLM) are fine-tuned using this approach. Two instances of pre-trained models (M_θ: top 2 layers learnable and M_ϕ: frozen) are employed. The input to SCORE is a pair of perturbed speech and original speech, randomly assigned to either M_θ or M_ϕ, as depicted in Fig. 1. To generate the perturbed speech, data augmentations used in ASR [1] such as speech perturbation and pitch shift are applied. The sequences obtained from both models (X and X') differ in length due to these perturbations. Therefore, a dynamic time warping-based differentiable loss function, soft-Dynamic Time Warping (soft-DTW) [1][2], is used to align sequences of unequal lengths. This learning framework ensures that the model learns speed and pitch invariant representations for the same spoken content.

This work was supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1]. This work was also funded in part by LivePerson, Inc.

Model	ASR (WER)↓	PR (PER)↓	QbE (MTWV)↑
HuBERT	6.42	5.02	7.19
WavLM	6.17	4.85	9.15
HuBERT+SCORE	6.35	4.84	8.10
WavLM+SCORE	6.15	4.72	9.22

Table 1: *Performance of SCORE fine-tuned HuBERT and WavLM on SUPERB benchmark (ASR, PR, and QbE).*

2. Results and Discussion

LibriSpeech's [3] train-clean-100 dataset was used for SCORE fine-tuning. The fine-tuned models were evaluated on three content-related downstream tasks from the SUPERB benchmark: automatic speech recognition (ASR), phoneme recognition (PR), and query-by-example spoken term discovery (QbE). SCORE fine-tuned models outperformed their vanilla versions on PR and QbE (and marginal improvement for ASR). We believe that employing stronger data augmentation techniques on speech waveforms could yield improvements for ASR as well.

3. References

[1] T. Ko, V. Peddinti, P. D, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. Interspeech 2015*, 2015, pp. 3586–3589.

[2] M. Cuturi and M. Blondel, "Soft-DTW: a differentiable loss function for time-series," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 894–903. [Online]. Available: https://proceedings.mlr.press/v70/cuturi17a.html

[3] 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.

[1] https://github.com/Maghoumi/pytorch-softdtw-cuda

Exploring Dominant Paths in CTC-like ASR Models: Unraveling the Effectiveness of Viterbi Decoding

Zeyu Zhao, Peter Bell, Ondřej Klejch

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

zeyu.zhao@ed.ac.uk, peter.bell@ed.ac.uk, o.klejch@ed.ac.uk

Abstract

Connectionist Temporal Classification (CTC) has emerged as a fundamental technique in Automatic Speech Recognition (ASR), renowned for its ability to marginalise all possible alignments between input and target sequences. This study reevaluates the traditional dependency on prefix beam search, which typically considers multiple alignments for each hypothesis, in CTC-based models. Our findings initially indicate the absence of a single dominant path in the lattices. However, we discover that the Viterbi decoder can effectively emulate the performance of a prefix beam search, as the hypotheses it identifies tend to be dominant in the lattices. This leads us to propose that an effective CTC-like model should not only aim for high accuracy but also align the optimal hypothesis with the most probable path, thereby enhancing the robustness of Viterbi decoding. Furthermore, our insights extend to a variety of topologies, demonstrating their applicability within a more comprehensive ASR framework.

Index Terms: ASR, E2E ASR, Differentiable WFST, Topology, Viterbi Decoding

1. Method

1.1. Decoding Methods

$$W^* = \arg\max_W [\log p(W) + \alpha \max_{\pi \in \Pi(W; \boldsymbol{S}^{\text{dec}})} \log p(\pi|X)] \tag{1}$$

$$W^* = \arg\max_W [\log p(W) + \alpha \sum_{\pi \in \Pi(W; \boldsymbol{S}^{\text{dec}})} \log p(\pi|X)] \tag{2}$$

1.2. Lattice and Metrics

Given the lattice L [1] and the best hypothesis W^*, we define *Best Path Proportion (BPP)* P_{path} as,

$$P_{\text{path}} = \frac{\max_{\pi \in \Pi(W^*; L)} p(\pi|X)}{\sum_{\pi \in \Pi(W^*; L)} p(\pi|X)}. \tag{3}$$

We define *Best Hypothesis Proportion (BHP)* as,

$$P_{\text{hypo}} = \frac{\sum_{\pi \in \Pi(W^*; L)} p(\pi|X)}{\sum_W \sum_{\pi \in \Pi(W; L)} p(\pi|X)}, \tag{4}$$

where $\sum_W$ is the summation over all the hypotheses in L. It is important to note that, $P_{\text{hypo}} > 0.5$ indicates that the best hypothesis identified by the Viterbi decoder in eq. (1) aligns with the full-sum decoder in eq. (2), where the summation is limited to the lattice.

2. Experiments

Table 1: *BPP (average) on WSJ and Librispeech*

	WSJ (eval92)	Librispeech (test-other)
S1-T1	0.285%	0.633%
S2-T1	0.119%	0.576%
S2-T1★	0.0970%	0.621%
S2-T2★	0.0113%	0.148%
S2-T2	0.0373%	0.311%
S3-T2	0.0844%	0.364%
S3-T2★	0.00643%	0.294%
S3-T2★★	0.00390%	0.473%

Table 2: *The ratio of $P_{\text{hypo}} > 0.5$ on WSJ and Librispeech*

	WSJ (eval92)	Librispeech (test-other)
S1-T1	87.0%	59.5%
S2-T1	92.7%	62.1%
S2-T1★	90.9%	61.1%
S2-T2★	91.5%	62.5%
S2-T2	89.7%	60.3%
S3-T2	93.6%	58.8%
S3-T2★	91.8%	45.4%
S3-T2★★	92.1%	39.8%

Table 3: *WER% in two subsets of eval92*

	$P_{\text{hypo}} > 0.5$	$P_{\text{hypo}} < 0.5$
S1-T1	1.89	8.12
S2-T1	2.48	6.83
S2-T1★	2.33	4.72
S2-T2★	2.24	5.71
S2-T2	2.28	5.16
S3-T2	2.35	4.75
S3-T2★	2.40	7.43
S3-T2★★	2.32	5.36

3. References

[1] D. Povey, M. Hannemann *et al.*, "Generating exact lattices in the WFST framework," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2012, pp. 4213–4216.

This work was supported by a Consolidated Studentships Award funded by Huawei Technologies Co., Ltd.

This word was accepted and presented in XAI-SA Workshop ICASSP 2024.

Detailed introduction will be provided with the poster.

Multilingual Integration in Lyrics Transcription: Data, Language Conditioning, and Transliteration Augmentation

Jiawen Huang, Emmanouil Benetos

Centre for Digital Music, Queen Mary University of London, London, UK

{jiawen.huang, emmanouil.benetos}@qmul.ac.uk

1. Introduction

Lyrics play an important role in music. Access to lyrics enhances user experiences in music listening and effectively communicates emotions. It also enables users to sing along and facilitates research in music content analysis. Moreover, singing is a valuable resource for language learners, offering an enjoyable learning experience. Multilingual automatic lyrics transcription (ALT) is challenging due to the limited availability of labelled data and the challenges introduced by singing, compared to multilingual automatic speech recognition. Although some multilingual singing datasets have been released recently, English continues to dominate these collections. Multilingual ALT remains underexplored due to the scale of data and annotation quality.

Through our research, we aim to advance the field of lyrics transcription by providing practical insights and innovative methodologies, ultimately enhancing the accuracy and interpretability of automated lyric transcription systems. We conduct a systematic study with a lightweight baseline on the multilingual nature of the task, revealing the connections between language and ALT performance, as well as the low-resource and imbalanced data situation, in the context of singing voice.

Our contributions are as follows: We provide practical guidelines covering training data, features, and language conditioning mechanisms. Our findings reveal the effectiveness of joint multilingual training and conditioning mechanisms. We develop a model that can perform inference with and without language information. Building upon these guidelines, we propose a transliteration training approach that combines joint training and conditioning to address the low-resource and imbalanced data problems.

2. Datasets and Baseline Model

Two datasets, MulJam v2 [1] and DALI v2 [2], are involved in training. MulJam v2 consists of 6 languages, which are also the languages with the most songs in DALI v2. For evaluation, we use the MultiLang Jamendo dataset [3]. It consists of 80 songs in 4 languages: English, French, Spanish, and German. Both training, validation and testing are performed at the utterance level.

The core experiments are conducted with a spectrogram-input baseline, which is a transformer adapted from a previous English ALT work [4]. We intentionally avoid using ALT models or feature extractors pretrained with speech data, such as wav2vec2 and Whisper, although we are aware that incorporating these could be beneficial to the performance. This is since using speech models would introduce impact from the data distribution of the pretrained models, making it difficult to estimate if any difference is attributed to training on more languages' data or the knowledge acquired during pretraining.

JH is a research student at the UKRI Centre for Doctoral Training in Artificial Intelligence and Music, supported jointly by UK Research and Innovation [grant number EP/S022694/1] and Queen Mary University of London.

EB is supported by RAEng/Leverhulme Trust Research Fellowship LTRF2223-19-106.

3. Method

Our study comprises three parts: Multilingual joint training, language conditioning, and transliteration augmentation.

Multilingual Joint Training Firstly, we highlight the impact of training on multilingual data compared to monolingual models. Low-resource language ALT suffers from having limited labelled data for training. Many languages share common phonemes, and those with the same writing system also share characters. We explore a straightforward way to utilize the commonalities among multilingual data through multilingual joint training.

We train our baseline model on subsets comprising one, four, five, and six languages. We observe that by training on more languages' data, overall WER is improved, as well as the individual WERs.

Language Conditioning Secondly, we discuss language conditioning methods, examining language embedding conditioning and FiLM conditioning. We condition our multilingual model with language information to investigate its impact. By providing the model with knowledge of the target language, it has the potential to learn language-specific features through the encoder and predict characters belonging to the target language alphabet through the decoder.

Results show that introducing language conditioning during training significantly enhances performance. Taking one step further, we extend the language-informed models to enable language-agnostic (blind) inference while retaining language configurability.

Transliteration Augmentation Thirdly, we integrate the two ideas above through transliteration augmentation. Transliteration is the process of converting text from one script into another script, while maintaining the phonetic correspondence between the original and the converted text. In the context of lyrics transcription, transliteration from one language to another could serve as a data augmentation method, as it allows us to improve low-resource language ALT with high-resource language data. Together with language conditioning, the model is equipped with the ability to generate transliteration according to the requirement for the target language.

By training on multilingual data with language conditioning and transliteration targets, we achieve superior performance and address the challenges posed by limited and imbalanced data.

4. References

[1] R. Yuan, Y. Ma, Y. Li, G. Zhang, X. Chen, H. Yin, L. Zhuo, Y. Liu, J. Huang, Z. Tian, B. Deng, N. Wang, C. Lin, E. Benetos, A. Ragni, N. Gyenge, R. B. Dannenberg, W. Chen, G. Xia, W. Xue, S. Liu, S. Wang, R. Liu, Y. Guo, and J. Fu, "MARBLE: music audio representation benchmark for universal evaluation," in *Proc. NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[2] G. Meseguer-Brocal, A. Cohen-Hadria, and G. Peeters, "Creating dali, a large dataset of synchronized audio, lyrics, and notes," *TISMIR*, vol. 3, no. 1, 2020.

[3] S. Durand, D. Stoller, and S. Ewert, "Contrastive learning-based audio to lyrics alignment for multiple languages," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[4] X. Gao, C. Gupta, and H. Li, "Polyscriber: Integrated fine-tuning of extractor and lyrics transcriber for polyphonic music," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 1968–1981, 2023.

Real-Time Spoken Language Processing in Conversational AI: Challenges and Future Directions

Arash Ashrafzadeh[1], *Arash Eshghi*[2], *Matthew Aylett*[3]

[123]The Interaction Lab, Heriot-Watt University, Edinburgh
[2]AlanaAI, Edinburgh, [3]CereProc, Edinburgh
{aa2070, A.Eshghi, M.Aylett}@hw.ac.uk

Humans process speech incrementally; meaning that we understand and generate utterances on a word-by-word basis in real-time [1]. Consequently, in order for spoken dialogue systems (also known as conversational AI) to be more robust towards natural, human-like dialogue, incremental processing (see fig. 1) is imperative in the backbone models of such systems. In addition to this cognitive motivation, processing of all the commonplace speech phenomena [2] such as turn-taking, interruptions, corrections and pauses, among others, has to happen on-the-fly to reflect their semantic effect on dialogue context, crucially in interaction with social robots or individuals with dementia [3, 4]. This abstract outlines the ongoing research within the Interaction Lab concerning this subject.

Current Research & Gaps Two primary approaches exist for dialogue processing: (1) *a grammar-based* approach where a grammar framework grounded in linguistic theory is used to parse dialogue, and (2) a *purely data-driven* approach where deep learning models, currently Large Language Models (LLMs) are utilized for dialogue processing. Since our focus is on incremental understanding and generation, for (1), an inherently incremental grammar such as Dynamic Syntax (DS, [5]) can be used directly to parse and produce utterances word-by-word [6, 7]. For (2), while LSTM-based models exhibit incremental processing capabilities, they are not state-of-the-art (SOTA) [8]; Alternatively, SOTA non-incremental LMs can be adapted in a restart-incremental manner; however, this adaptation incurs high computational complexity [9] and lacks pure incremental processing.

Both approaches have strengths and weaknesses: Grammar-based models are data-efficient and excel at compositional generalisation, yet they struggle with scalability and robustness. Conversely end-to-end neural approaches offer superior domain generalization through fine-tuning or prompt optimization, but they falter in tasks requiring planning, reasoning, compositional generalisation and control, among others. Moreover, the resource-intensive nature of data and compute power during training remains a challenge.

Future Glance Given the current landscape of NLP research, the following are some of the promising pathways towards data-efficient, scalable, compositional models for incremental dialogue processing:

- Grounded lexical learning with DS from multi-modal interaction following [10, 11] (current research)
- Further development of incremental vector-space composition following [12]
- Curation of ML-ready datasets of fundamental incremental phenomena in dialogue (see e.g. [13]);
- Using these to evaluate and fine-tune LLMs (including recent recurrent architectures [14]) in incremental settings.

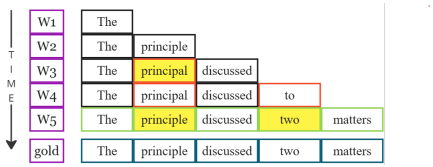


Figure 1: *An example of incremental ASR behaviour. Revokes are shown in yellow. Correct and incorrect prefixes of the gold sentence are specified in green and red*

1. References

[1] C. Howes, M. Purver, P. G. T. Healey, G. J. Mills, and E. Gregoromichelaki, "On incrementality in dialogue: Evidence from compound contributions," *Dialogue Discourse*, vol. 2, 2011.

[2] L. Burnard, "Reference guide for the british national corpus (world edition)," *October*, 2000.

[3] M. Aylett, A. Carmantini, and D. Braude, "Why is my social robot so slow? how a conversational listener can revolutionize turn-taking," 10 2022.

[4] A. Addlesee, I. Konstas, and A. Eshghi, "Current challenges in spoken dialogue systems and why they are critical for healthcare applications," in *Proceedings of the Dialogue for Good (DiGo) 2019 workshop*, 2019.

[5] R. Cann, R. Kempson, and L. Marten, *The Dynamics of Language*, 2005. [Online]. Available: https://www.researchgate.net/publication/298544659_Dynamics_of_Language_An_Introduction

[6] A. Eshghi, J. Hough, and M. Purver, "Incremental grammar induction from child-directed dialogue utterances," 6 2013.

[7] A. Eshghi and A. Ashrafzadeh, "Learning to generate and corr-uh i mean repair language in real-time," 2023.

[8] B. Madureira and D. Schlangen, "Incremental processing in the age of non-incremental encoders: An empirical assessment of bidirectional models for incremental nlu," 2020.

[9] B. Madureira, P. Kahardipraja, and D. Schlangen, "When only time will tell: Interpreting how transformers process local ambiguities through the lens of restart-incrementality," 2024.

[10] Y. Yu, A. Eshghi, and O. Lemon, "Learning how to learn: An adaptive dialogue agent for incrementally learning visually grounded word meanings," 2017.

[11] J. Mao, H. Shi, J. Wu, R. Levy, and J. Tenenbaum, "Grammar-based grounded lexicon learning," *Advances in Neural Information Processing Systems 34 pre-proceedings (NeurIPS 2021)*, 2021.

[12] M. Purver, M. Sadrzadeh, R. Kempson, G. Wijnholds, and J. Hough, "Incremental composition in distributional semantics," *Journal of Logic, Language and Information*, vol. 30, no. 2, pp. 379–406, 2021.

[13] A. Addlesee and A. Eshghi, "You have interrupted me again!: making voice assistants more dementia-friendly with incremental clarification," *Frontiers in Dementia*, vol. 3, 2024.

[14] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," 2023.

Self-supervised models for dysarthric speech: Understanding representations through visual analysis and probing

Ariadna Sanchez, Simon King

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

ariadna.sanchez@ed.ac.uk, Simon.King@ed.ac.uk

Detection of pathologies from speech has great potential in early diagnosis of neurodegenerative diseases, which can improve prognosis and care for patients. Recently, applications in pathological speech, e.g. pathology detection or speech recognition, have used self-supervised learning (SSL) models as feature extractors. In this work, we explore how SSL models represent a type of pathological speech, dysarthric speech, without explicitly fine-tuning the model to this type of speech. We compare English healthy and dysarthric speech of two different severities (mild and severe) from the TORGO database [1] through visualisation and probing techniques.

First, we visualise features extracted with the XLSR-53 SSL model [2] using t-SNE, both with utterance- and phoneme-averaged embeddings. Second, we probe each layer of the XLSR-53 model by training logistic regression classifiers that classify speech as healthy or dysarthric, similar to previous work for healthy speech [3]. Finally, we analyse the units discovered through k-means clustering, following a methodology originally applied to healthy speech [4].

Results suggest that dysarthric speech is not necessarily represented in a separable part of the latent space, but occupies a reduced part of the same space as healthy speech, as shown in Figure 1 for nasal sounds). Results from the probing task show in Figure 2 show poor performance in detecting mildly dysarthric speakers using the frame-level features extracted by XLSR-53. This could be due to the lack of granularity in the dysarthria severity levels in the dataset used: a speech therapist we consulted reported no audible dysarthric traits in the majority of mild dysarthric speakers in this database. This could be due to the participants in this database having been diagnosed but not presenting any symptoms in their speech, e.g. if their disease onset is limb-based instead of bulbar.

In future work, we aim to explore, with the methodology used, languages other than English.

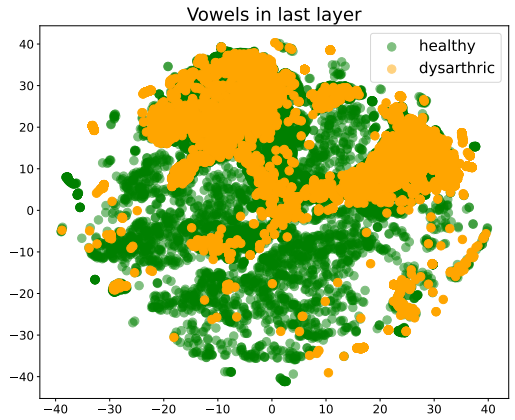


Figure 1: *Visualisation of vowel sounds with the last layer (24) of the XLSR-53 model. Each point corresponds to a two-dimensional representation of an average embedding over the aligned frames of a vowel in the dataset. Dysarthric vowel examples are plotted on top of healthy examples to compare how each type of speech spreads in the latent space.*

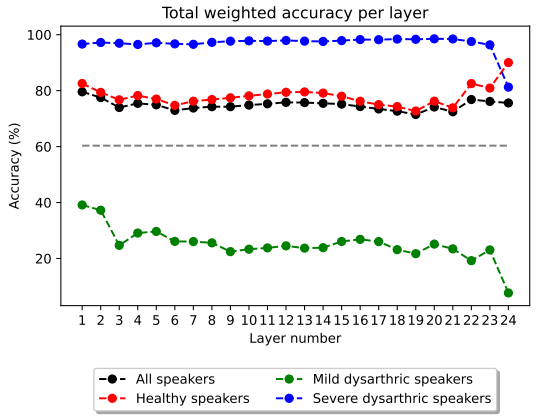


Figure 2: *Weighted accuracy results for speech type binary classification (healthy/dysarthric). The grey line indicates the chance level in this dataset.*

1. References

[1] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, pp. 523–541, 2012.

[2] A. Conneau *et al.*, "Unsupervised Cross-Lingual Representation Learning for Speech Recognition," in *Proc. Interspeech 2021*, 2021, pp. 2426–2430.

[3] M. de Seyssel *et al.*, "Probing phoneme, language and speaker information in unsupervised speech representations," in *Proc. Interspeech 2022*, 2022, pp. 1402–1406.

[4] D. Wells, H. Tang, and K. Richmond, "Phonetic analysis of self-supervised representations of english speech," in *23rd Annual Conference of the International Speech Communication Association, INTERSPEECH 2022*. ISCA, 2022, pp. 3583–3587.

How Much Context Does My Attention-Based ASR System Need?

Robert Flynn, Anton Ragni

Department of Computer Science, University of Sheffield, 211 Portobello, Sheffield S1 4DP, UK
rflynn2@sheffield.ac.uk, a.ragni@sheffield.ac.uk

Abstract

For the task of speech recognition, the use of more than 30 seconds of acoustic context during training is uncommon and under-investigated in literature. In this work, we conduct an empirical study on the effect of scaling the sequence length used to train/evaluate (dense-attention-based) acoustic models on speech recognition performance. For these experiments, a dataset of roughly 100,000 pseudo-labelled Spotify podcasts is used, with context lengths of 5 seconds to 1 hour being explored. Zero-shot evaluations are presented on the long-format datasets: Earnings-22, Tedlium and Rev16. Results demonstrate a benefit from training with up to 21.8 minutes of acoustic context, showing up to a 14.5% relative improvement from a baseline trained with 10 seconds of context. We find that the model's width/depth, positional encoding scheme and number of attention heads impact its ability to use longer contexts.
Index Terms: speech recognition, long-context, self-attention

1. Experimental Overview

For this work, a CTC [1] Conformer [2] based model used and trained on a dataset of Spotify podcasts [3] with context lengths ranging from 10 seconds to 1 hour of audio. Context length refers to the length of audio that is used as input to the model. We then evaluate each of the models zero-shot on a selection of datasets at the context length they were trained at, and examine how the performance changes as the context size is increased. The datasets used for evaluation are: Earnings-22 [4], Tedlium [5] and Rev-16 (originally used in [6].

2. Results

The following section overviews some of our key findings.

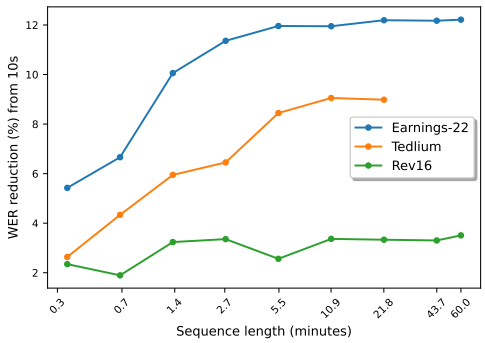


Figure 1: *WER reduction compared to a model using sequence length of 10s*

As shown in figure 1, the models benefit from up to 20 minute of context on Earnings-22, our most challenging dataset. On both Earnings-22 and Tedlium most of the benefit was seen from having at least 5 minutes of context with gains of up to around 12% compared to a baseline trained/evaluated with 10 seconds of context.

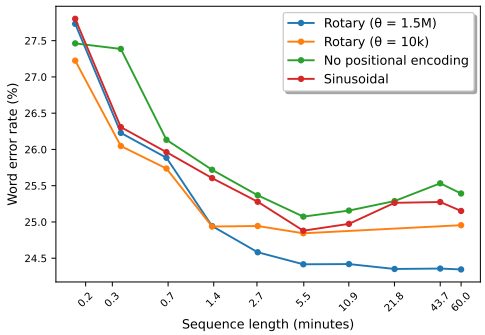


Figure 2: *WER at various sequence length for different positional encoding methods on Earnings-22*

From figure 2 we can see that the choice of positional encoding is important when working with longer contexts. At short sequence lengths, all choices of positional encoding explored perform similarly. However, rotary encodings with $\theta = 1.5\text{M}$ (θ controls the distance bias) was the only setting to fully benefit from sequence lengths greater than 5 minutes.

3. References

[1] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006.

[2] A. Gulati, J. Qin, C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[3] 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 *COLING*. ICCL, Dec. 2020, pp. 5903–5917. [Online]. Available: https://aclanthology.org/2020.coling-main.519

[4] M. Del Rio, P. Ha, Q. McNamara, C. Miller, and S. Chandra, "Earnings-22: A practical benchmark for accents in the wild," *arXiv preprint arXiv:2203.15591*, 2022.

[5] F. Hernandez, V. Nguyen, S. Ghannay, N. Tomashenko, and Y. Esteve, "Ted-lium 3: Twice as much data and corpus repartition for experiments on speaker adaptation," in *SPECOM*. Springer, 2018, pp. 198–208.

[6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *ICML*. PMLR, 2023, pp. 28 492–28 518.

Focused Discriminative Training For Streaming Automatic Speech Recognition

Adnan Haider[*1], *Xingyu Na*[2], *Tim Ng*[2], *Zhen Huang*[2], *and Xiaodan Zhang.*[2]

[1]Apple Inc, UK
[2]Apple Inc, USA

adnan_haider@apple.com

The combination of connectionist temporal classification (CTC) [1] and attention-based encoder-decoder (AED) [3, 2, 10] loss has become a powerful approach in training end-to-end (E2E) models to address both streaming and non-streaming speech recognition tasks [5].

This work aims to address the challenging task of further improving a streaming word piece ASR model that has converged with respect to CTC training. For neural network models whose output units represent phonetic units, sequence discriminative training approaches such as MMI, sMBR, EMBR, etc. [4, 8, 9], which embed the models within an Hidden Markov Model(HMM) framework, have been shown to be effective in achieving further WER improvements. However, the implementation of discriminative sequence training is a time-consuming and challenging process. Implementations of both lattice-based [4, 7] and lattice-free sMBR/MMI [6] techniques necessitate careful design decisions regarding the choice of HMM topology, lexicon, and graph generation. This results in a slower evolution of the training pipeline compared to the rapidly evolving training frameworks for E2E models. In particular, lattice-based MMI introduces an additional computational overhead by requiring pre-generated lattices, potentially diminishing the training framework's flexibility. Furthermore, in the LF-MMI paper [6], the authors also acknowledge the difficulty of further improving models trained with CTC and thus employ their discriminative training approach to fine-tune models trained with Cross-Entropy loss. The current variant of LF-MMI trains models from scratch. It should be mentioned that aside from the mentioned approaches, MWER [10] presents an HMM-independent approach that is well-established in training both streaming and non-streaming models. However, there is no evidence in the literature suggesting that it can be applied in conjunction with CTC loss. Our attempts to use MWER to enhance models trained with CTC were unsuccessful.

This work introduces an alternative novel approach called Focused Discriminative Training (FDT) to fine-tune models trained using CTC or an interpolation of CTC and AED (CTC+AED) loss. The algorithm consists of two essential components: implicit data selection coupled with an error region detection system and a CTC-based contrastive segment-level loss. The first component enables the algorithm to focus on challenging utterances. For such utterances, the error detection component aims to identify the regions where mis-recognition occurs. Subsequently, by employing the contrastive segment-level loss, the algorithm aims to enhance recognition performance within these identified error regions. The proposed method is HMM-independent, eliminating the requirement for various components, like HMM topology and lexicon, and utilizes N-best lists. Thus, the method is also lattice-free.

The efficacy of the algorithm is demonstrated by its ability to improve streaming models trained on the LibriSpeech dataset trained with CTC and CTC+AED loss for a fixed number of epochs. Subsequent fine-tuning for an additional training epoch reveals that FDT achieves a greater reduction in WER compared to additional fine-tuning by CTC or MMI. This study also assesses the algorithm's efficacy in improving a converged model trained on a large-scale dataset from a distinct domain. Results demonstrate that when applied to further fine-tune a converged word piece streaming E2E model that was trained on a dataset comprising 600k hours of assistant and dictation data with CTC+AED loss, the algorithm achieves additional improvements in WER.

* corresponding author

1. References

[1] A. Graves, S. Fernandez, F.Gomez and H. Schmidhuber "Connectionist temporal classification: labelling unsegmented sequence data with re- current neural networks" in *Proc. ICML*, 2006, pp. 369–376.

[2] H. Miao, G. Cheng, C. Gao, P. Zhang and Y.Yan "Transformer based online ctc/attention end-to-end speech recognition architecture" in *Proc. ICASSP*, 2020, pp 6084–6088.

[3] S. Karita NEY Soplin, S Watanabe, M Delcroix, A Ogawa and T Nakatani, "Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration" in *Proc. Interspeech*, 2019, pp. 1408–1412.

[4] P.C. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition," *Computer Speech and Language*, vol. 16, no. 1, pp. 25–47, 2002.

[5] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi. "Hybrid CTC/attention architecture for end-to-end speech recognition." *IEEE Journal of Selected Topics in Signal Processing* , vol 11, no. 8, pp 1240-1253, 2017.

[6] D. Povey, V.Peddinti, D. Galvez, P. Ghahremani, V.Manohar, X. Na , Y.Wang and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Proc. Interspeech*, 2016.

[7] D. Povey and P.C. Woodland, "Minimum phone error and i-smoothing for improved discriminative training," in *Proc. ICASSP*, 2022, pp. 105–108.

[8] B. Kingsbury, T. Sainath and H. Soltau, "Scalable minimum bayes risk training of deep neural networks acoustic models using distributed hessian-Free optimization," in *Proc. InterSpeech*, 2013, pp. 10–13.

[9] M. Shannon, "Optimizing expected Word Error Rate via Sampling for Speech Recognition," in *Proc. InterSpeech*, 2017, pp. 10–13.

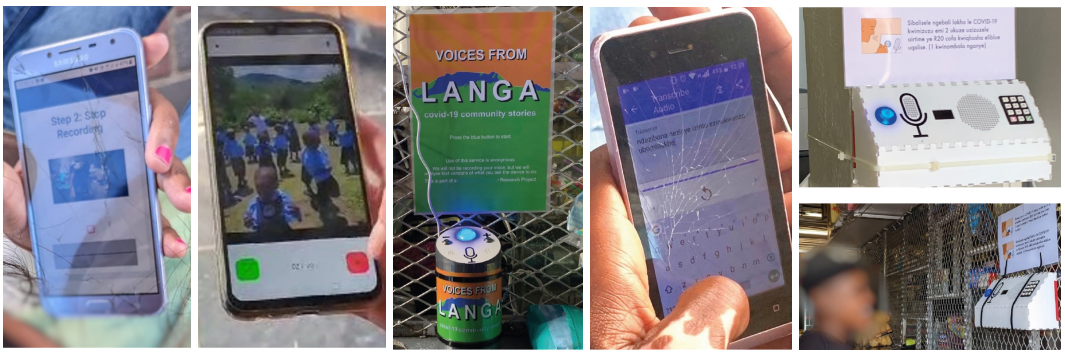
[10] R. Prabhavalkar, T. N. Sainath, Y. Wu, P. Nguyen, Z. Chen, C.-C. Chiu, and A. Kannan, "Minimum word error rate training for attention-based sequence-to-sequence sodels," in *Proc. ICASSP*, 2018, pp. 4839–4843.

Speech Interactions Designed With Minoritised Language Speakers

Emily Nielsen[1], *Electra Wallington*[2], *Ondřej Klejch*[2], *Dani Kalarikalayil Raju*[3], *Nina Markl*[4], *Gavin Bailey*[1], *Thomas Reitmaier*[1], *Jennifer Pearson*[1], *Matt Jones*[1], *Peter Bell*[2], *Simon Robinson*[1]

[1]Swansea University, UK [2]University of Edinburgh, UK
[3]Studio Hasi, India [4]University of Essex, UK

e.e.nielsen@swansea.ac.uk, electra.wallington@ed.ac.uk, o.klejch@ed.ac.uk, daniel@studiohasi.com, nina.markl@essex.ac.uk, g.bailey@swansea.ac.uk, thomas.reitmaier@swansea.ac.uk, j.pearson@swansea.ac.uk, matt.jones@swansea.ac.uk, peter.bell@ed.ac.uk, s.n.w.robinson@swansea.ac.uk



(a) (b) (c) (d) (e)

Figure 1: *Prototypes for co-creating and situating spoken-language technologies. (a-b) Speech-based media retrieval application (a) audio data collection and (b) retrieval application; (c) A speech probe deployed to engage minoritised language speakers in voice data-collection; (d) TranscriptTool interface to involve community members in the transcription of collected speech responses (e) The SpeechBox system to gather spoken community responses on their experience of COVID-19 lockdowns.*

Speech is one of the richest, and most natural, forms of communication. As such, speech-based interactive systems can facilitate easier interactions with technology [1]. However, speech-based systems are not currently available for the vast majority of languages, leaving diverse language communities digitally 'unheard': from those who speak vernacular languages that are not commonly written to those that are entirely unwritten [2]. The dearth of representative datasets, as well as communication practices that feature rich and unconstrained codeswitching [3], serve to further entrench this digital and linguistic inequality and exclude millions of people worldwide.

Our research project – UnMute – aims to empower underserved voices from 'low-resource' language communities. This project takes an interdisciplinary approach with researchers in Human-Computer Interaction (HCI), Natural Language Processing (NLP), and Linguistics to engage with, co-design, and develop technology for the digitally under-heard. In particular, we curate a portfolio of probes and prototypes that progressively vary on the theme of co-creating and situating spoken-language technologies to suit the needs and ways of speaking of diverse, resource-constrained, and under-heard, communities in South Africa [3, 4] and India [5]. These prototypes also demonstrate the benefits of human-centred methodologies, co-creating datasets and use-cases directly with communities, and ultimately illustrate how language technologies can broaden digital participation in communities of minoritised language speakers.

1. References

[1] C. Munteanu and G. Penn, "Speech-based interaction: Myths, challenges, and opportunities," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 1196–1199. [Online]. Available: https://doi.org/10.1145/3027063.3027117

[2] S. Bird and D. Yibarbuk, "Centering the Speech Community," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Y. Graham and M. Purver, Eds. St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 826–839.

[3] T. Reitmaier, E. Wallington, D. Kalarikalayil Raju, O. Klejch, J. Pearson, M. Jones, P. Bell, and S. Robinson, "Opportunities and Challenges of Automatic Speech Recognition Systems for Low-Resource Language Speakers," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, ser. CHI '22. New York, NY, USA: Association for Computing Machinery, Apr. 2022, pp. 1–17.

[4] T. Reitmaier, E. Wallington, O. Klejch, N. Markl, L.-M. Lam-Yee-Mui, J. Pearson, M. Jones, P. Bell, and S. Robinson, "Situating Automatic Speech Recognition Development within Communities of Under-heard Language Speakers," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, ser. CHI '23. New York, NY, USA: Association for Computing Machinery, Apr. 2023.

[5] T. Reitmaier, D. K. Raju, O. Klejch, E. Wallington, N. Markl, J. Pearson, M. Jones, P. Bell, and S. Robinson, "Cultivating Spoken Language Technologies for Unwritten Languages," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, ser. CHI '24. New York, NY, USA: Association for Computing Machinery, May 2024.

MyChat: a "written speech" corpus rich in conversational features

Mai Hoang Dao, Catherine Lai, Peter Bell

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

mai.hoang.dao@ed.ac.uk, c.lai@ed.ac.uk, peter.bell@ed.ac.uk

Abstract

Conversational speech is rich in phenomena that augment the purely lexical content of the interaction, enhancing understanding of the message or providing insight into the speakers' cognitive state. However, such features are not confined to spoken communication alone but also appear in text-based communication between humans, particularly in instant-messaging, a style of interaction that we term "written speech" on account of its many similarities to the spoken form. We believe it is important to better understand the relationship between these two forms of communication, and to this end we present the first "written speech" corpus comprising instant messages between friends, which we have specifically designed to explore patterns of conversational phenomena in written form. We find that it aligns more closely with conversational speech corpora than existing text-based chat datasets in respect of the incidence of features such as laughter and filled pauses.

Table 1: *Statistics of typing disfluency in our dataset. m denotes messages. AS denotes "active status". TD denotes "typing disfluency".*

Statistics	Counts
total messages	2279
TD, no AS change	1291
TD, AS change	29
delays	633
deletions	1141
insertions	122
all three TD types	64

Table 2: *Example of a typed disfluency sequence. Delay, deletion and insertion might occur multiple times during the typing process.*

Type	Examples
	They're friendly and loveable
delay	They're friendly and loveable
deletion	friendly and loveable
insertion	Dogs are friendly and loveable
insertion	Dogs are loyal and friendly and loveable

A: ya know that highlighter I have that has two tones?
B: ...mmaybe?? not sure
B: is it like two tones at the same time or one on either end
A: it's in a little tube cause its liquid
B: ohhh like makup highliter
A: yeah, at the same time
B: lmaoooo
A: hahaha yesssss
B: hahahhaa <3

A: yay! the babys was so cute pahah
B: It was!!
B: Although its hands must have hurt
B: Ahhhh the rugby!!!
B: Wooooooo
B: Wooooooooooooo
A: awww pahaha imagine having to clean it up
B: Omg yeah, that must have been annoying

Figure 1: *Examples of conversational patterns in our dataset. We add red for laughter, green for emojis, blue for intentional misspellings, and pink for filled pauses.*

The UKIS 2024 Committee would like to thank ISCA and Interspeech 2023 organisers for kindly providing this template.

Discovery of Grammatical Matrix Language Markers in Code-Switched Text

Olga Iakovenko, Thomas Hain

The University of Sheffield, UK

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

1. Introduction

Code-switching (CS) is the process of speakers switching between several languages in spoken or written language. In order to better describe the process of code-switching the Matrix Language Frame (MLF) theory was formulated [1]. It introduced the concept of a main, i.e. dominant language and a secondary, inserted language to describe CS utterances. These languages are called Matrix Language (ML) and Embedded Language (EL), respectively. The MLF theory introduces two methods for ML determination: *The Morpheme Order Principle* - ML will provide the surface morpheme order for a CS utterance, if it consists of singly occurring EL lexemes and any number of ML morphemes; and *The System Morpheme Principle* - all system morphemes which have grammatical relations external to their head constituent will come from ML.

In this work a novel system for predicting the ML of CS text based on the System Morpheme Principle of the MLF theory is introduced. A system was trained to determine Parts of Speech and Language Identity (POS+LID) tags on simulated CS text. The resulting POS+LID tags were used to train an ML determiner yielding 98% accuracy on a simulated CS dataset. The system was further tested on the SEAME corpus transcriptions and yielded a Matthew's Correlation Coefficient value of 0.66 when compared to another rule-based ML determination framework. Feature importance analysis shows that the POS-based features that most influence the decision of the system ML for Mandarin/English CS are Determiners for both languages, Proper Nouns, Auxiliaries and Verbs for Mandarin, and Subordinating Conjunctions, Interjections and Adverbs for English. The resulting methods can accurately perform LID or dominant language detection of multilingual and CS utterances and help gain insight into the speaker's manner of language mixing.

2. Methods for ML determination

2.1. Morpheme Order Principle (MOP)

The ML provides the context for singly occurring words from the EL. Suppose $\mathbf{x}$ - CS utterance, ε - empty morpheme, then $\mathbf{x}' = [\varepsilon, x_1, .., x_n, \varepsilon]$ can be split into subsequences of morphemes $[x_i, .., x_{j-1}, x_j, .., x_{k-1}, x_k, .., x_o]$
$l_i, .., l_{j-1}, l_k, .., l_o \in L_1, L_\varepsilon, l_j, .., l_{k-1} \in L_2$ - corresponding LIDs, and $[x_i, .., x_{j-1}]$, $[x_j, .., x_{k-1}]$, $[x_k, .., x_o]$ - constitute words. In this case the language of the context L_1 is ML while L_2 is EL.

Example: 开眼睛了 melissa 讲什么

2.2. System Morpheme Principle (SMP)

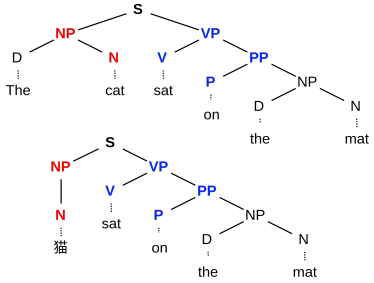
ML is the language that provides the system morphemes for a CS utterance $\mathbf{t}$. Suppose then that $\mathbf{g} = [g_1, .., g_n]$, $g_i \in G$ - grammatical categories of morphemes $\mathbf{t} = [t_1, .., t_n]$, $t_i \in T_{sys} \cup T_{cont}$ - morpheme types according to the 4-M model, $\mathbf{x} \rightarrow \mathbf{g} \rightarrow \mathbf{t}$ and $\mathbf{x} \rightarrow \mathbf{l}$, meaning that $P(L|\mathbf{t}, \mathbf{l}, \theta) = P(L|\mathbf{g}, \mathbf{l}, \theta)$.

Example: im okay with the 蛋黄

2.3. Feature importance

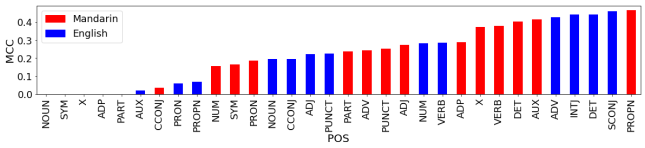
$P_t(L|g, l, \theta_t)$ - model trained with one grammatical category + LID as an input feature (g_t, l_t). Then for dataset $D_t = [(g_1, l_1, L_1), .., (g_m, l_m, L_m)]$ one can calculate feature importance f_t: $f_t = \prod_{i=1}^{|D|} P_t(L = L_i|g_i, l_i, \theta)$

2.4. Synthetic CS generation



3. Results

POS+LID joint task prediction: 94% and 93% Accuracy, 94% and 92% F1 for validation and test set respectively. MCC between MOP and SMP determiners on the test subset of real CS data using single-feature DT:



4. References

[1] C. Myers-Scotton, *Duelling Languages: Grammatical Structure in Codeswitching*. Clarendon Press, 1997. [Online]. Available: https://books.google.co.uk/books?id=NuYdnTyKkdQC

Behavioral evidence for higher articulation rate convergence following natural than artificial time altered speech

Jérémy Giroud[1], *Jessica Lei*[1], *Kirsty Phillips*[1], *Matthew H. Davis*[1]

[1]MRC Cognition and Brain Sciences Unit, University of Cambridge, UK

Jeremy.Giroud@mrc-cbu.cam.ac.uk, Matt.Davis@mrc-cbu.cam.ac.uk

Abstract As AI progresses, our exposure to artificially generated spoken content varying in naturalness and speed increases. This trend is amplified by AI-powered personal assistants' proliferation, multiplying our interactions with intelligent systems. Research is crucial to understand if phenomena observed in human-human interactions can inform these new interactions. For instance, in everyday conversation, people adjust their speaking pace to match their partner's, a phenomenon known as speech rate convergence [1]. It is crucial for effective communication, occurs automatically and is present in more artificial interaction scenarios. In this study, we investigated how the nature (natural vs. artificial) and the presentation rate (normal vs. fast) of the speech signal impact speech rate convergence. Data from 116 participants across two experiments reveal higher convergence towards naturally produced speech compared to artificially time altered speech. Implications for human-machine interactions are discussed.

Methods 61 and 55 participants took part in a speech perception and production experiment composed of 5 and 4 tasks respectively. The design allowed to assess participants speech rate convergence to naturally/artificially produced speech at both normal and fast presentation rates. We also recorded participants' speech rate following a convergence period which enabled to investigate long term changes on participants' speech production. In experiment 2, we removed a listening task which served as a perceptual adaption period enabling participants to get familiar with the acoustic characteristic of the spoken material. Spoken material was taken from the CHAINS corpus containing sentences produced at both normal and fast (54% faster) speed by the same speaker [2]. Artificial fast speech was created through linearly compressing the natural normal stimuli, and artificial normal speech was created through linearly expanding the natural fast stimuli. The resulting stimuli had the same articulation rates while being of different nature (artificial/natural). Participants' articulation rates were computed for each condition and entered into a linear mixed effects analysis.

Results Our results indicate that audio recordings are sufficient to elicit speech rate convergence. Moreover, the magnitude of the convergence is dependent on the presentation rate and naturalness of heard speech as illustrated in fig1. We also found evidence for modulation of participants' articulation rate beyond the initial convergence period; i.e. sustained speech-rate after-effects (fig2). Finally, convergence depends on participants' familiarity with the specific characteristics of the speech signal acquired during a perceptual adaptation period.

Discussion Overall, our findings contribute to the broader scientific understanding of speech convergence. Our results suggest that speakers apply principles from human-human interaction while engaging with artificially generated speech. Critically, our data shows that participants interact in distinct ways depending on the rate and naturalness of the speech they are hearing. The findings from this study also have real-life implications for the design of voice-controlled human-machine interfaces and AI assistants. This study suggests that mimicking the spectro-temporal features of natural speech may yield better engagement from individuals.

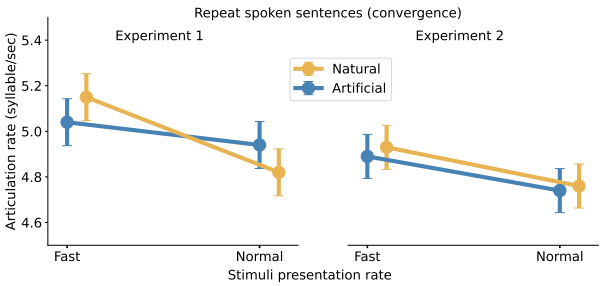


Figure 1: *Speech rate convergence from both experiments. Articulation rate data were extracted during the repeat spoken sentences task from n = 61 and n = 55 participants. Error bar represents standard error to the mean.*

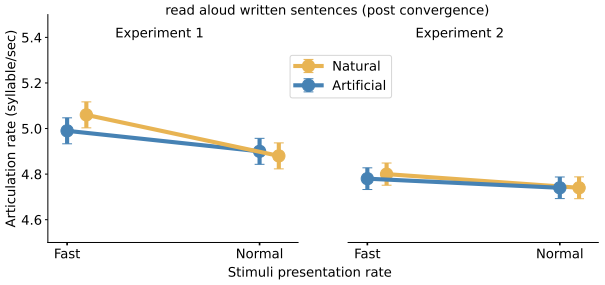


Figure 2: *post convergence after effects from both experiments. Articulation rate data were extracted during the read aloud written sentences task from n = 61 and n = 55 participants. Error bar represents standard error to the mean.*

1. References

[1] D. Freud, R. Ezrati-Vinacour, and O. Amir, "Speech rate adjustment of adults during conversation," *Journal of fluency disorders*, vol. 57, pp. 1–10, 2018.

[2] F. Cummins, M. Grimaldi, T. Leonard, and J. Simko, "The chains speech corpus: Characterizing individual speakers," in *Proc of SPECOM*, 2006, pp. 1–6.

Dynamic Time Warping as an Alignment Technique for Voice Puppetry

Ella Crocker

CereProc Ltd, Edinburgh

ella@cereproc.com

Expressivity and natural-sounding prosody are challenges for text-to-speech (TTS); synthetic voices can sound monotonous or have prosodic realisations which are not context-appropriate. Voice puppetry is a technique that uses a reference voice recording to direct the prosody of TTS, enabling the creation of expressive synthetic speech which is finely controllable by nonexperts.

Voice puppetry requires an accurate, time-aligned phonetic transcription of the reference recording in order to extract pitch and energy information on the phone level and apply this when generating synthetic speech. This typically relies on an alignment system pre-trained on the language variety being used. Dynamic Time Warping (DTW) was suggested as a language-independent solution for generating a transcription for an unlabelled reference recording in [1]. A TTS output waveform with a time-aligned transcript and the unlabelled reference are each converted to MFCC vectors, and a DTW alignment between them determines how the transcript is copied to the reference recording, shown in Figure 1.

To obtain a quantitative measure of alignment accuracy, the DTW-generated transcript and a Hidden Markov Model (HMM)-generated transcript were compared against manual annotations for the CSTR US KED Timit dataset [2]. A similarity measurement identified mutually closest phone boundary pairs as matches, assuming that they demarcate the same phone transition regardless of label (due to phoneset differences). The DTW alignment matched 78% of predicted phone labels to human annotations, while the HMM method matched 89% of phone labels. These results should be considered with caution given the inherent subjectivity of manual annotations.

Qualitative evaluation consisted of a MUSHRA-style listening test comparing synthesised speech samples from four conditions: the two voice puppetry methods with DTW- and HMM-generated transcriptions, unmodified TTS output, and tuned TTS output for which phone-level pitch, durations and energy were manually edited to sound appropriate. Short texts were synthesised from 3 different genres: a wedding speech, a sonnet, and some newspaper text. The reference speaker was a male RP speaker and the TTS voice was chosen to match.

Participants rated the DTW alignment condition as significantly less natural than the HMM alignment condition, suggesting that the difference in alignment accuracy had a strong effect on the naturalness of the resulting speech. Both puppetry methods were rated as significantly less natural than tuned and unmodified TTS. This may be due to alignment errors in both puppetry conditions resulting in unnatural extracted pitch and duration values (although without human annotations for the reference speaker audios, this could not be confirmed).

The tuned TTS did not significantly differ from the unmodified TTS, despite the fact that it had been deliberately edited to sound more expressive and genre-appropriate. Preferences may differ between the participants and the editor, or sentences presented in isolation may lack context, and thereby create a preference for more neutral prosody.

Only the DTW condition exhibited a significant interaction of the genre of the text, with synthesised newspaper text being preferred over the sonnet and wedding speech genres. The more neutral rendition of this by the reference speaker may have been more similar to the TTS audio, resulting in fewer alignment errors.

1. References

[1] E. Van de Vreken, K. Richmond, and C. Lai, "Voice puppetry with FastPitch," in *Proceedings of the Annual Conference of the International Speech Communication Association*, ser. Interspeech, vol. 2022-September. ISCA, Sep., pp. 5219–5220. [Online]. Available: https://interspeech2022.org/

[2] Carnegie Mellon University and Edinburgh University, "CSTR US KED Timit," http://festvox.org/dbs/dbs_kdt.html, 1997-2002.

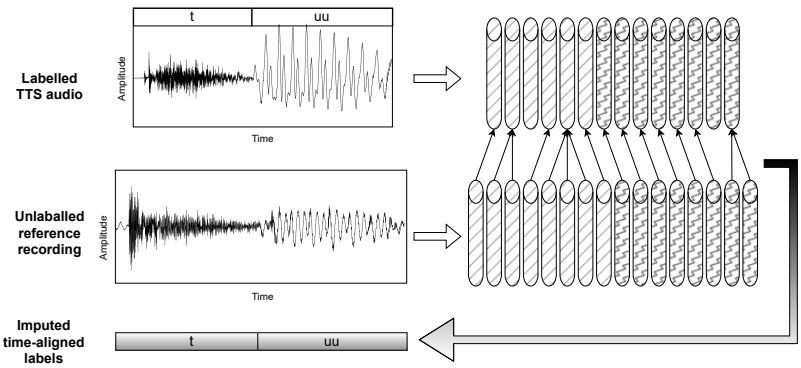


Figure 1: *Diagram illustrating the use of DTW to impute labels for unlabelled audio*

Bias in Sparse Subnetworks for Multilingual Automatic Speech Recognition

Edward Storey[1,2], *Naomi Harte*[2], *Peter Bell*[1]

[1]School of Informatics, The University of Edinburgh, UK
[2]Sigmedia Lab, ADAPT Centre, School of Engineering, Trinity College Dublin, Ireland
storeyed@tcd.ie, nharte@tcd.ie, peter.bell@ed.ac.uk

Large multilingual Automatic Speech Recognition (ASR) models are trained on hundreds of thousands of hours of data over hundreds of millions of parameters to achieve low error rates on many different languages within a single model. However, many of these large models are over-reliant on high-resource languages, resulting in worse results on low-resource languages. We concentrate on XLS-R for this study, which has been trained on 128 different languages, although the amount of data for each language in pretraining is greatly unbalanced [1]. The 24 highest-data languages that XLS-R was trained on contain 98% of total data and English alone accounts for 15.9%.

To explore the effect the imbalance of training data has had on XLS-R we utilise the Lottery Ticket Hypothesis (LTH) [2]. Through LTH we can identify language-specific subnetworks [3] and test the effectiveness of each language separately. To do this we train these subnetworks on various downstream languages and measure the Character Error Rate (CER). Figure 1 shows the CERs of the English subnetwork in XLS-R when trained on other high-resource languages. We compared these results to other language-specific subnetworks and found that English gave consistently lower error when compared to other language-specific subnetworks. We also see that error does not increase substantially across any languages until approximately 70% sparsity. As such we concentrate on sparsities of 70%, 80% and 90% in Figure 2.

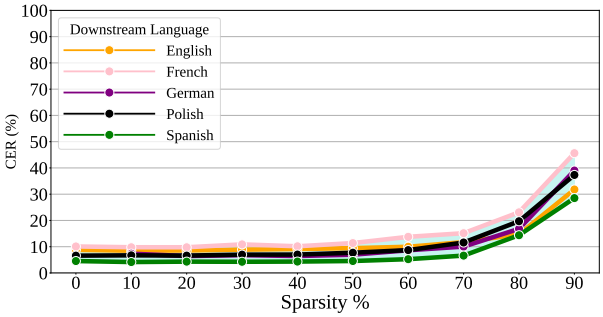


Figure 1: ***English subnetworks to multiple downstream Languages*** *English subnetworks are fine-tuned to downstream English, French, German, Polish and Spanish while pruning from 0% up to 90% sparsity*

Figure 2 shows the average CER when multiple language-specific subnetworks are trained on languages that aren't their own. We see in Figure 2 that the English subnetwork performs better than any of the other subnetworks tested at high sparsities. Further work showed that training from an English subnetwork could yield comparable or even lower CERs compared to training the whole model on the same data.

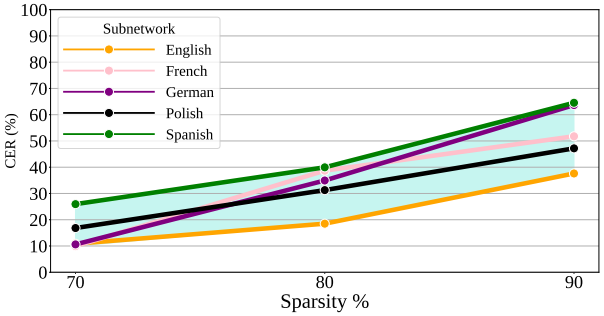


Figure 2: ***Mean average results for each language-specific subnetwork when tested on all other downstream languages*** *each subnetwork is fine-tuned to the four other languages, these results are then averaged and plotted at 70%, 80% and 90% sparsity*

English is the highest resource language in the XLS-R pretraining data and appears to have a prominent effect on downstream training. Our work has shown that when training on a new language it is more optimal to train from an English subnetwork than a subnetwork from any other language. It can even be more optimal than fine-tuning the whole XLS-R model.

1. References

[1] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," 2021.

[2] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," in *International Conference on Learning Representations*, 2019.

[3] Y. Lu, M. Huang, X. Qu, P. Wei, and Z. Ma, "Language adaptive cross-lingual speech representation learning with sparse sharing sub-networks," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6882–6886.

SALMONN: Towards Generic Hearing Abilities for Large Language Models

Guangzhi Sun[1]

[1]University of Cambridge

gs534@cam.ac.uk

1. Introduction

Hearing is arguably an essential ability of artificial intelligence (AI) agents in the physical world, which refers to the perception and understanding of general auditory information consisting of at least three types of sounds: speech, audio events, and music. In this paper, we propose SALMONN, a speech audio language music open neural network, built by integrating a pre-trained text-based large language model (LLM) with speech and audio encoders into a single multimodal model. SALMONN enables the LLM to directly process and understand general audio inputs and achieve competitive performances on several speech and audio tasks used in training, such as automatic speech recognition and translation, auditory-information-based question answering, emotion recognition, speaker verification, and music and audio captioning *etc.* SALMONN also has a diverse set of emergent abilities unseen in the training, which includes but is not limited to speech translation to untrained languages, speech-based slot filling, spoken-query-based question answering, audio-based storytelling, and speech audio co-reasoning *etc*. The presence of cross-modal emergent abilities is studied, and a novel few-shot activation tuning approach is proposed to activate such abilities. To our knowledge, SALMONN is the first model of its type and can be regarded as a step towards AI with generic hearing abilities.

2. SALMONN Architecture

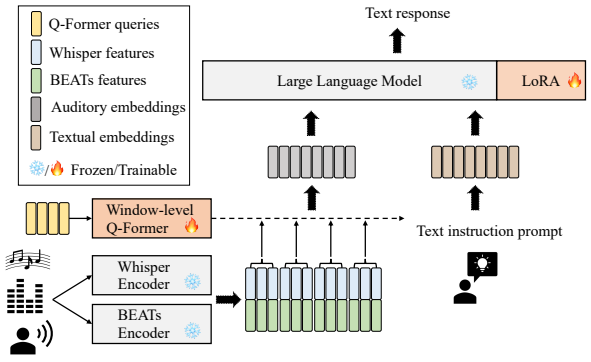


Figure 1: *The model architecture of SALMONN.*

The architecture of SALMONN is shown in Fig. 1. The output features of the two complementary auditory encoders are synchronised and combined. Q-Former is used as the connection module and applied at the frame level, whose output sequence is integrated with the text instruction prompt and fed into the LLM with LoRA adaptors to generate the text response.

3. SALMONN Training

A three-stage cross-modal training method of SALMONN is introduced in this section. Besides the pre-training and instruction tuning stages used by recent visual LLMs, an additional activation tuning stage is proposed to resolve the issue of over-fitting to the speech recognition and audio captioning tasks in instruction tuning.

Task Over-fitting: Although SALMONN built with only the first two stages of training can produce competitive results on the tasks trained in instruction tuning, it exhibits limited or almost no ability to perform untrained cross-modal tasks, especially the tasks that require cross-modal co-reasoning abilities. In particular, the model sometimes violates the instruction prompts and generates irrelevant responses as if it received an instruction related to a task commonly seen in training (*e.g.* speech recognition or audio captioning). We refer to this phenomenon as *task over-fitting*

Activation Tuning Stage: An effective approach to alleviating task over-fitting is to regularise the intrinsic conditional LM $P_{\Lambda}(\mathbf{Y}|\mathbf{X})$. An easy way to achieve this is to fine-tune SALMONN on tasks with longer and more diverse responses, such as auditory-information-based question answering and storytelling. Paired training data for such tasks can be generated based on the text paired with speech recognition or audio and music caption data, either manually by human annotators or automatically by prompting a text-based LLM.

4. Results

Task	ASR (WER)	AAC (CIDEr)	EN2Zh (BLEU4)
w/o Activation	2.1	34.4	47.6
w/ Activation	2.1	33.1	40.3

Table 1: *Results on trained tasks.*

Task	ER (Acc.)	Music Caps. (CIDEr)	SV (Acc.)
w/o Activation	63%	22.1	93%
w/ Activation	2.1	33.1	40.3

Table 2: *Zero-shot performance on audio or speech tasks.*

Task	Accuracy	Following Rate
w/o Activation	2%	4%
w/ Activation	50%	73%

Table 3: *Speech-audio co-reasoning abilities.*

Speech Synthesis for Sarcasm in Low-Resource Scenarios

Zhu Li, Yuqing Zhang

University of Groningen, the Netherlands
{zhu.li, yuqing.zhang}@rug.nl

Sarcasm is a unique form of communication that involves expressing the opposite of what is intended, often accompanied by exaggerated intonation, emphasis, and unexpected pauses. Traditional text-to-speech (TTS) models, trained on general speech data, often struggle to capture the subtleties and expressive nature of sarcasm. The limited availability of sarcastic speech datasets further exacerbates this challenge. Beyond data limitations, synthesizing sarcastic speech presents a considerable challenge due to the distinct characteristics that differentiate it from neutral speech. To tackle these challenges, we explore the feasibility of transfer learning and Large Language Models (LLMs) for data collection and annotation to develop more natural and expressive sarcastic speech synthesis systems capable of capturing the intricacies of sarcasm in speech.

1. Transfer learning

Figure 1 shows how we leveraged knowledge transfer on TTS by fine-tuning a pre-trained model on the sarcasm dataset. We first pre-trained the Fastspeech 2 model on a multi-speaker reading style dataset, LibriTTS, and then fine-tuned it using a multi-speaker sarcastic dataset, MUStARD [1]. We used Hifi-GAN universal as the vocoder and fine-tuned it on the MUStARD to make it compatible with laughter, which is an important feature of sarcasm. Due to the background noise and laughter, the Mean Opinion Score (MOS) of real data only approximately reaches 3.9. The MOS of synthesized sounds from ground-truth mel are lower than those from the fine-tuned vocoder ($3.38 < 3.82$), suggesting that the fine-tuned vocoder can adapt to laughter. The MOS of sounds generated by fine-tuned Fastspeech 2 and vocoder is much lower than those synthesized by fine-tuned vocoder only ($2.84 < 3.82$), indicating that transfer learning alone can't achieve a satisfactory performance.

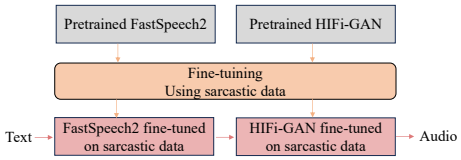


Figure 1: *Pipeline for fine-tuning*

2. Sarcastic data collection by LLMs

We proceeded to synthesize sarcastic speech by pre-training a TTS model using data annotated by LLMs. Specifically, we curated a large dataset of sarcastic textual data identified by LLMs, extracted corresponding speech data, and then trained a model capable of synthesizing sarcastic speech (see Figure 2).

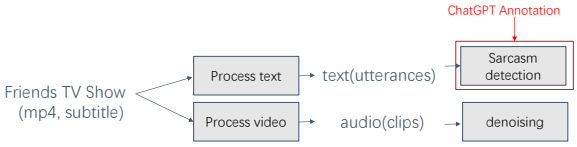


Figure 2: *Pipeline for data collection*

Table 1 shows sarcasm detection results from LLM annotations conducted on the first 100 samples from MUStARD.

LLMs	Recall	Precision	F1 score	Kappa
ChatGPT 3.5	.778	.823	.800	.579
ChatGPT 4	.814	.830	.822	.618
Llama 2-70B	.796	.605	.688	.193

Table 1: *Result of sarcasm detection*

After we obtained the sarcasm labels, we trained a sarcastic speech synthesis model, as shown in Figure 3.

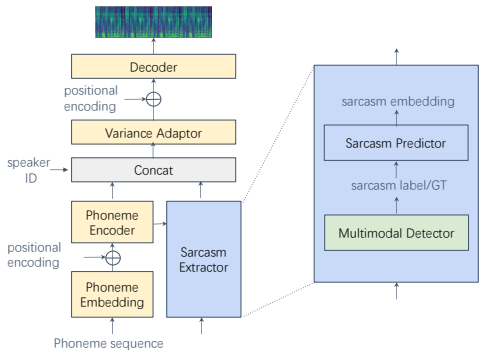


Figure 3: *Architecture of controllable sarcastic speech synthesis*

This model consists of a Fastspeech 2 model, and a sarcasm extractor, which extracts sarcasm representations from raw data. During the training phase, model optimization is guided by the loss between the labels provided by LLMs and those predicted by the multimodal detector. We are currently evaluating this model's effectiveness in generating sarcastic speech.

3. References

[1] S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, and S. Poria, "Towards multimodal sarcasm detection (an _obviously_ perfect paper)," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Zero-shot Audio Topic Reranking using Large Language Models

Mengjie Qian, Rao Ma, Adian Liusie, Erfan Loweimi, Kate M. Knill, Mark J.F. Gales

ALTA Institute, Machine Intelligence Lab, Department of Engineering, Cambridge University, UK

{mq227,rm2114,al826,el584,kmk1001,mjfg100}@cam.ac.uk

The Multimodal Video Search by Examples (MVSE) project investigates using video clips as the query term for information retrieval, rather than the more traditional text query. This enables far richer search modalities such as images, speaker, content, topic, and emotion. A key element for this process is highly rapid and flexible search to support large archives, which in MVSE is facilitated by representing video attributes by embeddings. This work aims to mitigate any performance loss from this rapid archive search by examining reranking approaches. In particular, zero-shot reranking methods using large language models (LLMs) are investigated as these are applicable to any video archive audio content. Performance is evaluated for topic-based retrieval on a publicly available video archive, the BBC Rewind corpus.

Specifically, the search process was split into two stages: in the first retrieval stage, each video is represented by vector embeddings to support very rapid retrieval for large archives. Subsequently, the top-N documents are reranked using a more powerful and computationally expensive model. The baseline approach for audio topic retrieval follows a multi-step process involving an initial foundation ASR system to transcribe the audio, an (optional) summarisation system to generate a more concise text for the ASR transcriptions, and a pre-trained topic classification model to extract topic embeddings. These embeddings are then used to rank the audio based on cosine similarities between the audio embeddings and the query embedding. The overall system architecture is illustrated in Figure 1. With this pipeline, we can develop a retrieval system in the wild without the need for domain-specific training data.

Reranking is employed in the second stage to boost retrieval performance and two zero-shot reranking methods with powerful LLMs are utilised: listwise reranking (LRL) [1] and pairwise reranking (PRL) [2]. LRL accepts a query along with a list of candidate documents as input and aims to generate a reordered list of document identifiers based on their relevance to the query. Conversely, PRL compares pairs of responses in each round to determine which one is considered better and the win ratio is used to rank the documents. Our results indicate that both GPT and Flan-T5 models improve retrieval performance as their size increases, with PRL showing efficiency on par with LRL while utilising fewer model parameters.

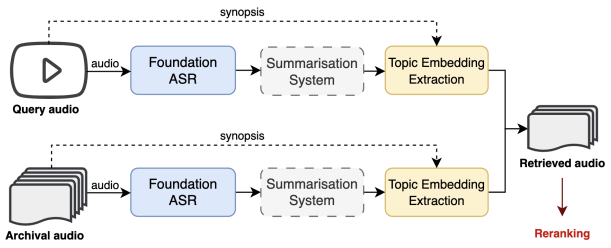


Figure 1: *Pipeline of the MVSE audio topic retrieval system.*

Additionally, we compared three types of text as input for reranking methods: ASR transcriptions (ASR), automatic summary generated in a zero-shot fashion using LLMs (AutoSum), and Synopses crafted by human transcribers that may contain external information not presented in audio content. Results highlight the performance degradation between ASR transcriptions and Synopses. One hypothesis for this degradation is that synopses contain external information not present in the audio, which are consequently missing from ASR transcriptions. To investigate this, we adopt a zero-shot fact-checking approach to assess information consistency. This approach first breaks text content into atomic facts with LLMs and then computes the percentage of atomic facts supported by a knowledge source. This method confirmed that synopses contain richer information than ASR transcriptions and highlighted a mismatch between AutoSum and Synopses, posing a challenge for AutoSum to match the performance of synopses despite its cleaner and more concise summaries.

In conclusion, this work demonstrates the effectiveness of zero-shot reranking methods in enhancing audio topic retrieval systems, with LLM-generated summarisation showing marked improvements over the baseline retrieval approach.

Table 1: *Retrieval and reranking using different text sources.*

Retrieval	Rerank	nDCG @3	nDCG @5	nDCG @10	Precision @1	Precision @3	Precision @5
ASR	-	0.47	0.46	0.45	0.48	0.47	0.45
ASR	ASR	0.52	0.50	0.47	0.59	0.50	0.49
ASR	AutoSum	0.54	0.51	0.47	0.58	0.52	0.49
AutoSum	ASR	0.54	0.52	0.49	0.59	0.52	0.51
Synopsis	Synopsis	0.73	0.68	0.59	0.76	0.71	0.65

This paper reports on research supported by EPSRC Project EP/V006223/1 (Multimodal Video Search by Examples) and Cambridge University Press & Assessment, a department of The Chancellor, Masters, and Scholars of the University of Cambridge.

1. References

[1] X. Ma, X. Zhang, R. Pradeep, and J. Lin, "Zero-shot listwise document reranking with a large language model," *arXiv preprint arXiv:2305.02156*, 2023.

[2] A. Liusie, P. Manakul, and M. J. Gales, "LLM Comparative Assessment: Zero-shot NLG Evaluation through Pairwise Comparisons using Large Language Models," *EACL*, 2024.

Exploring Category Specific Disorder in Chinese GAD, BD, and MDD with Verbal Fluency Task using Functional Near-Infrared Spectroscopy

Yufei Ren[1], *Gang Cui*[1], *Pozi Liu*[2], *Graham Murray*[3]

[1]Department of Foreign Languages and Literatures, Tsinghua University, Beijing, China
[2]School of Clinical Medicine, Tsinghua University, Beijing, China
[3]Department of Psychiatry, University of Cambridge, Cambridge, UK
ryf19@mails.tsinghua.edu.cn

Introduction

- The qualitative assessment of psychiatric disorders remains critically under-explored.
- The combination of the Verbal Fluency Task (VFT) and functional near-infrared spectroscopy (fNIRS) has emerged as a popular diagnostic tool in clinical settings across China[1].
- Differentiated neural activation patterns among Generalized Anxiety Disorder (GAD), Bipolar Disorder (BD), and Major Depressive Disorder (MDD) with fNIRS-VFT paradigm are found[2].
- However, analyses focusing on specific word categories may provide deeper insights into the neural processing differences among these disorders.

This study investigates real-time processing of different word categories in GAD, BD, and MDD using fNIRS. By examining the phenomenon of Category Specific Disorder (CSD)[3]—where only certain word categories are impaired while others remain unaffected.

Objectives:

1.To further examine the semantic network involved in the diagnosis test VFT from category specific perspective.

2.To enhance understanding of complex language processing in different psychiatric conditions for future language-specific assessments and treatments tailored to individual patient needs.

Key words:

Category Specific Disorder, Verbal Fluency Task, Language processing, Psychiatric Patients, fNIRS

Methodology

fNIRS:

- Functional near-infrared spectroscopy (fNIRS) is an emerging functional neuroimaging method, which non-invasively detects the brain activity near the brain surface by monitoring the relative concentration changes of oxy-hemoglobin ([oxy-Hb]) and deoxy-hemoglobin ([deoxy-Hb]) in the microvasculature of brain tissue during cognitive activity in real time[4].
- fNIRS is favored for its low operational cost, silent operation, and non-restrictive nature, making it accessible to a diverse range of patients.

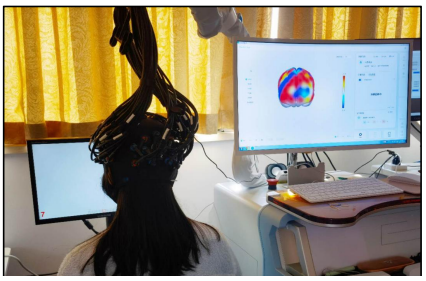


Figure 1: *fNIRS*

VFT:

- It is widely used to evaluate cognitive deficits in mental disorders by measuring word generation efficiency and the effectiveness of language use in information transfer.
- It assesses key executive functions such as working memory, response inhibition, and cognitive flexibility.

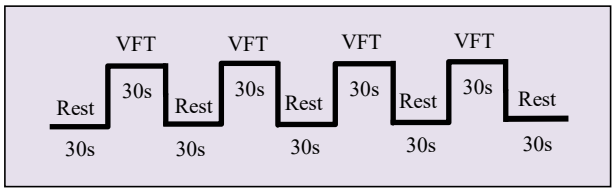


Figure 2: *fNIRS procedure of VFT*

Participants:

120 Chinese participants (30 MDD patients, 30 GAD patients, 30 BD patients and 30 HC) diagnosed using DSM-5 with a senior psychiatrist, from outpatients of the Department of Psychiatry, Yuquan Hospital, Beijing, China

References

[1] Ren et al. (2024). A scoping review of utilization of the Verbal Fluency Task in Chinese and Japanese clinical settings with near-infrared spectroscopy. *Frontiers in Psychiatry,* 15:1282546.

[2] Hu et al. (2021). Prefrontal cortex alterations in major depressive disorder, generalized anxiety disorder and their comorbidity during a Verbal Fluency Task assessed by multi-channel near-infrared spectroscopy. *Psychiatry Research,* 306:114229.

[3] Caramazza & Shelton (1998). Domain-specific knowledge systems in the brain: The animate-inanimate distinction. *Journal of Cognitive Neuroscience*, 10(1), 1-34.

[4] Obrig & Villringer (2003).Beyond the visible-imaging the human brain with light. *Journal of Cerebral Blood Flow & Metabolism*, 23:1-18.

Investigating Listener's Perception of Conversational Speaking Style using an Interview-Based Approach

Adaeze Adigwe, Sarenne Wallbridge, Simon King

The Centre for Speech Technology Research, University of Edinburgh, United Kingdom
{adaeze.adigwe, sarenne.wallbridge, Simon.King}@ed.ac.uk

Though the demand for TTS systems that can generate conversational-sounding speech is steadily increasing, we are far from producing consistent results. Previous research has focused on training TTS models on elicited or found conversational speech then measuring an improved listener preference. However, preference ratings cannot pinpoint *why* TTS voices fall short of conversational expectations. This underscores not only our limited understanding of conversational speaking styles, but also the limitations of typical speech synthesis evaluation methods. In this study, we pilot interview-based evaluation methods to understand the acoustic cues that listeners use to identify conversational speech.

Materials Inspired by [1], we propose two interview-based speech evaluations. We recruited 16 native English speakers; after hearing a speech stimulus, participants were asked to rate its conversationality and to provide explanations for their rating.

We utilise two types of speech stimuli. For Task 1, we select ten utterance pairs from a parallel corpus of read and spontaneous-conversational speech produced by the same speaker [2]. For Task 2, we use synthetic speech generated using a controllable FastSpeech2 model [3] which we trained on LJSpeech. This model allows us to synthesize prosodically varied realizations of the same text by controlling sentence-level prosodic features: $F0$ mean, $F0$ range, phone duration, energy, and spectral tilt. We apply eight predefined prosodic variation types (labelled Var A-H) to a set of test utterances. Our goal was to generate speech that varies along known axes rather than accurate simulations of conversational speech.

Task 1: Natural Speech Discrimination We began by asking whether listeners can identify conversational speaking style in the human speech. The interviewer instructed each participant: "Your task is to determine which sample in each pair is taken from an actual conversation and which one is the read-aloud speech, and comment on what in the speech led to your choice."

Task 2: Effect of Prosodic Manipulations We then asked how our prosodic manipulations of read speech affect conversationality ratings. Participants were asked to "Rate whether the given utterance sounds like it was said in a conversation and why.".

Results In Task 1, participants correctly identified the conversational sample 87% of the time. Given that the lexical content of the paired stimuli was identical, this demonstrates that listeners could differentiate between conversational and read speech based solely on differences in the acoustic realisation.

The results of Task 2 in Figure 1 show that even simple modifications by a TTS model can increase the perceived conversationality of utterances by 30% on average. Notably, samples generated with prosodic variation E—fast speaking rate, wide F0 range, and low mean F0—were judged as conversational 64% of the time. However, these results are incidental to our primary goal, which is to discover the reasons for listeners' judgements, beyond mere preferences, ratings, or accuracy. For this, we analysed participants' explanations to explore what factors contribute to their perception of conversationality. Common themes that emerged from semi-structured interviews included rhythm, pacing, degree of articulation and speaking rate as stylistic markers of conversational speech. Other dimensions included intonational variation, perceived spontaneity, prosodic boundaries, affect, propositional content and disfluency.

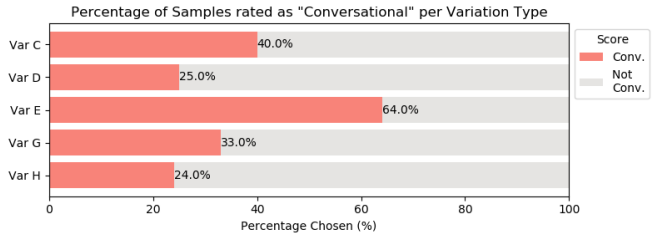


Figure 1: *Percentage of Stimuli rated as "Conversational" across Prosodic Variation Types*

The findings from Tasks 1 and 2 indicate that the acoustic-prosodic realisation of even *isolated* utterances contributes to perception of conversational speaking style. Qualitative interviews provide richer findings which can help us understand what constitutes speaking style: similar to previous research [1], we found that listeners rely heavily on imagining a conversational context to rate utterances, though its value somewhat reduced in contrastive perceptual tasks. Additionally, we observe that rating conversationality based solely on acoustics is complex, as listeners tend to rely on the conversationality of lexical content in their evaluations.

Acknowledgement Project is funded by European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 859588.

1. References

[1] Z. Hodari, C. Lai, and S. King, "Perception of prosodic variation for speech synthesis using an unsupervised discrete representation of F0," in *Proc. Speech Prosody 2020*, 2020, pp. 965–969.

[2] R. Dall, "Statistical parametric speech synthesis using conversational data and phenomena," Ph.D. dissertation, The University of Edinburgh, UK, 2017.

[3] T. Raitio, J. Li, and S. Seshadri, "Hierarchical prosody modeling and control in non-autoregressive parallel neural tts," in *ICASSP*. IEEE, 2022, pp. 7587–7591.

Towards End-to-End Spoken Grammatical Error Correction

Stefano Bannò, Rao Ma, Mengjie Qian, Kate M. Knill, Mark J.F. Gales

ALTA Institute, Department of Engineering, Cambridge University, UK
{sb2549, rm2114, mq227, kmk1001, mjfg100}@cam.ac.uk

Abstract

Grammatical feedback is crucial for L2 learners, teachers, and testers. Spoken grammatical error correction (GEC) aims to supply feedback to L2 learners on their use of grammar when speaking [1]. This process usually relies on a cascaded pipeline comprising an ASR system, disfluency detection (DD) and removal, and GEC [2, 3, 4], with the associated concern of propagating errors between these individual modules.

In this paper, we introduce an alternative "end-to-end" approach to spoken GEC, exploiting a speech recognition foundation model, Whisper [5]. This foundation model can be used to replace the whole framework or part of it, e.g., ASR and disfluency removal. These end-to-end approaches (see Figure 1) are compared to more standard cascaded approaches on the data obtained from a free-speaking spoken language assessment test, Linguaskill. Whisper$_{\text{dsf}}$, Whisper$_{\text{flt}}$, and Whisper$_{\text{gec}}$ are trained on disfluent, fluent, and grammatically correct manual transcriptions, respectively.

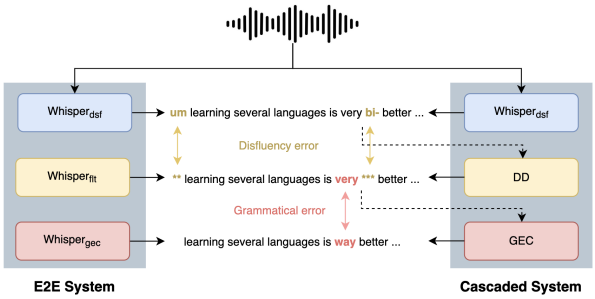


Figure 1: *Illustration of an end-to-end spoken GEC system and a cascaded system.*

As shown in Table 2, results measured in terms of word error rate (WER) and translation edit rate (TER) demonstrate that end-to-end spoken GEC (Whisper$_{\text{gec}}$) is possible within this architecture, but the lack of available data limits current performance compared to a system using large quantities of text-based GEC data. Conversely, end-to-end disfluency detection and removal (Whisper$_{\text{flt}}$), which is easier for the attention-based Whisper to learn, does outperform cascaded approaches (Whisper$_{\text{dsf}}$+DD), as shown in Table 1.

Additionally, the paper discusses the challenges of providing feedback to candidates when using end-to-end systems for spoken GEC.

This work reports on research supported by Cambridge University Press & Assessment, a department of The Chancellor, Masters, and Scholars of the University of Cambridge. Mengjie Qian is supported by EPSRC Project EP/V006223/1 (Multimodal Video Search by Examples).

	dsf	flt	gec
Whisper$_{\text{dsf}}$	**5.92**	9.97	19.17
Whisper$_{\text{dsf}}$+DD	—	6.31	—
Whisper$_{\text{flt}}$	9.22	**5.77**	14.89
Whisper$_{\text{gec}}$	13.73	10.37	**13.49**

Table 1: *Evaluation of Whisper performance against different manual references on Linguaskill in terms of %WER.*

System	ASR Model	gec WER	gec TER
Baseline	Whisper$_{\text{dsf}}$	19.17	18.74
	Whisper$_{\text{flt}}$	14.89	14.49
Cascaded	Whisper$_{\text{dsf}}$+DD+GEC	13.34	12.96
	Whisper$_{\text{flt}}$+GEC	**12.96**	**12.54**
E2E	Whisper$_{\text{gec}}$	13.49	13.08

Table 2: *GEC results with baseline, cascaded, and end-to-end systems.*

1. References

[1] K. M. Knill, D. Nicholls *et al.*, "Annotation of L2 English Speech for Developing and Evaluating End-to-End Spoken Grammatical Error Correction," in *Proc. 9th Workshop on Speech and Language Technology in Education (SLaTE)*, 2023, pp. 146–150.

[2] Y. Lu, M. J. F. Gales, and Y. Wang, "Spoken language 'grammatical error correction'," in *Proc. Interspeech 2020*, 2020, pp. 3840–3844.

[3] Y. Lu, S. Bannò, and M. J. F. Gales, "On assessing and developing spoken 'grammatical error correction' systems," in *Proc. 17th Workshop on Innovative Use of NLP for Building Educational Applications*, 2022, pp. 51–60.

[4] S. Bannò, M. Rais, and M. Matassoni, "Grammatical Error Correction for L2 Speech Using Publicly Available Data," in *Proc. 9th Workshop on Speech and Language Technology in Education (SLaTE)*, 2023, pp. 136–140.

[5] A. Radford *et al.*, "Robust speech recognition via large-scale weak supervision," in *Proc. 40th International Conference on Machine Learning*, 2023, pp. 28 492–28 518.

Hider-Finder-Combiner: Voice Conversion and Voice Privacy through Adversarial Information Hiding

Jacob J Webber[1] *Oliver Watts*[2] *Gustav Eje Henter*[3] *Jennifer Williams*[4] *Simon King*[1]

[1]The University of Edinburgh, UK; [2]SpeakUnique Ltd., UK;
[3]KTH Royal Institute of Technology, Sweden; [4]University of Southampton, UK

j.j.webber@ed.ac.uk

Privacy-preserving voice conversion aims to remove only the attributes of speech audio that convey identity information, keeping other speech characteristics intact. This paper presents a mechanism for privacy-preserving voice conversion that allows controlling the leakage of identity-bearing information using adversarial information hiding. This enables a deliberate trade-off between maintaining source-speech characteristics and modification of speaker identity. As such, the approach improves on voice-conversion techniques like CycleGAN and StarGAN, which were not designed for privacy, meaning that converted speech may leak personal information in unpredictable ways. Our approach is also more flexible than ASR-TTS voice conversion pipelines, which by design discard all prosodic information linked to textual content. Evaluations show that the proposed system successfully modifies perceived speaker identity whilst well maintaining source lexical content.

Our contributions in this work address the idea of controllable privacy in a new way. We build on work introduced with the Hider-Finder-Combiner (HFC-F_0) [1] model, which in turn has some similarities to Fader Networks [2]. The HFC-F_0 model was devised to modify F_0 contours of source speech, a control parameter which, like speaker identity, is deeply entangled with all other aspects of a speech signal. HFC-F_0 works by using an adversary (a *Finder*) to force an autoencoder to remove information about specific features of a speech signal. This disentangled representation can then be combined with new values for the control parameter, therefore achieving control through disentanglement. HFC-F_0 had component networks that were specific to the task of adding F_0 harmonics to a spectrogram, specifically using a Parallel Bank of Transposed Convolutions, which has since been shown to be very well suited for this in the domain of singing synthesis [3]. The use of models that are specifically equipped to do tasks like this invokes *inductive bias* and similar approaches have been used for voice privacy. That is, machine learning modules are constrained by design in a way that is useful for privacy.

Drawing this analogy between speech signal disentanglement and privacy control has been considered before in speaker anonymisation[4]. Systems have applied inductive biasing, using models that only have the power to learn specific attributes. For example, bottlenecks have been used to guide synthesis by accounting for features that are desirable to preserve [5, 6, 7, 8]. An example is preserving the lexical content of utterances alongside such bottleneck features. Other privacy-preserving approaches [7] have applied Fader Networks [2] to modify specific scalar attributes of the speech signal. Our system, HFC-VP, works by creating a disentangled hidden representation which can then be recombined with an arbitrary speaker embedding. To the best of our knowledge, the techniques that we introduce represent the first effort to achieve anonymization solely through minimising the mutual information between a predicted probability distribution of speaker identity and a prior distribution of speaker identity, from a large dataset of 904 speakers. Our principled and "unguided" approach means that our single system is capable of modifying different aspects of speech, such as prosody or lexical content, exactly where it is necessary to preserve privacy.

Our work differs substantially from the original HFC-F_0 system[1]. Changes include an improved treatment of the HFC-F_0 training losses. Our proposal also replaces the component networks, now using powerful transformer encoders and residual convolutional networks in place of RNNs. The resulting system outperforms a strong neural source-filter baseline in our evaluations. It also provides the possibility of sharing an anonymised speech representation that can be resynthesised at a later date. Audio examples are available `https://hfcvp.github.io/`.

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

1. References

[1] J. J. Webber, O. Perrotin, and S. King, "Hider-finder-combiner: An adversarial architecture for general speech signal modification," in *Proc. of Interspeech 2020*, 2020, pp. 3206–3210.

[2] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer *et al.*, "Fader networks: Manipulating images by sliding attributes," in *Advances in Neural Information Processing Systems*, 2017, pp. 5963–5972.

[3] T. Jayashankar, J. Wu, L. Sari, D. Kant, V. Manohar, and Q. He, "Self-supervised representations for singing voice conversion," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[4] P. Champion, D. Jouvet, and A. Larcher, "Are disentangled representations all you need to build speaker anonymization systems?" in *Proc. of Interspeech 2022*, 2022.

[5] K. Qian, Y. Zhang, S. Chang, M. Hasegawa-Johnson, and D. Cox, "Unsupervised speech decomposition via triple information bottleneck," in *International Conference on Machine Learning*. PMLR, 2020, pp. 7836–7846.

[6] J. Mosiński, P. Biliński, T. Merritt, A. Ezzerg, and D. Korzekwa, "AE-Flow: Autoencoder Normalizing Flow," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[7] P.-G. Noé, A. Emînî, M. Driss, T. Parcollet, A. Nautsch, and J.-F. Bonastre, "Adversarial Disentanglement of Speaker Representation for Attribute-Driven Privacy Preservation," in *Proc. Interspeech 2021*, 08 2021.

[8] S. Kovela, R. Valle, A. Dantrey, and B. Catanzaro, "Any-to-Any Voice Conversion with F0 and Timbre Disentanglement and Novel Timbre Conditioning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

A Comparison of Synthesis Method Impact on Listener Perception of Play-Acted Speech

Emily Lau[1], *Brechtje Post*[1], *Kate Knill*[2]

University of Cambridge, United Kingdom,
[1]Phonetics Laboratory, [2]Department of Engineering
{ehynl2, bmbp2, kmk1001}@cam.ac.uk

Performative speech is a nuanced and complex phenomenon. Nonetheless, this speech domain has been the subject of much interest to both Linguistics and Artificial Intelligence researchers, since it has been shown that performative speech has characteristics distinct from natural speech [1]. While much of the research modelling expressive prosody directly maps acoustic parameters to different emotions [2], other work has examined emotional expression in context of the evolution of human communicative functions [3]. Furthermore, [4] has proposed that affective speech is controlled by so-called Bio-informational Dimensions (BIDs) which control the vocal signal of the speaker in order to influence the behaviour of the receiver. The two BIDs that are the most explored in literature [5] and hence investigated in this work were those of 1) size projection, the body size projected by the speaker to the listener, and 2) dynamicity, the vigorousness of the speaker's voice.

A first listening test targeted the emotion of "anger" [6]. The stimuli were utterances spoken by an SSBE-speaking male, resynthesised along the BIDs of size projection and dynamicity to varying degrees, in order to simulate performative expressions of anger. Pitch range and median were manipulated simultaneously on each individual pitch point in the utterance. 30 SSBE-speaking participants listened to pairs of these resynthesised utterances, and were asked to rate their differences in performative expression. It was predicted based on [1] that listeners would find the stimuli with higher size projection and dynamicity more performatively angry. Combined ANOVAs indicated that size projection has a significant and positive effect on listener perceptions of performative anger. While dynamicity also had a significant effect, contrary to predictions, ratings consistently decreased when dynamicity increased.

In this work, a second listening test identical to the first one is currently being conducted, except with stimuli created with an alternate resynthesis method. In this method, the pitch range and median manipulations are performed in two separate steps, as opposed to one simultaneous process, with some additional scaling. The results confirmed the previous finding that size projection has a significant positive effect, while dynamicity had no significant effect. This also implied that pitch range is not a particularly impactful parameter on listener perception of vocal expression on its own. This calls into question listener expectations of dramatic anger and whether this was reflected in the stimuli.

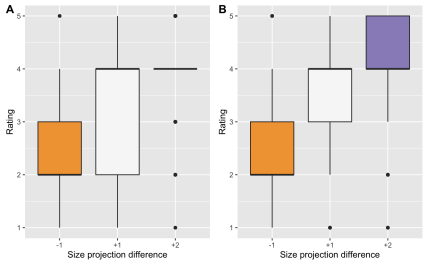


Figure 1: *Ratings of dramatic anger across all trials for Size Projection difference for the neutral synthesis base, where A indicates the ratings from [6], while B indicates the ratings for the current experiment.*

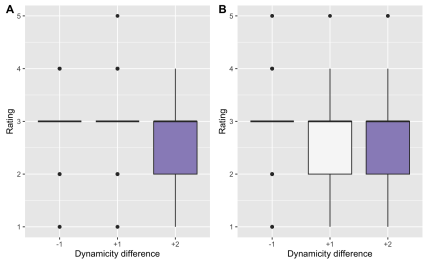


Figure 2: *Ratings of dramatic anger across all trials for Dynamicity difference for the neutral synthesis base, where A indicates the ratings from [6], while B indicates the ratings for the current experiment.*

1. References

[1] R. Jurgens, A. Grass, M. Drolet, and J. Fischer, "Effect of acting experience on emotion expression and recognition in voice: Non-actors provide better stimuli than expected," *Journal of Nonverbal Behavior*, vol. 39, pp. 195–214, 2015.

[2] K. Scherer, "Vocal correlates of emotional arousal and affective disturbance." pp. 165–197, 1989. [Online]. Available: https://psycnet.apa.org/record/1989-97735-007

[3] J. J. Ohala, "An ethological perspective on common cross-language utilization of f0 of voice," *Phonetica*, vol. 41, pp. 1–16, 1 1984. [Online]. Available: https://www.degruyter.com/document/doi/10.1159/000261706/html

[4] S. Chuenwattanapranithi, Y. Xu, B. Thipakorn, and S. Maneewongvatana, "Encoding emotions in speech with the size code," *Phonetica*, vol. 65, no. 4, pp. 210–230, 2008.

[5] Y. Xu, A. Lee, W.-L. Wu, X. Liu, and P. Birkholz, "Human vocal attractiveness as signaled by body size projection," *PLoS ONE*, vol. 8, p. e62397, 4 2013. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0062397

[6] E. Lau, B. Post, and K. Knill, "A biological investigation of performative speech through synthesis," in *Proceedings of the 20th International Congress of Phonetic Sciences*, R. Skarnitzl and J. Volín, Eds. Gaurant International, 2023, pp. 3111–3115.

The UKIS 2024 Committee would like to thank ISCA and Interspeech 2023 organisers for kindly providing this template.

A framework for flexible model combination of speech tasks

Shreyas Ramoji, Thomas Hain

LivePerson Center, Speech and Hearing Research Group,
Department of Computer Science, University of Sheffield, United Kingdom
{s.ramoji,t.hain}@sheffield.ac.uk

Abstract

Speech contains semantic and non-semantic information embedded in a one-dimensional physical signal, such as phonetic units, words, languages, speakers, gender, emotion, meaning, etc. Neural networks allow us to train models to sol. In this presentation, I intend to discuss my ideas and initial efforts towards building a framework that allows flexible model combination of speech tasks. By drawing attention to some limitations of currently popular approaches, we propose an approach that can potentially overcome these limitations and improve performance by enabling better information sharing and knowledge transfer across tasks.

In recent years, there have been two popular research directions in speech technologies. The first has focused on foundation models - deriving universal speech representations that aim to be flexible in adapting to various downstream tasks with minimal efforts. Several benchmarks such as SUPerB [1] aim to constrain the downstream models and limit their parameter sizes, allowing focused research aimed at improving the representation quality, using a variety of self-supervised, supervised and multitask objectives for this purpose. The second direction utilizes foundation models such as the Wav2Vec, HuBERT, and Whisper as means of extracting embeddings at regular time intervals, and focus on downstream training tailored to specific tasks. This approach relies on the encoded representations to capture the relevant information for every downstream task in a desired way. Both these research directions complement each other, and work together to progressively improve the state-of-art for each task. However, there are two potential drawbacks. Firstly, the information relevant to one task could be considered as noise to another, which only means this approach may impose overly restrictive structural constraints that may limit the performance. Secondly, the downstream models are trained independently, and the foundation models do not really capture the relationship between tasks.

A long-term goal of addressing these limitations points to discovering structures that capture information related to several tasks more efficiently, while also learning the relationship between them. Ideally, a neural network structure that tries to solve multiple tasks must have multiple entry and exit points for various tasks. A relevant research direction of interest is meta-learning [2], which aims to combine multiple model inferences in various different ways. As a first step, combining two models already trained on different tasks into a single structure, and jointly training this combined model with multitask and meta-learning objectives is a reasonable idea. In this case, the aim is to improve over the baseline performance achieved by individual models. This initial approach allows us to investigate the potential benefits of combining models and explore its connection to the long-term goal of a network with automatic exit point discovery.

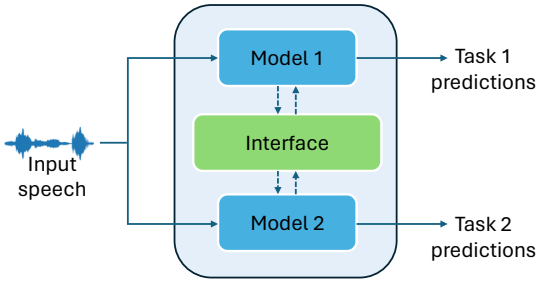


Figure 1: *An outline of the multitask framework*

While toolkits like SpeechBrain [3] offer convenient functionalities for training a variety of speech tasks leveraging pre-trained models and loss functions, implemented in a streamlined manner, they rely on pre-defined network structures built with multiple hierarchical levels of abstractions. This prevents the various nodes of the model's computation graph from being accessible directly. In our case, this makes it inconvenient to implement structural changes while combining two models in a flexible manner. To overcome this limitations, dedicated efforts in building this framework can allow us to combine models in a flexible manner.

Figure 1 illustrates an outline of our framework. The two models are connected by a generic structure called the interface. The framework is made configurable by enumerating the entry and exit points of the two models in a `YAML` file. These entry and exit points are connected by defining neural network modules, thereby allowing information transfer. This framework provides flexibility, convenience and also a means of controlling the gradient updates, which are crucial to explore multi-task and meta learning solutions.

1. References

[1] S. wen Yang et.al., "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech 2021*, 2021, pp. 1194–1198.

[2] S. Liu, A. Davison, and E. Johns, "Self-supervised generalisation with meta auxiliary learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[3] M. R. et.al., "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

Putting Expression in the Irish Synthetic Voice

Anna Maria Giovannini, Zihan Wang, Andrew Murphy, Maria O'Reilly, Ailbhe Ní Chasaide, Christer Gobl

Phonetics and Speech Laboratory, CLCS,
School of Linguistic, Speech and Communication Sciences, Trinity College Dublin, Ireland
giovanna@tcd.ie, zwang7@tcd.ie, murpha61@tcd,ie, moreil12@tcd.ie,
anichsid@tcd.ie, cegobl@tcd.ie

1. Introduction

Voice transforms involving primarily shifts in voice source and f_0 parameters were carried out on synthesised utterances of Irish to alter their perceived affective colouring. These transforms are intended in the first instance to provide for a degree of affect control in an Irish language AAC system, currently being used by non-speaking autistic children. The transforms were guided by voice source analysis of utterances produced by a male actor (a speaker of the Kerry dialect of Irish, featured as one of the voices in the ABAIR synthesiser at https://abair.ie) to portray the affects *angry, happy, sad, bored, relaxed* as well as a neutral rendition. The resynthesised neutral utterance was used as a baseline for three types of manipulations: i) global shifts applied to the entire utterance; ii) local shifts applied only to the accented syllables; iii) a combination of global and local changes. Stimuli targeting *sad* and *happy* included tempo changes and formant shifts were included for *happy.*

2. Methodology

The analysis and synthesis were carried out using the Voice Source Generator and Fant's transformed LF model (these are described in [1]). Parameter shifts from neutral for the three series of stimuli are summarised in Table 1 and are described in [2]. All 16 stimuli (3 manipulations × 5 affects + neutral) were presented to 23 listeners, all Irish speakers, in the form of a multiple-choice emotion recognition test.

Table 1: *Mean parameter shifts from neutral.*

Affect	f_0 diff (semitones)			E_e diff (dB)			R_d diff (%)		
	global	local	combined	global	local	combined	global	local	combined
angry	2.4	1.4	3.8	6.0	1.5	7.4	-14%	-5%	-18%
happy	4.1	4.6	8.7	4.1	2.6	6.7	-1%	4%	3%
bored	-2.0	0.0	-1.9	-3.4	-2.7	-6.1	0%	19%	20%
sad	-1.0	0.2	-0.8	-1.3	-1.3	-2.5	8%	13%	21%
relaxed	1.0	0.1	1.1	0.4	-0.3	0.1	16%	13%	30%

3. Results

Listeners' evaluations, presented in Figure 1, most positively identified the high activation affects *happy* and *angry*. Stimuli targeting *sad* were also effective, while those targeting *bored* and *relaxed* were not, although *bored* was positively associated with some of the sad-targeting stimuli. Results for low activations states are confounded by the fact that the *neutral* stimulus was to some degree biased towards *bored, sad* and *relaxed* affects. Of the three types of transforms, global, local and combined, the most effective appeared to vary with the targeted affect.

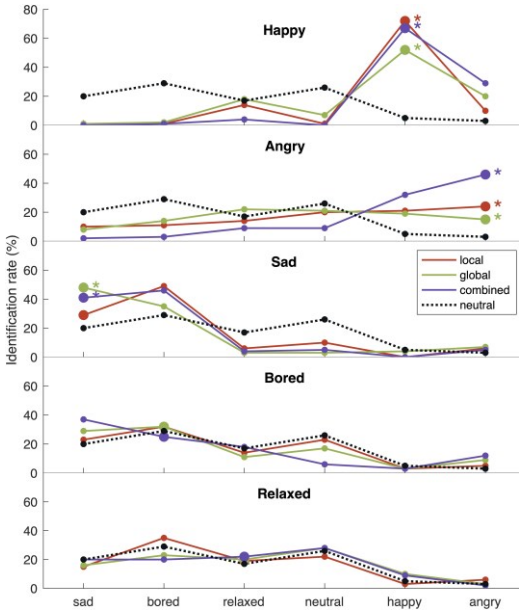


Figure 1: *Identification rates for each triad of affect-targeting stimuli: global (green), local (orange) & combined (purple) for both targeted and other affects. Asterisks mark values significantly higher than neutral (black dotted lines).*

4. Acknowledgements

The research conducted in this publication was funded by the Irish Research Council under award number GOIPG/2021/561 to the first author, and by the RóboGlór project, awarded by the Dept. of Tourism, Culture, Arts, Gaeltacht, Sports and Media, with funding from the National Lottery, as part of the *Straitéis 20 Bliain don Ghaeilge*.

5. References

[1] Z. Wang and C. Gobl, "A System for Generating Voice Source Signals that Implements the Transformed LF-model Parameter Control," in *Proceedings of INTERSPEECH* 2023, Dublin, Ireland, 2023, pp. 4738-4742.

[2] A. M. Giovannini, Z. Wang, M. O'Reilly, A. Ní Chasaide and C. Gobl, "Voice Transforms For Affect Control in Irish Speech Synthesis," in *Proceedings of Speech Prosody 2024, forthcoming.* Leiden, the Netherlands, 2024.

A New Standardized and Reproducible Benchmarking framework for Automatic Personality Perception Experiments over the SSPNet Speaker Personality Corpus

Nesreen Alareef[1], *Evangelia Fringi*[1], *Tanaya Guha*[1] *and Alessandro Vinciarelli*[1]

[1]University of Glasgow, UK

2699698A@student.gla.ac.uk, evangelia.fringi@glasgow.ac.uk, Tanaya.guha@glasgow.ac.uk, Alessandro.vinciarelli@glasgow.ac.uk

Abstract

This study aims at improving the reproducibility of the results obtained over the SSPNet Speaker Personality Corpus. This corpus is a dataset that, while having been repeatedly used in the literature (see [1] for a list of works using the data) and having served as a benchmark in an international evaluation campaign [2], still does not allow a fully rigorous comparison across multiple works. For this reason, this article proposes a benchmarking framework that makes the results obtained over the data fully reproducible (see Figuer 1). To such a purpose, the distribution includes not only data and annotations, but also features extracted with different methods (including foundation models that were not available at the time the Corpus was created), experimental protocols and baseline results to be used as a reference for comparison.

One of the most specific aspects of the Corpus is that it is designed for paralanguage-based APP. By design, the personality raters did not understand the language of the data and, therefore, they were mostly influenced by nonverbal cues. This made it possible to explore the relationship between paralanguage and personality perception, while confirming that the data is compatible with the findings of Psychology about the way listeners attribute traits to speakers [3]. However, such a specificity is a possible limitation too because it does not allow one to apply multimodal approaches that take into account both language and paralanguage. One way to address such a problem is to ask new raters, capable to understand the language of the data, to assess the speakers again and test whether there is any difference. Effective speech transcribers such as Whisper, not available when the data was originally collected, can then be used to transcribe the data and enrich the work on the Corpus with both language-based and multimodal approaches.

Besides the above, the Corpus provides the opportunity to investigate other problems that, to the best of our knowledge, have not been addressed extensively (the list is not exhaustive):

- The metadata about gender allow one to explore possible biases in personality perception and the development of gender specific models.
- Possible correlations among different traits can be exploited to improve the APP performance (e.g., through multi-task learning).
- The availability of all scores provided by individual raters allows one to model the variance in the personality assessments or to develop models specific to every rater.
- The state-of-the-art focused on the use of binarized scores, but it is possible to target the inference of the underlying continuous personality ratings.

In this respect, the combination of the Corpus and the proposed benchmarking framework promises to be a valuable resource for the speech community.

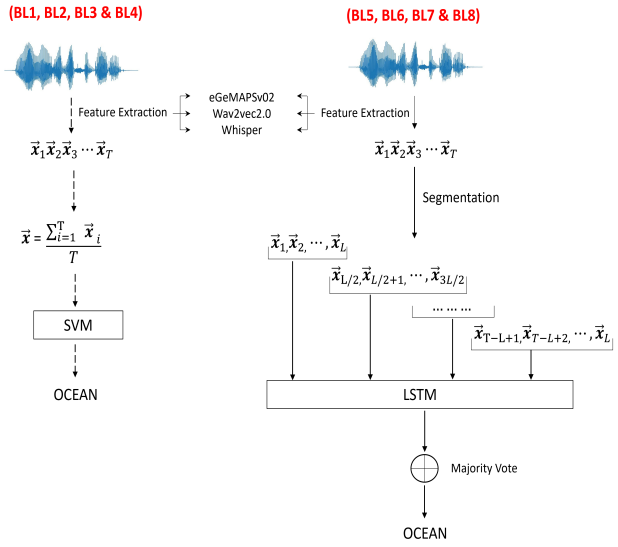


Figure 1: *Baseline approaches. The diagram illustrates the eight baseline approaches applied in the experiments, where 'OCEAN' represents the outcome of classifying the Big Five personality traits. The features utilized include eGeMAPSv02, wav2vec2.0, and Whisper, with feature extraction performed via wav2vec2.0 in two modes: output from the last convolutional layer and the last hidden layer. These features were employed alongside SVM and LSTM models to generate these approaches.*

Index Terms: Automatic Personality Perception, Big Five, Computational Paralinguistics, Benchmark data.

1. References

[1] Y. Li, P. Bell, and C. Lai, "Transfer learning for personality perception via speech emotion recognition," *arXiv preprint arXiv:2305.16076*, 2023.

[2] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. Van Son, F. Weninger, F. Eyben, T. Bocklet *et al.*, "The interspeech 2012 speaker trait challenge," in *INTERSPEECH 2012, Portland, OR, USA*, 2012.

[3] P. Ekman, W. Friesen, M. O'Sullivan, and K. Scherer, "Relative importance of face, body, and speech in judgments of personality and affect," *Journal of Personality and Social Psychology*, vol. 38, no. 2, pp. 270–277, 1980.

Oral Session A

Speech Modifications for Improved Listening Experience in ADHD Adults

Lucy Valls-Reed and Jennifer Williams

School of Electronics and Computer Science, University of Southampton, UK
lvr1g23@soton.ac.uk, j.williams@soton.ac.uk

1. Introduction

It is estimated that 2.5-4% of adults in the UK have been diagnosed Attention-Deficit/Hyperactivity Disorder (ADHD). ADHD presents many challenges for individuals in numerous aspects of their lives, especially communication. This makes it especially difficult for those who have ADHD to engage with online media content designed for neurotypical listeners such as lectures, TED Talks, tutorial videos, and news broadcast (McCarthy et al., 2012). In this work we propose to explore how ADHD-friendly speech can be developed, and eventually realised as an online tool plugin or speech modification 'filter' tailored to individual listener needs. This work answers the recent call to understand how different types of speech modifications, including those from speech synthesis, can affect listening effort (Simantiraki, Cooke, and King, 2018).

As a first step toward developing bespoke speech modifications for ADHD listeners, we propose to explore how *clear speech* is perceived for the ADHD community. By examining the features of clear speech that could improve intelligibility and listening comprehension for ADHD listeners, we can begin to assess the interplay between speech synthesis, voice modification, and ADHD cognitive processing mechanisms. In recent years, there has been growing research into clear speech and audio in general to diagnose ADHD (Bloomberg et al., 2019), alongside known benefits for individuals with hearing impairments and dyslexia (Smiljanic 2021; Breier et al., 2002). However, it remains unknown which features of speech contribute to enhancing intelligibility and reduced cognitive loading for individuals with ADHD. If such features can be identified, there is massive potential to use these features to enhance, modify, and synthesise speech in ways that benefit the ADHD community.

2. Speech Features

In neurotypical listeners, increased intelligibility from clear speech is due to a variety of different features (Uchanski, 2005) on the production side from the speaker, as well the interplay of intelligibility and cognitive effort on the part of the listener (Winn and Teece, 2021). As an initial first step to explore these issues for ADHD listeners, we propose to look into the following acoustic features, and may add additional phonetic features at a later stage:

1. **Speaking Rate**: Previous work has shown mixed results as to how speaking rate correlates with increased intelligibility and decreased listening effort, and do not specifically account specifically for ADHD listeners (Bradlow et al., 1996; Hazan and Markham, 2004).
2. **Raised F0**: From the music domain, showed that people with ADHD have better concentration when listening to sound with certain tones and higher pitch (Juiter and Wu, 2021), especially with increased energy in the 1000-3000 Hz range (Krause and Braida, 2004).
3. **Second and Third Formants**: higher spectral prominence in second and third vowel formants correlates with increased intelligibility in clear speech (Krause, 2001).

3. Methodology

The speech materials we plan to use are based on the nonsense sentences described in Picheny et al (1985). The purpose of these sentences is to provide no context to force listeners to focus on word identification. All of the sentences will follow the format of: article (adjective) noun (auxiliary verb) verb (preposition) article noun. Parentheses denote parts of the sentence which are optional, for example "A light oven should pivot the beam". We will record 300 nonsense sentences, spoken by one or more talkers, and split them into 20 lists of 15 utterances.

We propose to assess 20 ADHD listener participants and record the percentage of key words (adjective, noun and verb) correct to determine intelligibility scores. The speech will be presented via headphones in an isolated room at a fixed volume. Each participant listens to one of the control lists (15 clear sentences spoken in conversational speech) and one modified list for each feature, ensuring all lists are different so no participant listens to the same sentence twice. After each sentence, listeners would have to repeat the sentence they heard back into a microphone to be recorded. We would then analyse this audio to determine the percentage of key words correct intelligibility scores. We will assess the listening effort with pupilometry (Govender, Wagner, and King, 2019).

4. Discussion

The significance of tailoring speech for individuals with ADHD cannot be emphasised enough. In a growing digital age, the challenges individuals with ADHD face when they engage with online media is becoming increasingly common, despite ADHD listeners being under-represented in research. Both intelligibility and listening effort are not well understood for ADHD listeners. By developing ADHD-friendly speech synthesis, we have the potential to transform the way these individuals interact with digital media, providing them with greater accessibility in these experiences. Our work in-progress also highlights the broader importance of considering diverse neurocognitive profiles in the design of future assistive technologies

This work was supported by the UK EPSRC through the Trustworthy Autonomous Systems (TAS) Hub (EP/V00784X/1)

Variability of speech timing features across repeated recordings of non-pathological speech samples

Judith Dineley[1], Ewan Carr[1], Lauren L. White[1], Catriona Lucas[1], Zahia Rahman[1], Tian Pan[1], Faith Matcham[2], Johnny Downs[1], Richard Dobson[1,3], Thomas F. Quatieri[4], Nicholas Cummins[1]

[1]Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK
[2]School of Psychology, University of Sussex, Falmer, UK
[3]Institute of Health Informatics, University College London, London, UK
[4]MIT Lincoln Laboratory, Lexington, MA, USA

judith.dineley@kcl.ac.uk, nick.cummins@kcl.ac.uk

Introduction

Speech offers an accessible, objective way to monitor an individual's health. It has the potential to measure outcomes in clinical trials and, ultimately, be applied in clinical practice. Variations in speech timing features have been reliably linked to symptoms of various health conditions, promising clinical potential. However, replication challenges hinder their translation; extracted speech features are susceptible to methodological variations between recording and processing pipelines. This variability is problematic when assessing health state: it can mask or even mimic pathological changes.

Understanding the differences in extracted speech features that arise through differences in extraction pipeline design is key to developing reliable pipelines for clinical research and practice. Towards this aim, this study compares three established open-source timing feature extraction techniques to understand the strengths and limitations of different extraction approaches and inform future studies of pathological speech.

Methodology

We recorded speech samples from 28 healthy participants speaking in the morning, afternoon and early evening (*Day*), and 26 participants speaking on a Monday, Wednesday, and Friday at the same time (*Week*). Participants were pre-screened to exclude any hearing, speaking, neurological or mental health disorders that might affect their speech. Recordings were conducted in a room insulated for reverberation and with quiet surroundings. Recordings were made with an Audio Technica AT2020USB+ condenser microphone, with no built-in pre-processing, operated using Audacity. Participants recited one of three texts of similar length and complexity per session to avoid practice effects: *The Rainbow Passage* [long], *Your Rate of Oral Reading* and *Comma gets a Cure*.

From these recordings, we extracted five exemplar timing and pausing features using three methods. The first was a Praat script that uses intensity thresholding. The other two methods utilize word boundaries estimated from forced alignment, with transcripts generated using two different automatic speech recognition (ASR) engines: (i) Whisper and (ii) wav2vec 2.0. We compared how the features *Phonation Time Ratio*, *Pause Time Ratio*, *Pause Rate*, *Mean Pause Duration* and the *Coefficient of Variation (CoV) of Pause Duration* generated by the three methods differed, including how they are affected by within-speaker variation (over time). These features were selected to be robust to differences in reading content.

We use intraclass correlation coefficients (ICC) to explore the effect of non-pathological within-speaker variability on timing features derived from the three methods. ICCs were calculated using linear mixed effect (LME) models, with a separate model for each feature-study combination. The LME models were unconditional and included the speech feature as the outcome and a participant-level random intercept only.

Results and Discussion

Key results showed that different extraction methods resulted in different ICC values. This indicates there are differences between extraction methods in the quantification of within-person feature variability over repeated feature measures. Feature extraction methods that result in similar values would show similar feature variabilities over repeated recordings with similar ICC. Further, even if pipelines extracted different feature values from one another, if they did so systematically (e.g., mean pause duration is always 20ms longer using one method), we would also expect these to show similar variability across repeated recordings. However, we do not observe this.

We observed features extracted using Praat had markedly lower ICCs for *Day* compared to *Week*, indicating greater within-person variability. We speculate that as an intensity-based extraction approach, this method is more sensitive to non-pathological variation in participants' voice intensity. We also observed that W2F2-FA tended to exhibit lower within-person variability than Whisper-FA. As the only difference between the two is the transcription methods, this must be attributable to differences in transcription methodology.

Conclusion

Our results demonstrate that different extraction methods led to differing estimates of within-speaker feature variability over time. These findings highlight the importance of feature extraction in study design and interpretation, and the need for consistent, accurate extraction techniques for clinical research. When predicting health state using speech, differences in extraction pipelines must be understood and accounted for, particularly in clinical cohorts where differences attributable to extraction methods may be larger. For example, transcription tools that deliberately remove dysfluencies should be used with caution, as dysfluencies are more common in pathological speech and may also contain valuable health information. Differences in forced alignment performance additionally have implications for acoustic feature extraction, which we will investigate in future work.

Poster Session B

Automatic Speech Recognition System-Independent Word Error Rate Estimation

Chanho Park, Mingjie Chen, Thomas Hain

Speech and Hearing Research Group, University of Sheffield, UK
{cpark12, mingjie.chen, t.hain}@sheffield.ac.uk

Abstract

Word error rate (WER) is a metric used to evaluate the quality of transcriptions produced by Automatic Speech Recognition (ASR) systems. In many applications, it is of interest to estimate WER given a pair of a speech utterance and a transcript. Previous work on WER estimation focused on building models that are trained with a specific ASR system in mind (referred to as ASR system-dependent). These are also domain-dependent and inflexible in real-world applications. In this paper, a hypothesis generation method for ASR System-Independent WER estimation (SIWE) is proposed. In contrast to prior work, the WER estimators are trained using data that simulates ASR system output. Hypotheses are generated using phonetically similar or linguistically more likely alternative words. In WER estimation experiments, the proposed method reaches a similar performance to ASR system-dependent WER estimators on in-domain data and achieves state-of-the-art performance on out-of-domain data. On the out-of-domain data, the SIWE model outperformed the baseline estimators in root mean square error and Pearson correlation coefficient by relative 17.58% and 18.21%, respectively, on Switchboard and CALLHOME. The performance was further improved when the WER of the training set was close to the WER of the evaluation dataset.

Index Terms: Word error rate, WER estimation, hypothesis generation

1. Proposed methods

1.1. Word error rate estimation

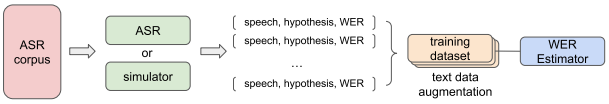
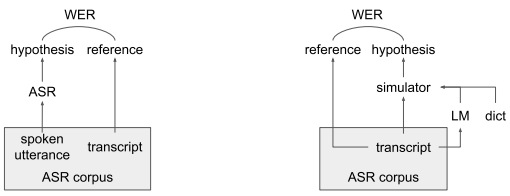


Figure 1: *Illustration of WER estimation*



(a) ASR hypothesis (b) Hypothesis generation

Figure 2: *Illustration of ASR hypothesis and Hypothesis generation*

1.2. Hypothesis generation

Table 1: *Hypothesis generation strategies and edit types. del.: deletion, sub.: substitution, ins.: insertion.*

strategy	del.	sub.	ins.
Random selection	✓	✓	✓
Phonetic similarity		✓	
Linguistic probability			✓

2. Results

Table 2: *Fe-WERn [1] is an estimator trained on HYPn train transcribed using ASRn. Evaluation metrics are Root Mean Square Error (RMSE) and Pearson Correlation Coefficients (PCC).*

model	**evaluated on**			
	HYP1 eval		**HYP2 eval**	
	RMSE↓	**PCC↑**	**RMSE↓**	**PCC↑**
Fe-WER1	**0.0926**	**0.8806**	0.1234	0.6524
Fe-WER2	0.1350	0.7210	**0.0928**	**0.7962**
SIWE7	0.1443	0.7282	0.1195	0.6739
SIWE8	**0.1268**	**0.7914**	**0.1133**	**0.7083**

3. Conclusion

Hypothesis generation and data augmentation for SIWE are proposed in this paper. The estimator trained on the hypotheses generated by this proposed method outperforms the baselines on out-of-domain (OOD) datasets. As a hypothesis generation strategy, linguistic information helps improve the performance of WER estimators when they are evaluated on in-domain datasets. The phonetic similarity for substitution improved the performance of the estimators on both in-domain and OOD datasets. On OOD datasets, the performance of estimators has further improved when the WER was targeted.

4. Acknowledgements

This work was conducted at the Voicebase/Liveperson Centre of Speech and Language Technology at the University of Sheffield which is supported by Liveperson, Inc..

5. References

[1] C. Park, C. Lu, M. Chen, and T. Hain, "Fast word error rate estimation using self-supervised representations for speech and text," *arXiv preprint arXiv:2310.08225*, 2023.

Frameworks for Assessing Privacy Risks in Affective Speech-Based Systems

Basmah Alsenani[1], *Tanaya Guha*[1] , *Alessandro Vinciarelli* [1]

[1]University of Glasgow, United Kingdom

b.alsenani.1@research.gla.ac.uk, tanaya.guha@glasgow.ac.uk,
alessandro.vinciarelli@glasgow.ac.uk

Lately, more AI applications have been analyzing speech data to understand users' emotions and mental health conditions, particularly focusing on depression detection. Automated depression detection includes identifying signs of depression in individuals by analyzing their tone, word choice, and speech patterns to provide insights into their mental well-being. This technology has the potential to revolutionise mental health care by providing early detection and intervention for individuals at risk for depression. Apart from the spoken content, speech audio captures details about the speaker's identity and demographic traits, posing significant privacy risks to users, including potential inference attacks as the most AI applications nowadays operate partially on the cloud and send the interemdiate representations to the cloud for further processing. The transmitted representations are at risk of attacks that may expose users' sensitive demographic details [1, 2, 3], especially in applications like depression detection, posing privacy risks. Previous efforts have primarily focused on developing deep models that preserve privacy such as differential privacy [4] and cryptographic technologies [5]. However, there is a critical need to systematically assess and quantify privacy risks in such systems to ensure the responsible development of AI technologies for mental health applications, as neglecting these risks can have serious implications for user privacy and data security.

Our work, which is submitted to an international conference, presents the first framework for systematically assessing privacy risks in a multimodal (audio-lexical) depression detection system, particularly looking at attribute inference attacks. Unlike past works that considered only white-box gender inference attacks against unimodal systems, our framework designs novel white-box and black-box attacks across multiple modalities against three protected speaker attributes: gender, age and education level. We present extensive results on a large, clinically validated dataset, demonstrating critical vulnerability of depression detection systems, where an adversary can infer speaker attributes with 59% - 68% accuracy even for inputs as short as 10 seconds of speech. Our findings provide valuable insights and guidelines to guide the development and evaluation of privacy-preserving models for detecting depression through speech analysis.

The proposed framework not only explores attacks on protected speaker attributes such as gender, age, and education level in both white-box and black-box scenarios across multiple modalities, but also sheds light on the vulnerability of depression detection systems. With our extensive results showing that speaker attributes can be inferred with high accuracy even from short 10-second speech inputs, it is evident that privacy-preserving models for speech-based depression detection need to be developed and benchmarked carefully to prevent such attacks. Our results offer valuable insights and recommendations

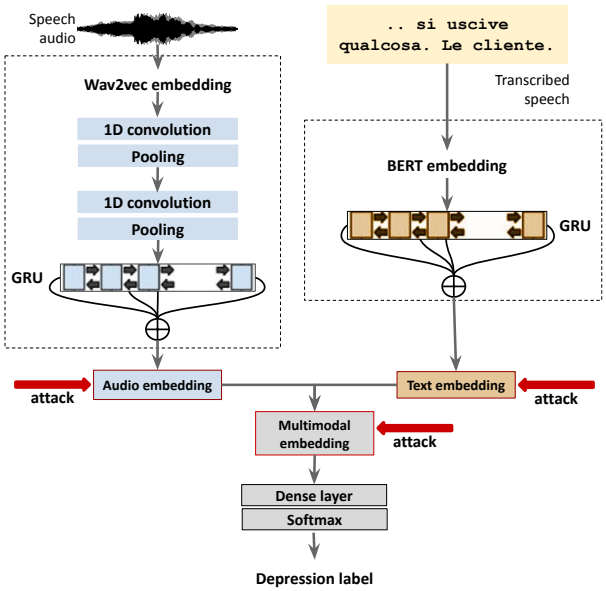


Figure 1: *Target model: ultimodal depression detection model*

to enhance the security and privacy of these systems, suggesting practical steps for improvement.

Figure 1 shows the architecture of our Target Model. It is a multimodal depression detection model that uses spoken data in the forms of both audio and text. The red 'attack' arrows indicate the layers from which the private embeddings are derived.

1. References

[1] Y. Elazar and Y. Goldberg, "Adversarial removal of demographic attributes from text data," in *Proceedings of EMNLP*, 2018, pp. 11–21.

[2] C. Biswas, D. Ganguly, P. Mukherjee, U. Bhattacharya, and Y. Hou, "Privacy-aware supervised classification: An informative subspace based multi-objective approach," *Pattern Recognition 2022*, vol. 122, p. 108301, 2022.

[3] B. M. L. Srivastava, A. Bellet, M. Tommasi, and E. Vincent, "Privacy-Preserving Adversarial Representation Learning in ASR: Reality or Illusion?" in *Interspeech 2019*, 2019, pp. 3700–3704.

[4] T. Feng, R. Peri, and S. Narayanan, "User-level differential privacy against attribute inference attack of speech emotion recognition in federated learning," in *Proc. Interspeech 2022*, 2022, pp. 5055–5059.

[5] M. Dias, A. Abad, and I. Trancoso, "Exploring hashing and cryptonet based approaches for privacy-preserving speech emotion recognition," in *Proc. ICASSP 2018*, 2018, pp. 2057–2061.

Just Because We Camp, Doesn't Mean We Should: The Ethics of Modelling Queer Voices.

Atli Sigurgeirsson[1,*], *Eddie L. Ungless*[2,*]

[1]The Centre for Speech Technology Research, University of Edinburgh, Scotland
[2]School of Informatics, University of Edinburgh, Scotland
s2063518@ed.ac.uk, s2112744@ed.ac.uk

Modern voice cloning models claim to be able to capture a diverse range of voices. We test the ability of a typical pipeline (30 second raw segments: **Segment** → shorter segmented raw utterances: **Raw-utt** → after voice enhancement: **VE-utt** → synthesis after voice cloning: **copy-synth**) to capture the style known colloquially as "gay voice". We use the phrase *gay voice* to refer to the phenomenon of "phonetic practices" [1] typically associated by listeners with gay men, for example sibilant quality [1]. We ask: can perceived *gay voice* be successfully modeled by end-to-end voice-cloning TTS models (Research Q1)? Any speaker identity modeling necessarily involves some loss, and we believe this will be exacerbated in the case of men with *gay voice*. We believe that because these models implicitly do not model *gay voice*, the resulting output will be poor in terms of fidelity to source voice.

We predict: (H1) Compared to the difference between ground truth and synthesised control speaker utterances, synthesised *gay voice* utterances will show a larger decrease in perceived gayness compared to ground truth utterances. We notice a homogenisation effect: synthesised speech is rated as sounding significantly "less gay" than its corresponding ground-truth for speakers with *gay voice*, but ratings actually increase for control speakers. In a sense, our prediction that `gay` speakers would show a larger decrease in ratings **(H1) is supported**, in that ratings actually increased for `control` speakers.

Because we believe perceived gayness is an important part of speaker identity, we also believe that this loss in perceived gayness will be associated with a loss in perceived speaker similarity. As such, we predict: (H2) Compared to the difference between ground truth and synthesised control speaker utterances, synthesised *gay voice* utterances will be rated as having a lower speaker similarity with ground truth utterances. We find that there is no significant difference in similarity accuracy between the groups, so **RQ1:H2 is not supported**. However, when we examine the correlation between *gay voice* rating and similarity, we notice for `gay` speakers there is a medium correlation between size of loss of *gay voice* and similarity accuracy and no such correlation for `control` speakers. It seems loss of perceived *gay voice* for `gay` speakers is more pertinent to similarity judgements than "loss" of "straight voice".

Finally, we explore which steps in a speech synthesis pipeline have the biggest impact on the fidelity of *gay voice* (Research Q2). Looking at average *gay voice* rating by clip type, we observe that ratings actually increase between **Segment** and **Raw-utt**, particularly for `gay` speakers. There is a large drop in ratings between **VE-utt** and **Copy-synth** for `gay` speakers and an increase for control speakers. In answer to **RQ2, synthesis has the biggest impact on *gay voice* ratings**, over using short, out-of-context clips, and voice enhancement.

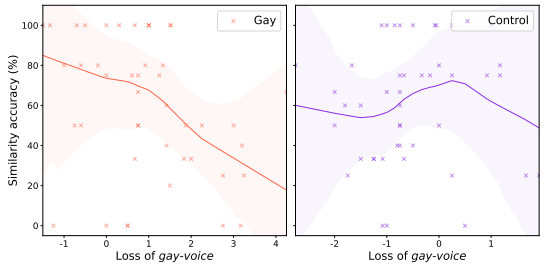


Figure 1: *Graphs showing the correlation between the similarity accuracy and reduction in gay voice between reference* ***Segments*** *and* ***Synth*** *(A positive value indicates a loss). Lines represent LOWESS-smoothing and shaded area indicates the 95% confidence intervals.*

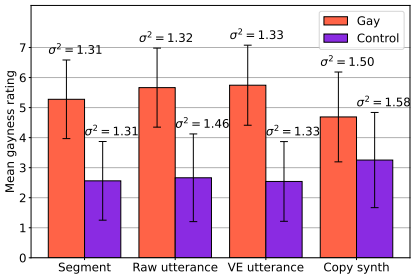


Figure 2: *Mean* gay voice *rating of speaker groups by clip type.*

We caution that improving the ability of such models to synthesise *gay voice* comes with a great number of risks. We use this pipeline as a starting point for a discussion on the ethics of modelling queer voices more broadly. Collecting "clean" queer data has safety and fairness ramifications, and the resulting technology may cause harms from mockery to death. Having a dataset of private individuals labeled for sexuality could be harmful if the data is made public. Further, a model may learn features of queer voices from training data, to create gaydars or "clockers", which could put queer individuals at fatal risk.

1. References

[1] S. Mack and B. Munson, "The influence of /s/ quality on ratings of men's sexual orientation: Explicit and implicit measures of the 'gay lisp' stereotype," *Journal of Phonetics*, vol. 40, no. 1, pp. 198–212, Jan. 2012.

This work was supported in part by Huawei and 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.

*These authors contributed equally to this work

Portability of Text- v Audio-based Auto-markers for Dialogic Assessment

Simon W. McKnight, Stefano Bannò, Mark J. F. Gales, Siyuan Tang, Kate M. Knill

ALTA Institute, Cambridge University Engineering Department,Trumpington Street, UK

{swm35, sb2549, mjfg100, st941, kmk1001}@cam.ac.uk

Dialogic, conversational, speaking tests are widely used to assess the ability of L2 English speakers. They more accurately reflect realistic conversational situations that candidates will encounter in English-speaking situations than monologic, prompt response, tests do. Several deployed speaking auto-markers for the latter exist, but dialogic auto-markers are still at the research stage. [1] proposed using neural model-based auto-markers for dialogic assessment following [2] for monologic assessment. In [1] a proprietary dataset was used for training and testing. The ICNALE Spoken Dialogue corpus [3] provides a publicly available evaluation set which should allow comparison of approaches from different groups. There is not enough data, however, for training a model. This research, therefore, investigates the portability of these neural auto-markers across datasets. Graders trained on a monologic test and on a dialogic test are compared.

Following [1], two types of regression auto-markers are primarily used here. Neural text graders take ASR transcripts as input to a text encoder followed by a linear layer. For the shorter monologic tests a BERT encoder (`BERT`) is used [2] and a Longformer (`LF`) for the longer dialogic tests [1]. A 5 model ensemble is used. The neural speech graders feed in relevant audio extracts to wav2vec2.0 with attention models (`wav2vec2`) with 2 layers of 4 attention heads (the first layer is intra-utterance and the second cross-utterance). The inputs are taken only from the candidate's part of the response or conversation. LLMs potentially allow grading based on very limited training data so zero-shot and few-shot trained GPT-4 based approaches using data from the ICNALE corpus are compared.

Several datasets are used in this research: (a) proprietary monologic datasets using the 5-part Linguaskill (LS) data [2]; (b) a proprietary dialogic dataset using Occupational English Test (OET) data, which are role-plays set in healthcare situations [1]; and (c) a public dialogic dataset using ICNALE Spoken Dialogue (ICNSD) data [3]. Linguaskill and OET data are split into separate train/dev/eval sets, and the auto-markers evaluated are trained on them (`BERT` on Linguaskill, `LF` on OET, `wav2vec2` on both). ICNSD is not a high-stakes test, but is particularly useful for evaluation as it contains 140 90s extracts of role-plays that are each marked by 80 raters and include strong/weak point comments by those raters. It also has accurate manual transcriptions. The OET dialogic sentence structure and style are more similar to ICNSD than the monologic Linguaskill structure and style are, but the Linguaskill vocabulary is more similar to the everyday ICNSD vocabulary than the healthcare-related OET vocabulary is.

Initial experiments showed that, unsurprisingly, neutral graders perform best when the training and evaluation data are matched. Table 1 shows selected pre-trained models performance on the unmatched case of ICNSD evaluation data (PCC = Pearson correlation coefficient, SRC = Spearman rank coefficient). The text-based models trained on Linguaskill port to ICNSD better than their corresponding audio-based models, but the converse is true for OET. Overall the LS open speaking neural graders do best, suggesting that for auto-marker portability the dialogic nature of the data is less important than the vocabulary used. Differing utterance lengths in the training and evaluation sets appear to be relatively unimportant.

Train Data		wav2vec2		BERT/LF	
Dataset	Type	PCC ↑	SRC ↑	PCC ↑	SRC ↑
LS Part 1	5 x 10-20s open responses	0.708	0.698	**0.797**	**0.823**
LS Part 2	Read speech	0.461	0.440	**0.710**	**0.741**
OET	2 x 5-min healthcare role-plays	**0.682**	**0.669**	0.476	0.513

Table 1: *Selection of pre-trained audio- v text-based model performance on 140 ICNSD participants.*

3 ICNSD participants were held out and used for few-shot training a recent decoder-only text-based model `GPT-4` [4], with mean ratings and all 80 strong/weak points submitted. Table 2 compares text-based model performances, showing: (a) `GPT-4` underperforms with zero-shot training but improves dramatically with 3 few-shot training; and (b) `GPT-4` is particularly robust to transcription errors.

Model	Few Shot	Eval Text	PCC ↑	SRC ↑
LS `BERT` Ens	-	Manual	**0.855**	0.844
		Automatic	0.785	0.815
`GPT-4` Ens	✗	Manual	0.599	0.578
	3	Manual	0.854	**0.881**
		Automatic	**0.853**	**0.876**

Table 2: *LS `BERT` and `GPT-4` performance on 137 ICNSD participants. 3 models in manual, 5 in automatic ensembles.*

1. References

[1] S. W. McKnight, A. Civelekoglu, M. J. F. Gales, S. Bannò, A. Liusie, and K. M. Knill, "Automatic assessment of conversational speaking tests," in *Proc. Workshop on Speech and Language Technol. in Education (SLaTE)*, 2023.

[2] S. Bannò, K. M. Knill, M. Matassoni, V. Raina, and M. J. F. Gales, "L2 proficiency assessment using self-supervised speech representations," in *Proc. Workshop on Speech and Language Technol. in Education (SLaTE)*, 2023.

[3] S. Ishikawa, "The ICNALE Spoken Dialogue: A New Dataset for the Study of Asian Learners' Performance in L2 English Interviews," *English Teaching*, vol. 74, no. 4, pp. 153–177, Dec. 2019.

[4] OpenAI, "OpenAI platform - text generation models," https://platform.openai.com, 2024.

This abstract reports on research supported by Cambridge University Press & Assessment (CUP&A), a department of The Chancellor, Masters, and Scholars of the University of Cambridge. The Linguaskill data was provided by CUP&A, the OET data by Cambridge Boxhill Language Assessment Trust (CBLA) and the ICNALE Spoken Dialogue data by Professor Shin'ichiro Ishikawa.

NOMAD: Unsupervised Learning of Perceptual Embeddings for Speech Enhancement and Non-Matching Reference Audio Quality Assessment

Alessandro Ragano[1], *Jan Skoglund*[2], *Andrew Hines*[1]

[1]University College Dublin, School of Computer Science, Ireland
[2]Google LLC, USA

alessandro.ragano@ucd.ie, jks@google.com, andrew.hines@ucd.ie

1. Method

Full-reference and no-reference speech quality metrics show some disadvantages. The former needs the clean counterpart while the latter assumes that the mean opinion score (MOS) is absolute, while studies have proven to be a relative scale [1]. To mitigate both no-reference and full-reference issues, we presented NOMAD (Non-Matching Audio Distance) at ICASSP 2024, a perceptual differentiable audio metric that predicts quality using non-matching references i.e., unpaired clean speech. Our method creates an embedding space where signals with similar degradation intensity are close. Our approach is based on contrastive learning with the triplet loss [2] guided by the Neurogram Similarity Index Measure (NSIM) [3] to model degradation intensity.

NOMAD relies on the assumption that audio quality is linked to degradation intensity. Given two values α_i and α_j, along with the corresponding degraded samples $\boldsymbol{x}_i = h(\boldsymbol{x}, \alpha_i)$ and $\boldsymbol{x}_j = h(\boldsymbol{x}, \alpha_j)$, the degradation parameter can be used as a label to learn a similarity function $\mathcal{D}(\cdot, \cdot)$ that follows the constraint:

$$\alpha_i > \alpha_j \implies \mathcal{D}(\boldsymbol{x}, \boldsymbol{x}_i) > \mathcal{D}(\boldsymbol{x}, \boldsymbol{x}_j) \quad (1)$$

To model this distance we employ the NSIM to choose anchor, positive, and negative samples for the triplet samples as shown in Figure 1. The model wav2vec 2.0 [4] is chosen as a backbone, given its capacity to adapt to several speech processing tasks.

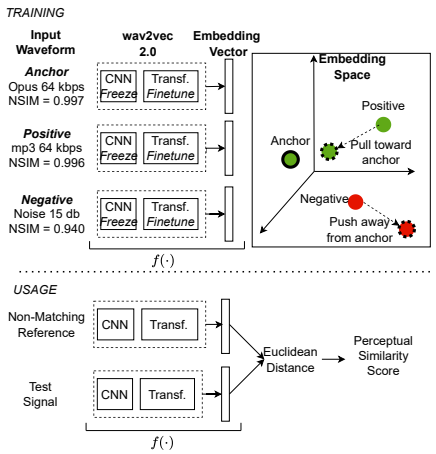


Figure 1: *Overview of the proposed method NOMAD.*

This work has emanated from research supported in part by the Google and research grants from Science Foundation Ireland (SFI) co-funded under the European Regional Development Fund under Grant Number 13/RC/2289_P2.

2. Results

NOMAD can be used with non-matching references since the output is a distance score emulating perception between any two samples. We choose to sample 899 clean sources and use them as non-matching references in our experiments.

Table 1: *PC and SC of non-matching reference (NMR) and full-reference (FR) models.*

		P23 EXP 1		P23 EXP 3		TCD-VoIP		GENSPEECH	
Type	**Model**	PC	SC	PC	SC	PC	SC	PC	SC
NMR	NOMAD	**-0.85**	**-0.88**	**-0.85**	**-0.75**	**-0.64**	**-0.64**	**-0.94**	**-0.90**
	w2v	-0.26	-0.27	-0.38	-0.36	-0.39	-0.54	-0.67	-0.90
	NORESQA	-0.24	-0.20	-0.46	-0.23	-0.11	-0.14	-0.69	-0.69
FR	NOMAD FR	-0.86	-0.87	-0.86	-0.73	-0.63	-0.65	**-0.96**	**-0.90**
	CDPAM	-0.48	-0.35	-0.39	-0.37	-0.76	-0.79	-0.93	**-0.90**
	ViSQOL	0.87	0.89	0.78	0.67	0.74	0.76	0.64	0.74
	WARP-Q	-0.88	-0.92	**-0.87**	-0.79	-0.90	**-0.92**	-0.89	**-0.90**
	PESQ	**0.91**	**0.96**	**0.87**	**0.87**	**0.91**	0.91	0.49	0.52

Results in Table 1 indicate that NOMAD is better than another non-matching reference model NOERESQA [5]. In addition, NOMAD shows competitive results with full-reference metrics without seeing MOS during training.

Since NOMAD is a differentiable metric, we also evaluate its performance as a perceptual training loss for speech enhancement. Results by finetuning the pre-trained DEMUCS model [6] (FT NOMAD) and multi-task learning with mean squared error (MT NOMAD) are shown in Table 2.

Table 2: *Speech enhancement performance evaluation.*

	PESQ↑				MUSHRA↑		
	2.5	7.5	12.5	17.5	2.5	7.5	12.5
Noisy	1.42	1.76	2.10	2.60	20 *(10,52)*	30 *(19,65)*	46 *(20,82)*
Demucs (*Baseline*)	2.40	2.83	3.06	3.31	58 *(39,78)*	78 *(51,90)*	84 *(58,90)*
FT Nomad (*Ours*)	**2.43**	**2.88**	**3.14**	**3.42**	70 *(50,82)*	80 *(50,90)*	88 *(57,91)*
MT Nomad (*Ours*)	2.42	2.84	3.10	3.36	**72** *(58,84)*	**90** *(63,94)*	**90** *(71,95)*

3. References

[1] S. Zielinski, F. Rumsey, and S. Bech, "On some biases encountered in modern audio quality listening tests - A review," *Journal of the Audio Engineering Society*, vol. 56, no. 6, pp. 427–451, 2008.

[2] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[3] A. Hines and N. Harte, "Speech intelligibility prediction using a neurogram similarity index measure," *Speech Communication*, vol. 54, no. 2, pp. 306–320, 2012.

[4] 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.

[5] P. Manocha, B. Xu, and A. Kumar, "NORESQA: A framework for speech quality assessment using non-matching references," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22 363–22 378, 2021.

[6] A. Défossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," *Proc. Interspeech*, pp. 3291–3295, 2020.

Enhancing ASR performance for dysarthric speech: Deep Learning approaches and dataset expansion

Leon Turner[1], *Eugenio Donati*[1]

[1]University of West London
leon.turner@uwl.ac.uk, eugenio.donati@uwl.ac.uk

Dysarthria is a neurological motor speech disorder that impacts an individual's ability to control the muscles involved in speech production, leading to reduced speech intelligibility. As a result, individuals with dysarthria often struggle with interacting and communicating with others in daily life. There are many types of dysarthria, each with its own unique characteristics and causes such as amyotrophic lateral sclerosis, Parkinson's disease, and cerebellar damage. The characteristics of dysarthria cause the produced speech to exhibit significant variation in its acoustics characteristics compared to non-dysarthric speech.

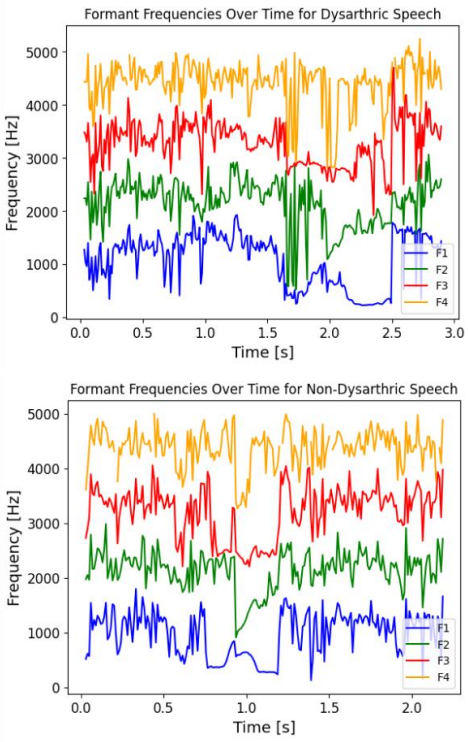


Figure 1: *Formant frequencies in dysarthric and non-dysarthric speech*

As illustrated in fig. 1, a greater variation in the formant frequencies is characteristic of dysarthric speech. The formants present significant anomalies and abrupt changes, reflecting the motor control issues associated with dysarthria when compared to typical speech motor control.

This study is based on the understanding that the existing speech recognition (ASR) systems are not optimized for dysarthria. These systems have potential to greatly improve the lives of individuals with dysarthria by providing them with a means to communicate and interact with others. Yet the current state of these technologies falls short in accurately recognizing dysarthric speech. This study aims to expand the available data and explore novel approaches to developing robust automatic speech recognition systems specifically designed for dysarthric speech. The first stage of this research will see the design and construction of the dataset, as it will provide the necessary materials to train and test the automatic speech recognition system. This will be done in partnership with organizations that specialize in treating individuals with dysarthria. The construction of the dataset will see the recording of keywords and general speech utterances from individuals affected by dysarthria at various degrees of severity. The collected data is then expected to be processed to isolate the relevant phonetic and acoustics characteristics. The second stage will use deep learning techniques such as convolutional neural networks and recurrent neural networks to create automated speech recognition models customized for dysarthric speech. It will explore enhancing existing model architectures or creating new ones that can capture the distinctive acoustic and phonetic features of dysarthric speech. Transfer learning will be used to make use of existing models trained on large-scale datasets of non-dysarthric speech, refining them specifically for dysarthria. These models will be optimized for real-world application by addressing computational efficiency and real-time processing constraints. Subsequently, untrained networks will be used to analyse their performances with dysarthric speech; this stage will involve the use of existing architectures normally employed in automatic speech recognition as well as investigating possible novel architectures tailored to the phonetic characteristics of dysarthria. The performance of the developed ASR models will be evaluated using standard evaluation metrics such as word error rate on both dysarthric and non-dysarthric speech datasets. This research is expected to have significant impact on the lives of individuals with dysarthria, allowing them to effectively communicate and interact with others. Furthermore, it can lead to impact in various domains, such as healthcare, education, and accessibility.

[1] The UKIS 2024 Committee would like to thank ISCA and Interspeech 2023 organisers for kindly providing this template.

Investigating the Role of Visual Information in Audio-Visual Speech Recognition

Zhaofeng Lin, Naomi Harte

Sigmedia Lab, School of Engineering, Trinity College Dublin, Ireland
linzh@tcd.ie, nharte@tcd.ie

Audio-Visual Speech Recognition (AVSR) combines auditory and visual speech cues to enhance the accuracy and robustness of speech recognition systems. While recent advancements in AVSR have achieved low word error rates in controlled, clean environments, state-of-the-art (SOTA) systems show improved performance in noisy environments compared to their audio-only counterparts. However, the extent to which these systems effectively leverage visual information remains unclear.

In this study, we assess the performance of three open-source SOTA AVSR systems: Auto-AVSR [1], Audio-Visual Efficient Conformer (AVEC) [2], and AV-RelScore [3]. These systems are evaluated on the LRS2 [4] test set under varying levels of pink noise and the results are shown in Figure 1. Our findings indicate that Auto-AVSR has the best overall performance, possibly due to its extra training dataset, which exceeds 4,000 hours, compared to AVEC with 818 hours and AV-RelScore with 195 hours.

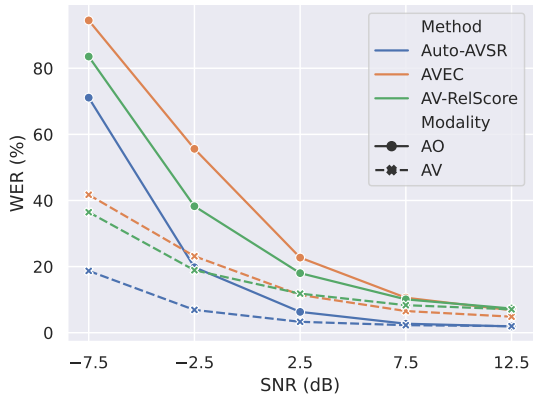


Figure 1: *Performances of six models (three methods, both audio-only (AO) and audio-visual (AV) systems) on the LRS2 test set under different noise levels when the noise is pink noise.*

To investigate the influence of visual information on WER, we categorise words based on their visual salience. Words containing more visible mouth movements (e.g., lip-to-teeth, lips puckered, lips together) are labelled 'Hi-Vis', while those with less visible articulation are labelled 'Low-Vis'. Table 1 displays the WERs for each model in clean and noisy environments, with results for both Hi-Vis and Low-Vis words. In clean environments, audio-only systems tend to perform better, suggesting that visual cues may not always be beneficial. However, in noisy conditions, AV systems generally exhibit lower WERs, indicating that visual information can improve recognition accuracy.

Table 1: *WERs (%) of six models on Hi-Vis and Low-Vis words under clean and noisy (pink noise -2.5 dB) environments*

		Auto-AVSR		AVEC		AV-RelScore	
		Hi-Vis	Low-Vis	Hi-Vis	Low-Vis	Hi-Vis	Low-Vis
Clean	AO	0.89	0.83	1.98	2.09	4.4	3.91
	AV	1.19	0.88	2.78	1.71	4.99	3.85
Noisy	AO	12.85	11.6	47.02	39.86	30.45	29.54
	AV	5.15	4.93	17.77	18.61	13.88	15.7

Interestingly, while Auto-AVSR has the lowest overall WER, it performs worse on Hi-Vis words compared to Low-Vis words. In contrast, both AVEC and AV-RelScore exhibit lower WERs for Hi-Vis words. This finding indicates that while Auto-AVSR benefits from a larger training dataset, this model might not exploit the visual information as well as the other two. These results highlight the complexities of integrating visual information into AVSR systems and suggest that visual cues can play a crucial role in improving speech recognition accuracy, particularly in noisy environments. However, how these cues are utilised varies across different models, indicating opportunities for further research and development. These results are still under research, and we are hoping to give more details in the presentation.

1. References

[1] P. Ma, A. Haliassos, A. Fernandez-Lopez, H. Chen, S. Petridis, and M. Pantic, "Auto-avsr: Audio-visual speech recognition with automatic labels," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[2] M. Burchi and R. Timofte, "Audio-visual efficient conformer for robust speech recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 2258–2267.

[3] J. Hong, M. Kim, J. Choi, and Y. M. Ro, "Watch or listen: Robust audio-visual speech recognition with visual corruption modeling and reliability scoring," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 783–18 794.

[4] J. Son Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6447–6456.

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.

Has Artificial Intelligence Rendered Language Teaching Obsolete?

Zoe Handley

[1]The University of York
zoe.handley@york.ac.uk

Abstract

In his recent article Twenty-first century technologies and language education, Kern [1] identifies a number of possible directions language learning and teaching might take as a result of recent developments in Artificial Intelligence (AI). With an interest in how AI, specifically speech technologies, might be harnessed to facilitate language learning and teaching that dates back to the first wave of AI, I was particularly interested in Kern's perspectives on the impact of the second wave of AI on language learning and teaching. As Kern highlights, language and technology are intrinsically related and this has never more so been the case than for AI. As such, Kerns suggests that AI poses an existential threat to language teaching and the need to learn a language in the first place. In this paper, I explore this claim. Having explored what makes a good language teacher [2 - 4], I examine how far the current state of the art in AI has come to replacing language teachers. A particular emphasis is placed on examining how popular language learning apps and software packages, namely Duolingo [5 - 7], Grammarly [8] and similar software [9, 10], and Rosetta Stone [6, 11], use AI technologies to carry out the different functions of a language teacher. Further to this, evidence for the effectiveness of these apps and software packages.

Index Terms: Computer-Assisted Language Learning (CALL), learning analytics, Automated Writing Evaluation (AWE), Automated Speech Evaluation (ASE)

1. What makes a good language teacher?

According to the COACTIV model of teachers' professional competence [2], effective teachers possess not only well-developed pedagogical knowledge, including subject specific as well as general pedagogical knowledge, but also organizational and counselling knowledge. Language teaching is distinct in terms of both the pedagogical knowledge and counseling knowledge required of the teacher because the subject is the medium of instruction and learning a language is acknowledged to be particularly anxiety-provoking [3 - 4].

2. Some popular language learning apps

2.1. Duolingo

Duolingo is an intelligent tutoring system that integrates generative AI to automatically generate exercises, machine translation to generate feedback, and learning analytics to tailor instruction to the individual learner. See [5 – 7] for evaluations.

2.2. Grammarly

Grammarly deploys a range of natural language processing to provided automated writing feedback. See [8 – 10] for evaluations.

2.3. Rosetta stone

Rosetta Stone deploys Automatic Speech Recognition (ASR) to provide feedback and pronunciation and interactive conversational practice. See [6, 11] for evaluations.

3. Evaluations of language learning apps

Despite advances in AI, the capacities of language learning apps remain limited compared to human language teachers. Intelligent language tutors not only lack the ability to engage [6, 8] and counsel learners, but are also limited in terms of pedagogy [6, 11] and their ability to represent language in a way that supports learners' understanding [7]. Current intelligent language tutors at best complement expert language tutors by offering unlimited repetitive practice and feedback on linguistic form that makes time for expert tutors to engage with knowledge and meaning, provide more creative opportunities to use language and engage with the whole learner at a human level.

4. References

[1] Kern (2024). Twenty-first century technologies and language education: Charting a path forward. The Modern Language Journal.
[2] Baumert, J., Kunter, M. (2013). The COACTIV Model of Teachers' Professional Competence. In: Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S., Neubrand, M. (eds) Cognitive Activation in the Mathematics Classroom and Professional Competence of Teachers. Mathematics Teacher Education, vol 8. Springer, Boston, MA.
[3] Borg, S. (2006). The distinctive characteristics of foreign language teachers. Language teaching research, 10(1), 3-31.
[4] Andrews, S. (2003). Teacher language awareness and the professional knowledge base of the L2 teacher. Language awareness, 12(2), 81-95.
[5] Vesselinov, R., & Grego, J. (2012). Duolingo effectiveness study. City University of New York, USA, 28(1-25).
[6] Loewen, S., Crowther, D., Isbell, D. R., Kim, K. M., Maloney, J., Miller, Z. F., & Rawal, H. (2019). Mobile-assisted language learning: A Duolingo case study. ReCALL, 31(3), 293-311.
[7] Marques-Schafer, G., & da Silva Orlando, A. A. (2018). Languages learning conceptions and Duolingo: A critical analysis on its proposals and learners experiences. Texto Livre-Linguagem e Tecnologia, 11(3), 228-251.
[8] Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study. Assessing Writing, 44, 100450.
[9] Ranalli, J. (2021). L2 student engagement with automated feedback on writing: Potential for learning and issues of trust. Journal of Second Language Writing, 52, 100816.
[10] Chapelle, C. A., Cotos, E., & Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. Language testing, 32(3), 385-405.
[11] Vesselinov, R. (2009). Measuring the effectiveness of Rosetta Stone. Rosetta Stone.

Effective Context in Self-Supervised Speech Models

Yen Meng, Hao Tang

The Centre for Speech Technology Research, University of Edinburgh, United Kingdom
{yen.meng, hao.tang}@ed.ac.uk

1. Effective context

LSTMs and Transformers are widely used for various speech processing tasks, owing to their ability to make use of long context. Though these models take the entire utterance (or longer) as input, how much context a model actually uses is largely unknown. In this work, we study the **effective context** of several speech models, supervised or self-supervised. Being able to measure the amount of effective context can not only improve our understanding of a model but also open up opportunities to improve models.

At a conceptual level, effective context of a function is the parts of an input that can significantly change the output. The effective context by this definition can be measured by the Jacobian matrix of the function of interest. Several prior studies have adopted similar ideas, including measuring the function change while manipulating the input (Zeiler and Fergus, 2014; Khandelwal *et al.*, 2018; O'Connor and Andreas, 2021). The closest to our work is Tang and Glass (2018), where effective context is measured by taking the gradient from the loss to the input. Gradients to the input are also heavily used in model attribution (Simonyan *et al.*, 2013). The Jacobian is also closely related model certification and adversarial robustness (where a function is overly sensitive to certain context).

Formally, for an input utterance $x = (x_1, \ldots, x_T)$ where each x_t is a frame of dimension K. We are interested in measuring the context of a function f such that $h_1, \ldots, h_T = f(x_1, \ldots, x_T)$ where $h_t \in \mathbb{R}^D$. We denote the **influence** of h_t to x_τ as $s(t, \tau)$ and define $s(t, \tau)$ as the Frobenius norm of the Jacobian of h_t at x_τ. In other words,

$$s(t,\tau) = \sqrt{\sum_{k=1}^{K}\sum_{d=1}^{D}\left(e_d^\top \frac{\partial h_t}{\partial x_{\tau,k}}\right)^2} \tag{1}$$

where e_d is d-th element of the standard basis. This definition is general in several aspects: it can be applied to various neural architectures independent of output frame rates, such as LSTMs, Transformers, and CNNs; it can be applied to the loss function and would fall back to the usual ℓ_2 norm of the gradient; it can also be applied to intermediate layers and is independent of the loss function.

We accumulate the influence at various time shifts for all utterances in a data set. For a data set of N utterances, each of which has T_n frames with the measured influence $s_n(t, \tau)$ of f, for $n = 1, \ldots, N$, the accumulated influence at time shift τ is defined as

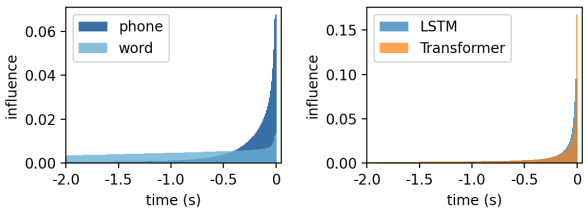
$$S(\tau) = \sum_{n=1}^{N}\sum_{t=1}^{T_n} s_n(t, t+\tau). \tag{2}$$

For example, if $S(-w)$ is high, then the influence of hidden vectors on the input that is w frames before is, on average, high. We accumulate the influence for time shifts $\tau = -W, -W+1, \ldots, W$. Finally, the effective context is defined as the time shift that covers a certain amount of the overall influence.

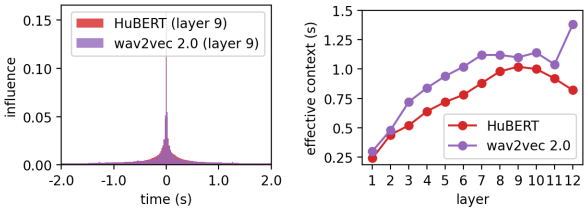
2. Results

We measure effective context on the `dev93` subset of the Wall Street Journal data set. We measure the effective context up to 2 seconds (200 frames for 10 ms models and 100 frames for 20 ms models). The effective context is defined as the time shift that covers 70% of the accumulated influence.

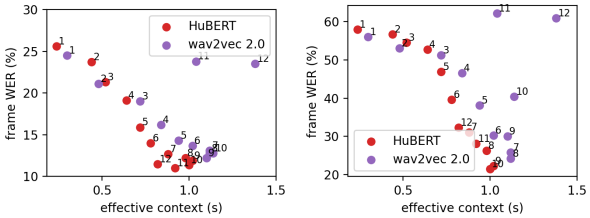
The figure on the left below compares the accumulated influence of two frame classifiers at their last layers, both based on 3-layer unidirectional LSTMs, one predicting phones, and one predicting words. The influence values are normalized to a distribution for the plots. It's clear that the model predicting words uses more context than the one predicting phones. The figure on the right below compares a 3-layer LSTM and a 6-layer causal Transformer trained with a self-supervised objective, autoregressive predictive coding (APC), predicting the 5th frame into the future. Both APC models behave similarly to a phone classifier (on the left).



The figure on the left below compares the 9th layer of HuBERT and wav2vec 2.0 (where phonetic information is the most prominent). The shape of the accumulated influence is again similar to a phone classifier but doubled. The figure on the right below shows how the effective context changes across Transformer layers. The effective context overall increases from low to high layers.



Finally, we probe the phone and word information across the Transformer layers of HuBERT and wav2vec 2.0. There are strong correlations between effective context and the probing error rates. There seems to be a minimum amount of effective context that a model needs to achieve in order to predict phones and words well.



Overall, the accumulated influence is extremely sharp for all models and has a long tail, showing that self-supervised models, despite having 12 layers of self-attention, still have deficiency in modeling context and only have a rudimentary understanding of words.

NLP and Speech Processing to study Mental Health Recovery Narratives from NEON (Narrative Experiences Online collection)

Shrankhla Pandey[1,a], *Sarah Morgan*[1,2], *Dr Stefan Rennick-Egglestone*[3], *Prof Mike Slade*[3]

[1]University of Cambridge, UK, [2]King's College London, UK, [3]University of Nottingham, UK

[a]sp2147@cam.ac.uk

Abstract: Mental health recovery narratives have potential to help people experiencing mental health problems. However, currently narratives need to be manually annotated before they can be recommended to potential readers. In this ongoing research project, I aim to build an automated system to annotate recovery narratives, using both NLP and acoustic speech features. To that end, I am using the NEON dataset- a large, diverse online repository of recovery narratives.

Background:

- Computational techniques (ML, DL, and NLP) show immense promise in revolutionizing mental health research, but there is a gap in self-management. A randomised controlled trial [1] found that receiving mental health recovery narratives provides some recovery-oriented benefits, including increases in quality of life and the presence of meaning in life.
- Majority of the NLP in mental health research is done on data from social media, electronic health records, surveys. Majority of speech research is based on picture description tasks, free-speech. There is a gap in research on ML models built on first-person narratives and interviews.

NEON Dataset:

The Narrative Experiences Online (NEON) [2] study has assembled a large and diverse online repository of mental health recovery stories (the NEON Collection). All narratives in the NEON Collection are manually annotated into characteristics such as genre, tone using the Inventory of Characteristics of Recovery Stories (INCRESE) [3] tool. The NEON team at the University of Nottingham devised methods of curating such collection and studied the impact of reading recovery narratives through randomized controlled trials (RCTs).

Our Goals:

- Build automated methods to predict INCRESE characteristics from the text stories and audio interviews for new recovery narratives in seven dimensions, such as genre, tone, trajectory using ML pipeline as shown in Figure 2

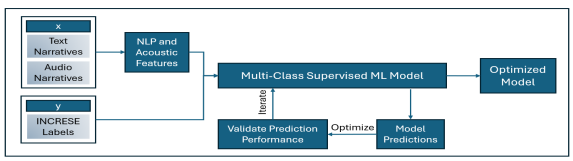


Figure 1: *Supervised Machine Learning Pipeline to predict INCRESE characteristics from NEON narratives*

Our Research:

Descriptive Statistics: The NEON collection currently has 602 stories containing text, 67 stories containing audio. The text stories have a mean length of 1520 words; the shortest story is 8 words long, and the longest is almost 20000 words. The audio stories have a mean length of 22 minutes; the shortest is 1 minute and the longest is 85 minutes

NLP Features:

- For our baseline model to predict one of the INCRESE characteristics, genre from text narratives (n=324), we used the bag of words with Naive Bayes [f1-score =0.39] and Logistic Regression [f1-score=0.54])
- Next, we will extract gradient measures from the trajectories of sentence wise change of the two features (1) cosine similarity score between the BERT embeddings of characteristic type and narratives (2) positive, neutral, negative sentiment

Speech Processing Features: We are using the following features to predict INCRESE characteristic of a narrative

- (1) Local and global speech graph (as shown in Figure features using netts [4]. (2) paralinguistic features (3) OpenSmile Compare2016 acoustic features, using dimensionality reduction techniques

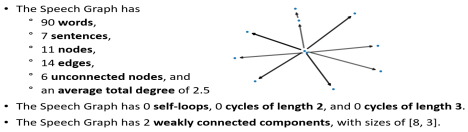


Figure 2: *Semantic Speech Graph, along with its global and local features, created from a transcript of a NEON narrative*

Conclusion: In conclusion, faster, automated annotations of recovery narratives could enable a wider range of narratives to be included in future recommender systems. It is important that the risks associated with automated annotations are addressed in this work and also discussed among patients, and their carers.

1. References

[1] Slade et al. Effectiveness and cost-effectiveness of online recorded recovery narratives in improving quality of life for people with non-psychotic mental health problems.

[2] Slade et al. Recorded Mental Health Recovery Narratives as a Resource for People Affected by Mental Health Problems.

[3] Llewellyn-Beardsley et al. INCRESE: Development of an Inventory to Characterize Recorded Mental Health Recovery Narratives.

[4] Nettekoven et al. Semantic Speech Networks Linked to Formal Thought Disorder in Early Psychosis.

On Robustness of Speaker Retrieval in the Wild: A Comparative Study of x-vector and ECAPA-TDNN Models

Erfan Loweimi, Mengjie Qian, Kate Knill, Mark Gales

Speech Group, Machine Intelligence Lab, University of Cambridge, UK
{el584,mq227,kmk1001,mjfg100}@cam.ac.uk

There is a growing abundance of publicly available or company-owned audio/video archives, highlighting the increasing importance of efficient access to desired content and information retrieval from these archives. In this extended abstract, we briefly discuss the challenges, solutions, effectiveness, and robustness of speaker retrieval systems developed "in the wild" which involves addressing two primary challenges: extraction of task-relevant labels from limited metadata for system development and evaluation, as well as the unconstrained acoustic conditions encountered in the archive, ranging from quiet studios to adverse noisy environments.

While we focus on the publicly-available BBC Rewind archive (specifically the Northern Ireland subset with 12,594 videos, 409 h data, spanning 1948 to 1979), our framework (Fig. 1) addresses the broader issue of speaker retrieval on extensive and possibly aged archives with no control over the content and acoustic conditions. Typically, these archives offer a brief and general file description, mostly inadequate for specific applications like speaker retrieval. In addition, manual annotation of such large-scale archives is unfeasible.

In the context of EPSRC-funded MVSE (multi-modal video search by example) [1] project which aims at speaker (as well as face/emotion/topic/content) retrieval with clips rather than text, we explore various aspects of system development such as diarisation, speaker embedding extraction, efficiency, combining various embeddings and usefulness of multimodality. We analyse the challenges, possible solutions, and their functionality.

To evaluate the performance, we conduct systematic experiments in both clean setup (Table 1) and against various distortions simulating real-world applications including various background noise at different SNRs (Fig. 2), sampling rate reduction (Table 2), reductions in bit-depth (Table 3) and reverberation using real and synthetic room impulse responses (Table 4).

Our findings demonstrate the effectiveness and robustness of the developed speaker retrieval systems, establishing the versatility and scalability of the proposed framework for a wide range of applications beyond the BBC Rewind corpus.

Index Terms: Speaker retrieval, speaker embedding, ECAPA-TDNN, development in the wild, noise robustness

Table 1: *Performance of the speaker retrieval in the clean setup.*

Feature	P@1	P@5	P@10	MAP@10	NDCG@10
x-vector	73.4	63.5	54.7	72.5	58.7
ECAPA	86.3	81.2	77.1	85.9	79.1

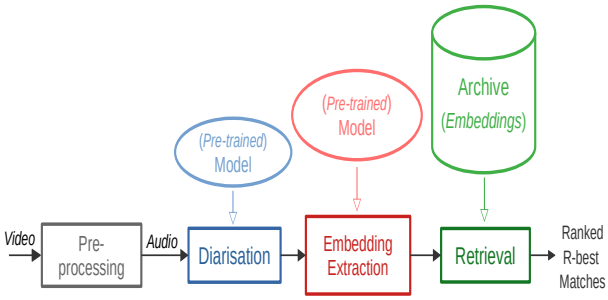


Figure 1: *Workflow of the speaker retrieval system.*

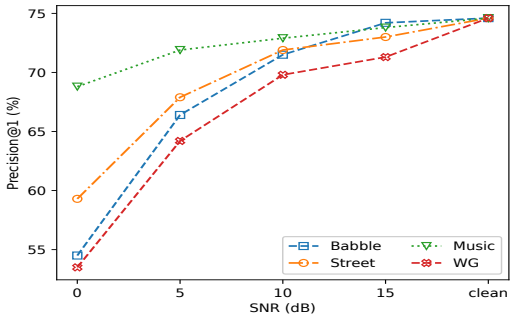


Figure 2: *P@1 vs SNR for ECAPA-TDNN on various noise.*

Table 2: *Robustness against 8 kHz sampling rate.*

	P@1	P@3	P@5	P@10	AvgRPR
xvector	38.4	31.6	28.0	22.7	-53.7
ECAPA	80.9	77.3	74.8	69.8	-7.8

Table 3: *Robustness against quantisation with bit-depth of 8.*

	P@1	P@3	P@5	P@10	AvgRPR
xvector	69.4	63.3	57.8	50.2	-6.9
ECAPA	83.4	81.6	80.0	75.7	-2.3

Table 4: *Robustness against reverberation. Database: AIR. RT60 of Office and Lecture rooms are 0.48 and 0.79 seconds.*

Embedding	Room (RT60)	P@1	P@3	P@5	P@10
xvector	Office	40.8	32.2	28.2	22.9
xvector	Lecture	25.7	19.1	16.4	12.6
ECAPA-SB	Office	83.4	80.5	77.9	73.3
ECAPA-SB	Lecture	81.6	78.0	75.4	71.2

1. References

[1] H. Wang, M. Mulvenna, R. Bond *et al.*, "Multimodal Video Search by Examples (MVSE)," EPSRC Grant Reference EP/V002740/2.

Wearable Audio-Visual AI for Use in Hearing Aids

Poppy Welch and Jennifer Williams

School of Electronics and Computer Science, University of Southampton, UK

p.welch@soton.ac.uk, j.williams@soton.ac.uk

Abstract

We present our work in-progress toward understanding the key issues of audio-visual (AV) hearing aid development. We identify these key issues as: 1) personalisation, 2) edge-computing constraints, and 3) ethical and privacy considerations. As an important research area with far-reaching societal consequences - both for the hearing impaired community and the public at large - addressing these key areas is critical. In this work, we discuss why these issues are challenging to address and their impact for speech research and beyond.

1. Introduction

Over 430 million people worldwide are affected by some degree of hearing loss [1], with the current uptake of hearing assistive devices estimated to be at approximately 11% [2]. This is typically due to the poor performance of commercially available hearing aids, especially in noisy scenarios. Human speech perception is inherently multimodal with visual cues being used to enhance the acoustic signal, in the form of lip reading for example [3]. This has led to an increase in research into AV speech enhancement, as many of these cognitively-inspired multimodal models outperform the audio-only ones [4].

Despite recent advancements [5, 6], there are still several open and unsolved issues to consider before these devices can be developed at scale, widely adopted, and made safe for users and society. Our approach to AV hearing aid development covers three key research questions:

RQ1: How can AV hearing aids be personalised to their users in order to achieve the best individual performance?

RQ2: How can an AV hearing aid be developed to fit with limited on-device constraints?

RQ3: What ethical and privacy considerations need to be made with AV wearable devices in order to protect secondary users?

The need for personalisation in speech technology has become especially clear during the past few years - from speaker-aware ASR to customised synthetic voices. In hearing aid technology, we must address the ways that different degrees of hearing loss affect each individual, and consider that preferences for using a hearing assistive device may differ among individuals in various sound environments. To achieve this, we need algorithms that can identify and adapt appropriately to different sound environments, allowing the user to have an element of control.

Wearable devices such as hearing aids have severe implementation requirements as edge-devices, and must be energy efficient and compact with low latency for real-time processing. Typical audio-only hearing aids can already be energy intensive and the use of multi-modal data exacerbates this problem further. Therefore low-power and low-complexity models must be developed for real-world application.

Privacy and ethical considerations are an essential component of AV hearing aids and should be at the forefront of design, in particular when considering secondary users. The use of these devices from a first-person perspective captures an extensive amount of rich information, with the potential risk of severe privacy violations. The use of both audio and visual streams allows for easy identification of individuals in the user's surroundings and the recording of potentially private or sensitive conversations. There is therefore a need to balance the invasion of secondary users' autonomy and the benefits of using an assistive medical device for a disability.

It is essential that significant advancements are made in this area, with a predicted 2.5 billion people expected to be affected by hearing loss by 2050 [1]. The research proposed has far reaching implications not only for the considerable number of potential users, but also for those interacting with them.

2. Discussion

It remains an open research question as to whether all three of these issues can be addressed in a single system architecture. While the issues we describe are important, it may be necessary to examine each of them independently before arriving at an integrated solution. At the same time, they are interrelated as privacy can be viewed on a spectrum of personalisation, and priorities toward privacy may differ across individuals.

3. References

[1] World Health Organization et al., "Hearing screening: Considerations for implementation," 2021.

[2] N. Bisgaard, S. Zimmer, M. Laureyns, and J. Groth, "A model for estimating hearing aid coverage world-wide using historical data on hearing aid sales," *International journal of audiology*, vol. 61, no. 10, pp. 841–849, 2022.

[3] L. D. Rosenblum, *Primacy of Multimodal 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] 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.

[5] L. A. Passos, J. P. Papa, A. Hussain, and A. Adeel, "Canonical cortical graph neural networks and its application for speech enhancement in audio-visual hearing aids," *Neurocomputing*, vol. 527, pp. 196–203, 2023.

[6] M. A. Almaiah, A. Ali, R. Shishakly, T. Alkhdour, A. Lutfi, and M. Alrawad, "A novel federated-learning based adversarial framework for audio-visual speech enhancement," *Journal of Theoretical and Applied Information Technology*, vol. 102, no. 4, 2024.

Towards Multimodal Turn-taking for Naturalistic Human-Robot Interaction

Sam O'Connor Russell, Naomi Harte

ADAPT Centre, Dept. of Electronic and Electrical Engineering, Trinity College Dublin, Ireland

russelsa@tcd.ie, nharte@tcd.ie

At key moments in conversation, the floor is either *held* by the speaker or *shifted* to the listener [1]. Thus, robots must make fast and accurate hold/shift (H/S) decisions to converse naturally with humans. Recent turn-taking models enable fast H/S decisions [2]. However, this involved telephone speech. As visual cues provide a wealth of turn-taking information when participants can see one another [1], our work investigates if they should be incorporated into turn-taking models.

We use the Switchboard corpus (260 hours of telephone speech) [3] and the Candor corpus (850 hours of videoconferencing audio and video recordings) [4]. We use the Speechmatics ASR tookit to transcribe Candor, as the provided alignment is inaccurate. We also obtained an equivalent ASR transcription for Switchboard to ensure a fair comparison.

We extract facial action units (FAUs) from the Candor videos using OpenFace [5]. We compute the percentage of frames in which each FAU is active in 500 ms windows. Windows are considered before and after shifts, before and after holds, and during random periods. We use a Mann-Whitney U test to compare median percentages. The visual channel is clearly used during shifts and holds. We see in Table 1, that interlocutor lip, jaw and mouth movements are more likely to occur just before a shift (i.e. before a speaker begins to speak) than at any other period when not speaking. This indicates that the visual modality is used to signal in advance that an interlocutor wishes to 'grab the turn'. We find less pronounced differences in FAU activations for holds. This could indicate that holds are less reliant on the visual channel.

	Comparing facial action units in Candor (videoconference) when				
Facial action unit	*Not speaking vs. when speaking*	*Not speaking vs. before a shift*	*Random speech vs. after a shift*	*Random speech vs. before a hold*	*Random speech vs. before a hold*
Inner brow raiser		+ 6%	+1%	+4%	4%
Cheek raiser		+8%	+9%	+2%	+2%
Nose wrinkler			+4%	+2%	+3%
Upper lip puller		+11 %	+9%		
Lip corner puller		+11 %	+12%		
Jaw drop	+ 10%	+17%	+8%		
Mouth stretch	+ 10 %	+8%			
Blink	+ 4 %	+6%			

Table 1: *This table shows that certain facial action units are more likely to be observed during turn-taking events. For example, the jaw drop FAU is observed 17% more when the interlocutor is about to speak (i.e. just before a shift) than when the interlocutor is not about to speak. Only statistically significant results are displayed.*

We investigate the performance of a state-of-the-art turn-taking model. The model was first introduced by Ekstedt and Skantze (see [2]). It is a causal, audio-only transformer-based turn-taking model. We compare performance using the audio from interactions where interlocutors can and cannot see one another (Candor and Switchboard). We train two separate models with: 1) Switchboard (telephone) audio and 2) Candor (videoconference) audio. We withhold 20% of sessions for testing.

Table 2: *F1 values for shift (S) and hold (H) predictions of models trained on the audio from the Switchboard or Candor corpus. We also compare the performance of a model trained on one corpus and deployed on another. The drop in performance drop indicates audio cues are used differently in these corpora.*

		Balanced F1 Score	
Model trained on	**Model deployed on**	**Hold**	**Shift**
Switchboard (telephone)	*Switchboard (telephone)*	0.89	0.57
Candor (videoconference)	*Candor (videoconference)*	0.82	0.55
Switchboard (telephone)	*Candor (videoconference)*	0.73 (↓ 11%)	0.52 (↓ 5%)
Candor (videoconference)	*Switchboard (telephone)*	0.85 (↓ 3.4%)	0.31 (↓ 40%)

We identify all periods of silence greater than 200 ms with a speaker change (a shift) and where the speaker stays the same (a hold). We report model performance in Table 2. We find a notable performance drop when we deploy the Switchboard-trained model on the Candor corpus and vice-versa (Table 2). The F1 score drops by 11% for holds (Switchboard-trained on Candor) and by 40% for shifts (Candor-trained on Switchboard, Table 2).

We are working on the following hypothesis to explain this finding: as interlocutors can see one another in Candor, they change the way in which they exploit the audio channel for signalling their intentions. Thus, a telephone-speech trained model cannot predict turn-taking accurately by using audio from a videoconferencing interaction and vice-versa.

1. References

[1] G. Skantze, "Turn-taking in Conversational Systems and Human-Robot Interaction: A Review," *Computer Speech & Language*, vol. 67, p. 101178, May 2021.

[2] E. Ekstedt and G. Skantze, "Voice activity projection: Self-supervised learning of turn-taking events," *INTERSPEECH 2022*, pp. 5190–5194, 2022.

[3] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, speech, and signal processing, ieee international conference on*, vol. 1. IEEE Computer Society, 1992, pp. 517–520.

[4] A. Reece, G. Cooney, P. Bull, C. Chung, B. Dawson, C. Fitzpatrick, T. Glazer, D. Knox, A. Liebscher, and S. Marin, "The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation," vol. 9, Mar. 2023.

[5] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *IEEE Winter Conference on Computer Vision*, 2016.

Parameter Efficient Finetuning for Speech Emotion Recognition

Nineli Lashkarashvili, Wen Wu, Guangzhi Sun, Phil Woodland

Department of Engineering, University of Cambridge, UK
{nl438, ww368, gs534, pw117}@cam.ac.uk

Speech emotion recognition (SER) plays a crucial role in human-computer interaction. SER has garnered growing interest in recent years. However, data scarcity remains a critical challenge for its advancement. The development of foundation models has paved the way to a new paradigm, enabling the customisation of general-purpose models for specific downstream tasks. Foundation models are large machine learning models trained on a very large quantity of data usually by self-supervised learning. Transferring knowledge from pre-trained foundation models can help alleviate the data sparsity issue of downstream tasks. However, finetuning the whole foundation model is resource-consuming and can lead to over-fitting and catastrophic forgetting. Parameter-efficient fine-tuning (PEFT) algorithms have been proposed to help mitigate these issues, which only fine-tune a small number of (extra/internal) model parameters while keeping most pre-trained foundation model parameters frozen. PEFT has been shown to be successful for natural language processing tasks and also some speech tasks. However, the efficacy of PEFT methods for SER hasn't been extensively explored.

We investigate the use of different PEFT adaptors and their combinations for SER. Various adaptors are studied including the Bottleneck Adaptor (BA) [1], Low-Rank Adaptation (LoRA) [2], the Weighted Sum (WS) [3], and a proposed new adaptor called the Weight-Gating (WG) adaptor which assigns a learnable gating vector $\mathbf{g} \in \mathbb{R}^d$ to each hidden state of dimension d which is passed to a sigmoid activation function. PEFT modules are inserted in the Transformer encoders in the foundation models as shown in Figure 1.

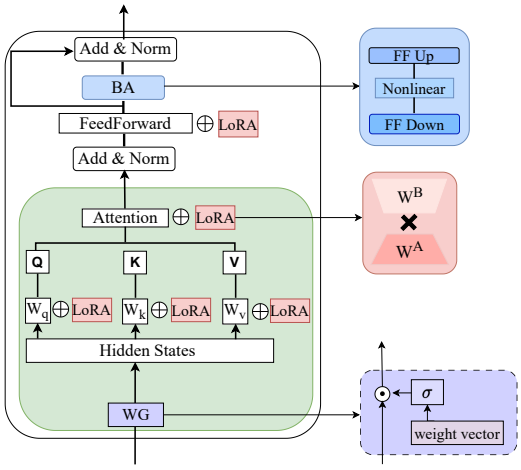


Figure 1: *A Transformer block with various PEFT adaptors.*

Both the wav2vec 2.0 base [4] and the HuBERT large [5] foundation models were tested. Experimental results on IEMOCAP [6] are shown in Figure 2. It can be seen that all PEFT methods, except for WG in HuBERT, exhibit superior performance compared to the pre-trained baseline. Augmenting the combination of BA, LoRA, and WS with WG yields an overall performance boost that surpasses fine-tuning (FT) while using 2-3% of the total trainable parameters for both models. In conclusion, PEFT can achieve comparable performance to full fine-tuning while requiring far fewer updated parameters. The combination of various PEFT adaptors results in enhanced system performance, demonstrating comparable or even superior outcomes to state-of-the-art models without relying on external features beyond raw audio data.

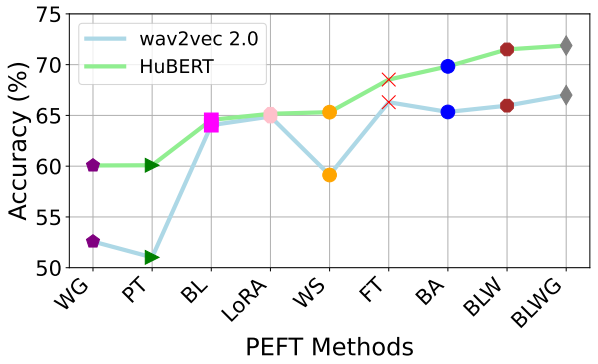


Figure 2: *BL (BA + LoRA), BLW (BA + LoRA + WS), BLWG (BA + LoRA + WS + WG), and others.*

1. References

[1] N. Houlsby *et al.*, "Parameter-efficient transfer learning for NLP," in *Proc. ICML*, 2019.

[2] E. J. Hu *et al.*, "LoRA: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[3] S.-w. Yang *et al.*, "SUPERB: Speech processing universal performance benchmark," *arXiv preprint arXiv:2105.01051*, 2021.

[4] A. Baevski *et al.*, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Proc. NeurIPS*, 2020.

[5] W.-N. Hsu *et al.*, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *TASLP*, vol. 29, pp. 3451–3460, 2021.

[6] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.

Tracking Articulatory Feature Transition Phenomena in Speech Embeddings

Patrick Cormac English[1,2], *Erfan A. Shams*[2], *John D. Kelleher*[3], *Julie Carson-Berndsen*[2]

[1] SFI Centre for Research Training in Digitally-Enhanced Reality (d-real), Ireland
[2] ADAPT Research Centre, School of Computer Science, University College Dublin, Ireland
[3] ADAPT Research Centre, School of Computer Science and Statistics, Trinity College Dublin, Ireland

patrick.english@ucdconnect.ie, erfan.shams@ucd.ie, john.kelleher@adaptcentre.ie, julie.berndsen@ucd.ie

The transformer network architecture [1], originally proposed for language models, has been adapted for automatic speech recognition (ASR) and achieved state-of-the-art performance. However, interpreting the fine-tuned parameters of these models is challenging [2]. Post-hoc explainable AI (XAI) methods[3], such as domain-informed probing tasks, are used to evaluate the latent representations and reveal whether human-interpretable information is encoded within the model.

This paper analyses the latent representations of a transformer-based wav2vec 2.0 base model, fine-tuned on the Librispeech corpus, using a domain-informed probing task based on phonetic articulatory features from the International Phonetic Alphabet (IPA). The probing task evaluates whether the model captures the finer details of articulatory feature transitions, such as place of articulation (POA) and manner of articulation (MOA). A visualisation tool was proposed to allow frame-by-frame inspection of the probed model, illustrating the probabilities of various articulatory features across an utterance and contributing to the explainability of these black box models in relation to known phonetic and phonological phenomena.[1] The TIMIT corpus, with timestamped phone-level annotations, provided supervised data for the probing task. Phonetic features were derived from individual phone labels based on IPA classifications for vowels and consonants at different levels, including category, vowel features (height, front, rounding), and consonant features (MOA, POA, voicing). The original phone labels were also used for a separate phone identification probe. TIMIT utterances were fed into wav2vec 2.0, and latent representations from 12 transformer layers and the initial convolution layer were extracted. The representations are 13×N×768 tensors, where N is the number of 25 ms encoded frames, with a training set of 175,232 frames per layer.

Multi-layer perceptron probes with 768 input units, a hidden layer of 200 ReLU units, and softmax output units were trained on the aggregated representations. Probes trained on randomised datasets failed to predict features effectively, supporting that probe performance was not influenced by chance correlation.

A visualisation tool, w2v2viz, was developed to facilitate analysis and manual inspection of probe outputs. For the visualisation task, probes were provided with time-step embeddings, unaveraged and without the need for labels. Probe outputs for each time-step representation across output layers are visualised as a 3D terrain, with X and Y axes representing MOA and POA categories, and the Z axis corresponding to the product of probability distributions for each feature intersection in consonant frames. The transition frames between the same two phones in different utterances were examined, focusing on the N-F transition. As shown in Figure 1, example A demonstrated bilabial assimilation, with the probes detecting articulatory features of a P phone, exemplifying epenthesis. In contrast, example B showed a direct transition without bilabial assimilation, with the probes identifying the persistence of the POA feature as the speaker transitioned to the production of the fricative F.

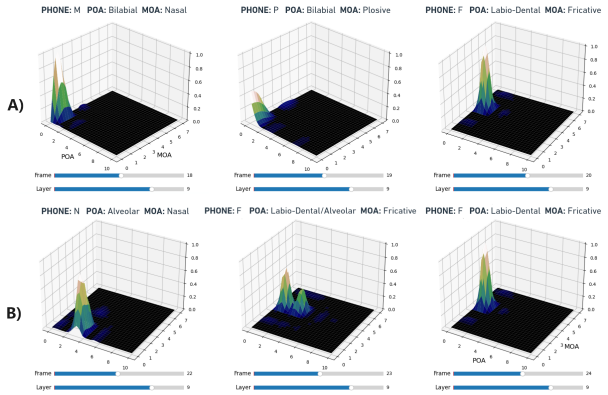


Figure 1: *Visualisation of N-F transition with: A) Bilabial assimilation and epenthesis of P. B) No* ***bilabial*** *occurrence.*

Our probing methodology successfully identifies nuanced articulatory phenomena at the sub-phonemic level, as well as articulatory transitions not recorded in the TIMIT dataset, particularly constraints on POA transition. The visualisation tool facilitates the investigation of single frames across model layers, tracking the emergence of features. We plan to expand this work to assess a wider range of information, allowing further evaluation of phonetic theory and explainability within model embeddings.

1. References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017.

[2] W. Saeed and C. Omlin, "Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities," *Knowledge-Based Systems*, vol. 263, p. 110273, 2023.

[3] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[1]This abstract is a shortened version of our published paper "Following the Embedding: Identifying Transition Phenomena in wav2vec 2.0 Representations of Speech Audio", in ICASSP 2024.

3rd COG-MHEAR Audio-Visual Speech Enhancement Challenge (AVSEC-3)

Andrea Lorena Aldana Blanco, Ondrej Klejch, Cassia Valentini-Botinhao, Peter Bell

The Centre for Speech Technology Research, University of Edinburgh, United Kingdom
lorena.aldana@ed.ac.uk

The Audio-Visual Speech Enhancement Challenge (AVSEC) aims to improve the intelligibility and quality of speech signals using visual information of the target speaker, hence improving performance of audio-only speech enhancement technologies. Building upon the two previous successful editions of the challenge: AVSEC-1 (SLT 2022) and AVSEC-2 (ASRU 2023), we are running the third edition (AVSEC-3), which will take place as a satellite event of Interspeech 2024.

To overcome the scarcity of audio-visual datasets targeted at speech enhancement, we built the challenge dataset from in-the-wild videos (i.e. TED Talks). Speech from the target speaker is mixed with one interferer that can be either speech or noise. Signal-to-noise ratios in the train/dev sets range from -15 dB to $+5$ dB for competing speakers and from -10 dB to $+10$ dB for noise interferers.

Furthermore, we designed a human intelligibility evaluation protocol for in-the-wild stimuli that does not require a specific sentence structure and is suitable for crowd-sourcing platforms.

In this edition of the challenge we aim to gain deeper understanding about the system's performance across a wider range of conditions. To do that, we increased the number of noise categories and increased the granularity of signal-to-noise ratios used to evaluate the systems.

Figure 1 shows results from AVSEC-2 and from the best performing system in AVSEC-1.

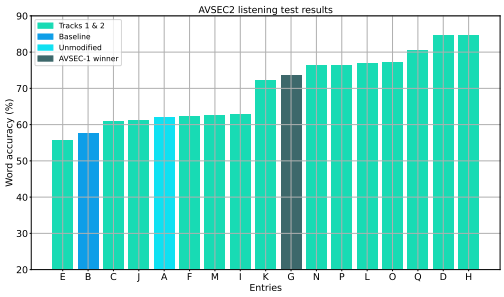


Figure 1: *Overall word accuracy scores (%) AVSEC-2 and best performing system of AVSEC-1. Original non-enhanced (A) and baseline (B) are in blue.*

We expect results of AVSEC-3 to show a general trend of system improvement as in the previous two editions, but also to provide further insight about audio-visual listening in the context of speech enhancement systems.

Research plan: Neural encoding of AI-generated speech prosody by L1 and L2 speakers

Linda Bakkouche[1], *Stephanie Cooper*[1], *Xinbing Luo*[1], *Madeleine Rees*[1],

Emily Lau[1], *Charles McGhee*[3], *Kai Alter*[4], *Brechtje Post*[1], *Julia Schwarz*[2]

lb983@cam.ac.uk; js2275@cam.ac.uk

Background

Artificial intelligence (AI)-powered voice clones are increasingly utilized in educational and clinical settings, offering significant benefits for second language learning e.g., IELTS [6]; Duolingo [5], audiobooks (e.g., Elevenlabs), and assistive technology (e.g.,ESTA,Socially Assistive Robots;[2]). Despite their advantages, AI-generated voices often fall short in replicating human-like prosody [4], such as fundamental frequency (F0) and rhythm [1], elements that are crucial for conveying emotions and ensuring speech intelligibility, especially in noisy environments (Dupuis et al., 2016; De Looze & Hirst, 2014). The effects of such prosodic differences on speech perception, however, are unclear.

Objective

The objective of this research is to explore how prosodic discrepancies in AI-generated speech affect neural responses and behavioral performance among listeners, with a particular focus on comparing monolingual and bilingual populations. This investigation is essential for uncovering the potential challenges that second language (L2) learners face when interacting with AI speech technologies in educational and assistive settings, aiming to enhance the effectiveness and accessibility of these sophisticated tools.

Methodology

The study addresses three main research questions:

RQ1: To what extent does AI speech impact perception accuracy and behavioural performance in monolingual and bilingual groups?

RQ2: Does AI-generated speech trigger neural compensation, as indicated by amplitude deviations of Event-Related Potentials (ERPs)?

RQ3: Does the variation in intonation (F0 and speech rhythm) between AI-generated and human speech affect speech entrainment?

The study integrates behavioral measures with electroencephalography (EEG) to analyze listeners' responses to narrative passages produced by AI and human voices. Prior to initiating the main experiment, we will select optimal Voice Cloning (VC) tools and an Automatic Speech Recognition (ASR) system to ensure the utilization of high-quality, natural-sounding AI voices. To evaluate the naturalness and similarity of these AI voices to human voices, thirty native English speakers will undergo MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) tests. Following this preparatory phase, the main experimental component will involve 90 participants categorized into three groups based on their language proficiency (monolingual native speaker, advanced bilingual, intermediate bilingual; *Figure 1*). Participants will listen to stories in both quiet and noisy conditions (multi-talker babble) to mimic real-life learning environments. Cognitive processing during these listening tasks will be measured through EEG, focusing on ERPs, oscillations, and behavioral tasks assessing recognition and attention.

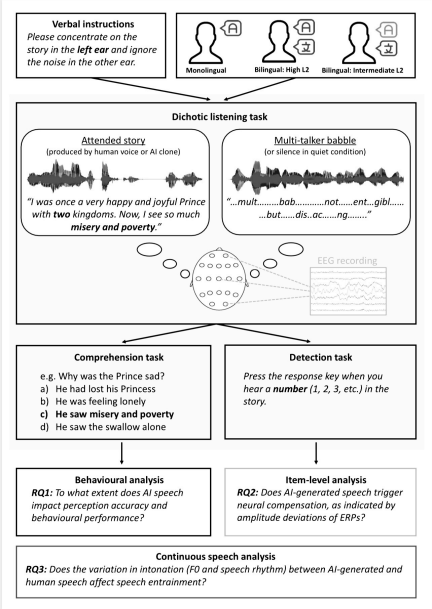


Figure 1: *Planned experiment design*

Expected Outcomes and Implications

We anticipate that listeners will show varying levels of adaptability to AI-generated speech, with potential difficulties for intermediate proficiency bilinguals. Neural adaptation might be evident, particularly in ERP components associated with auditory attention and working memory. Differences in prosody are also expected to affect speech entrainment, potentially impacting the synchronization of neural activity with prosodic characteristics of speech. Results from this study will guide further development of AI-powered educational tools and inform strategies to optimize AI speech for better learning outcomes in L2 education. Additionally, this research will contribute to the broader understanding of human speech processing in the context of emerging AI technologies.

References

[1] Cohn, M., Segedin, F., & Zellou, G. (2022). Acoustic-phonetic properties of Siri- and human-directed speech. *Journal of Phonetics, 90*, 101123. https://doi.org/10.1016/j.wocn.2021.101123
[2] Georgieva-Tsaneva, Andreeva et al. (2023). Exploring the Potential of Social Robots for Speech and Language Therapy: A Review and Analysis of Interactive Scenarios. *Machines (Basel), 11*(7), 693. https://doi.org/10.3390/machines11070693 (https://doi.org/10.3390/machines11070693)
[3] Hirst, D. J., & de Looze, C. (2021). Measuring Speech. Fundamental frequency and pitch. In R. A. Knight & J. Setter (Eds.), *Cambridge Handbook of Phonetics* (1st ed., pp. 336-361). Cambridge University Press. https://doi.org/10.1017/9781108644198
[4] Kim J., Xu K., Merrill K. (2022). Man vs. machine: Human responses to an AI newscaster and the role of social presence. *Soc. Sci. J.* 1–13. 10.1080/03623319.2022.2027163
[5] Rachels, J. R., & Rockinson-Szapkiw, A. (2017). The effects of a mobile gamification app on elementary students' Spanish achievement and self-efficacy. *Computer Assisted Language Learning, 31*(9), 1-18. https://doi.org/10.1080/09588221.2017.1382536 (https://doi.org/10.1080/09588221.2017.1382536)
[6] Zou, B., Reinders, H., Thomas, M., & Barr, D. (2023). Editorial: Using artificial intelligence technology for language learning. *Frontiers in Psychology, 14*, 1287667-1287667.

[1]Phonetics Laboratory, University of Cambridge, [2]Department of Psychology, University of Cambridge, [3]Department of Engineering, University of Cambridge, [4]Biosciencies Institute, Newcastle University

Exploring individual speaker characteristics within a forensic automatic speaker recognition system

Chenzi Xu[1], *Vincent Hughes*[2], *Paul Foulkes*[2], *Philip Harrison*[2], *Poppy Welch*[2], *Jessica Wormald*[2], *Finnian Kelly*[3] *and David van der Vloed*[4]

[1]University of Oxford, United Kingdom [2]University of York, United Kingdom
[3]Oxford Wave Research, United Kingdom [4]Netherlands Forensic Institute, Netherlands

chenzi.xu@ling-phil.ox.ac.uk, {firstname.lastname}@york.ac.uk,
finnian@oxfordwaveresearch.com, d.van.der.vloed@nfi.nl

Automatic speaker recognition (ASR) systems are increasingly used in forensic voice comparison (FVC) [1] casework around the world. A key issue for automatic speaker recognition (ASR), particularly for forensics, is our lack of understanding about why certain voices prove more or less of a challenge for systems. In this paper, we focus on variability in individual speaker performance within an x-vector ASR system and examine this variability as a function of the phonetic content within speech samples.

We first conducted an initial validation [2] to assess the overall performance of an x-vector ASR system using GBR-ENG, a dataset of forensically-realistic recordings provided by the UK Government. Subsets of 48 male (160 files total) and 46 female (154 files total) speakers were used as test sets, given that there were at least three non-contemporaneous samples for each speaker. All selected samples were good quality (determined based on over 30s of net speech, WADA SNR of over 24dB, and less than 1% clipping), mobile phone recordings, to reduce confounding variables across speakers. The calibration sets contained 50 male and 50 female speakers selected at random, with two good quality mobile phone samples recorded on separate days. Analysis was conducted using sets of male and female speakers separately.

Then we assess the ASR performance in two ways: (1) the extent of variability in individual performance, both in terms of discrimination and calibration error; (2) the variability across speakers as a function of the phonetic content within the speech samples, using the method described in [3]. For speech samples, orthographic transcriptions were created using the large-V2 model in Whisper-timestamped, and time-aligned phone boundaries were obtained via the Montreal Forced Aligner. Phones were then categorised into five phoneme classes: (a) vowels, (b) nasals, (c) approximants, (d) fricatives, and (e) plosives. We performed experiments where the speech material corresponding to a given class was removed from the two speech recordings of each trials. The impact of a specific phoneme class was evaluated using the metric C_{llr}^{R} defined in [3], a means of comparing effects across segment categories whilst not skewing results due to the amount of phonetic material excluded. Acoustic-phonetic data were also extracted which capture long-term, laryngeal properties of the voices, providing an additional source of phonetic information to examine individual speaker performance.

The inclusion of vowels generally improved performance, but not for all speakers (see Figure 1). Indeed, some speakers produced broadly the same C_{llr} irrespective of the phonetic content in the speech samples. Poor ASR performance was not well correlated with long-term laryngeal features (f0 and laryngeal voice quality) and these features may provide additional speaker discriminatory information for some speakers. We discuss the implications of these findings in terms of developing a speaker quality metric for flagging potentially problematic speakers prior to ASR comparison.

1. References

[1] D. van der Vloed and T. Cambier-Langevel, "How we use automatic speaker comparison in forensic practice." *International Journal of Speech, Language & the Law*, vol. 29, no. 2, 2022.

[2] G. S. Morrison, E. Enzinger, V. Hughes, M. Jessen, D. Meuwly, C. Neumann, S. Planting, W. C. Thompson, D. van der Vloed, R. J. Ypma *et al.*, "Consensus on validation of forensic voice comparison," *Science & Justice*, vol. 61, no. 3, pp. 299–309, 2021.

[3] A. Moez, B. Jean-François, B. K. Waad, R. Solange, and K. Juliette, "Phonetic content impact on forensic voice comparison," in *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 210–217.

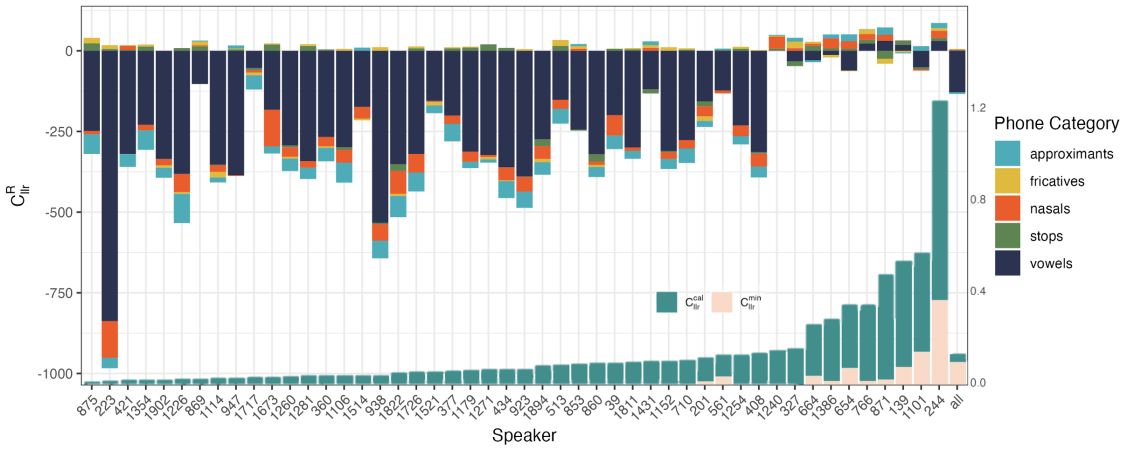


Figure 1: *Bottom: By-speaker and overall C_{llr}^{cal} (green) and C_{llr}^{min} (pink) values (C_{llr} is the overall height of the bar) for males. Upper: By-speaker (speaker numbers on the x-axis) and overall C_{llr}^{R} values (negative C_{llr}^{R} indicates that a given category contributes to improving C_{llr} while positive C_{llr}^{R} values indicate that performance gets worse with the inclusion of a given category) for each of the five phone categories for the male sets (*note differences in the scales on the y-axes)*

An Audio-Based Depression Tracking Model Using Machine Learning

Andrea Vitullo[1], *Dr Eugenio Donati*[1]

[1]University of West London

21506988@student.uwl.ac.uk | andrea.vitullo27@gmail.com,
eugenio.donati@uwl.ac.uk

This study provides a detailed account of the process and results of an extensive investigation into detecting depression through speech using neural network models. It presents an end-to-end methodology that includes data collection from the DAIC-WOZ dataset [1][2][3], audio pre-processing, feature extraction, and the application of advanced machine-learning techniques. Key features such as Mel-Frequency Cepstral Coefficients (MFCCs), Chromagrams, Mel Spectrograms, and Spectrograms were extracted and analysed for their effectiveness in pattern recognition indicative of depressive states. Neural network architectures, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), were leveraged to understand speech data's spatial and temporal dynamics.

The methodology focused on audio pre-processing, tackling issues such as background noise, non-contributing sections, and gender representation. Audio files were converted to a uniform format and subjected to noise removal and silence deletion to ensure high-quality data. Participants' speech was isolated from interview segments. Key features were extracted and analysed for their effectiveness in capturing speech patterns indicative of depressive states.

The neural network models were constructed and trained using balanced data to mitigate potential biases. Specifically, the architecture of the Chromagram-based model, which achieved a test accuracy of 0.63, included layers for convolution and pooling to extract and abstract features, followed by GRU recurrent layers to capture temporal dynamics. The optimisation process involved fine-tuning parameters such as learning rate and batch size, utilising frameworks like Optuna for efficient hyperparameter tuning.

Results revealed that the Chromagram-based model slightly outperformed others, suggesting that tonal qualities of speech are potentially more indicative of depression. MFCC-based models were less effective in capturing emotional nuances, highlighting the importance of selecting appropriate features for depression detection. Comparative analysis with existing literature demonstrated the model's robustness and alignment with current research, while also offering improvements in terms of feature extraction.

This research's findings align with and extend current literature on Automatic Depression Detection (ADD). They provide valuable insights into the challenges and considerations of deploying machine learning in clinical settings. The balanced approach in class representation and participant consistency underscores the potential of these models to be used in real-world scenarios, offering a promising outlook for more effective mental health care solutions. Additionally, the project emphasises the critical need for maintaining well-balanced and ethically sourced datasets, ensuring the development of socially conscious models that guarantee fair and successful solutions for mental health care. This study contributes to research in the application of neural network models in diagnosing mental health disorders, advocating for continuous improvement and refinement of these technologies. Future research directions could include studying more advanced neural network architectures, such as Transformers, to enhance the model's accuracy and robustness, and exploring their integration into broader clinical workflows, ensuring they are robust, fair, and effective in diverse real-world settings.

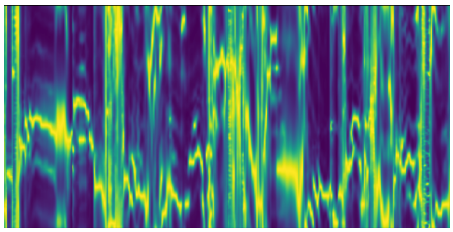


Figure 1: *Chromagram Example*.

1. References

[1] Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, Louis-Philippe Morency, "*The Distress Analysis Interview Corpus of human and computer interviews*", Proceedings of Language Resources and Evaluation Conference (LREC), 2014

[2] DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., Lucas, G., Marsella, S., Morbini, F., Nazarian, A., Scherer, S., Stratou, G., Suri, A., Traum, D., Wood, R., Xu, Y., Rizzo, A., and Morency, L.-P. (2014). "*SimSensei kiosk: A virtual human interviewer for healthcare decision support*". In Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'14), Paris

[3] Degottex, G.; Kane, J.; Drugman, T.; Raitio, T.; and Scherer, S., *COVAREP - A collaborative voice analysis repository for speech technologies*. In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014), pages 960-964, 2014.

[1] The UKIS 2024 Committee would like to thank ISCA and Interspeech 2023 organisers for kindly providing this template.

Characterizing Code-switching: Applying Linguistic Principles for Metric Assessment and Development

Jie Chi, Electra Wallington*, Peter Bell*

The Centre for Speech Technology Research, University of Edinburgh, United Kingdom
{jie.chi, electra.wallington, peter.bell}@ed.ac.uk

Handling code-switching (CS), the alternation between two or more languages – or language varieties – within a single conversation, is a growing focus in speech technology. This trend is driven by the expansion of multilingual automatic speech recognition (ASR) methodologies, which have in turn enabled ASR systems to be deployed in diverse contexts and domains where CS is likely to occur. However, to develop such effective systems in real-world scenarios, it is crucial to recognize that not all code switching behaviour is equal. Consider the examples below: all come from CS datasets [1], but contrast the singular English segment in an otherwise monolingual Gaelic sequence in (1), with the far more frequent switches between Xhosa and English in (3). Additionally, observe that the usage of English in (1) and (2) is constrained to technical concepts and named entities, in contrast to to the far freer employment of English syntactically and semantically in (3).

(1) bhiodh na gaidheil dìreach math fhèin air ready steady cook s tha mi a smaointinn gu bheil e mar phàirt den chultar againne cuideachd bidh thu [2]
Gaelic people would be good at ready steady cook **and I think it's part of our culture too.**

(2) और विभिन्न crystal systems के crystal structures को दिखाना उदाहरण के लिए cubic hexagonal
'**and showing the** crystal structures **of different** crystal systems **for example** cubic, hexagonal'

(3) i worry yam kengoku it's the next meeting on friday and kengoku I'm going out chomi nomfana wam
my worry **now** its the next meeting on Friday and **now** Im going out **friend with my man'**

A system built or trained on CS data primarily resembling (1) would likely not be sufficient to handle input in the style of (3). We believe that treating CS as a monolithic phenomenon thus poses a significant risk of domain mismatch. Given such CS style variation, it may not be possible to guarantee that a system or methodology developed on one CS dataset would generalise to other languages or contexts.

Motivated by this conundrum, there have been efforts to develop metrics to measure the richness, level, or complexity of CS in datasets. Popular metrics include M-index[1], I-Index[2], and Code-Mixing Index[3], which rely on counts of tokens or switching points, as well as the Burstiness and Memory metrics[4], which rely on characterizing the distribution of participating languages.

That it is becoming habitual to quote at least one such metric when introducing novel CS data into the research space is, theoretically, facilitating more robust CS ASR development. However, we believe caution is warranted: to rely on these metrics to gauge the complexity of a dataset is to assume that they are observing features that accurately and *sufficiently* characterize the space of CS variation. Testing this requires a clear definition of what constitutes a meaningful division of CS variation, but there is a circularity issue: we cannot measure a metric's success at identifying CS styles across datasets when the same metric is used to annotate CS occurrences.

We argue that what is needed are methods to analyze and assess the sufficiency of such metrics, whether in isolation or in combination, that are grounded in external factors. Linguistic research provides valuable insights in this regard. Extensive linguistic literature suggests that surface variability in CS is not random but influenced by latent variables related to meta-properties of the speech situation and setting.

Our contributions are twofold: we introduce a method of analysing CS richness metrics which leverages linguistic findings as makeshift ground-truths. We propose to assess metrics by considering the extent to which their outputs align with linguistic research, such as the observation that CS tends to be more complex in conversational settings than in formal or technical settings. Second, we introduce the "T-index", which uses a pre-trained machine translation system to better model the relationship between languages at each switching point.

*Equal Contribution

[1] When we use the term *CS datasets*, we refer to the transcripts of CS speech datasets, rather than written text.

[2] https://learngaelic.scot

1. References

[1] R. Barnett, E. Codó, E. Eppler, M. Forcadell, P. Gardner-Chloros, R. Van Hout, M. Moyer, M. C. Torras, M. T. Turell, M. Sebba *et al.*, "The lides coding manual: A document for preparing and analyzing language interaction data version 1.1–july 1999." *International Journal of Bilingualism*, vol. 4, no. 2, pp. 131–271, 2000.

[2] G. A. Guzman, J. Serigos, B. Bullock, and A. J. Toribio, "Simple tools for exploring variation in code-switching for linguists," in *Proceedings of the second workshop on computational approaches to code switching*, 2016, pp. 12–20.

[3] V. Srivastava and M. Singh, "Challenges and limitations with the metrics measuring the complexity of code-mixed text," *CoRR*, vol. abs/2106.10123, 2021. [Online]. Available: https://arxiv.org/abs/2106.10123

[4] G. Guzmán, J. Ricard, J. Serigos, B. E. Bullock, and A. J. Toribio, "Metrics for Modeling Code-Switching Across Corpora," in *Proc. Interspeech 2017*, 2017, pp. 67–71.

LEVERAGING LANGUAGE AFFINITIES IN WAV2VEC FINETUNING FOR LOW RESOURCED LANGUAGES

Jeffrey Michael, Oscar Saz

Emotech Ltd, London, UK

jeffrey@emotech.co, oscar@emotech.co

Abstract

This paper proposes enhancing Automatic Speech Recognition for low–resource languages using automatic language similarity metrics and wav2vec models. Urdu and Catalan are target languages, with English, Hindi, and French as pretraining languages. The hypothesis suggests that pretraining with similar languages improves ASR. Experiments confirm a correlation between language similarity and ASR performance, showing that closely related languages outperform multilingual models. The methodology proposed measures similarities based on lexical, phonological, and geographical factors; and the experimental setup fine–tunes models on limited data, ensuring balance. Findings stress base language selection's significance, advocating for context-aware decisions considering linguistic affinity and data diversity. The research offers insights for accurate ASR systems in underrepresented languages, stressing a thoughtful approach to pretraining language selection for various language pairs.

1. Introduction

Advancements in Automatic Speech Recognition systems have started leveraging unlabelled data by pretraining stages. However, selecting an optimal pretraining strategy, for low-resource languages, remains a challenge. This paper proposes a novel approach to enhance ASR performance by leveraging language similarity metrics to inform pretraining strategy selection.

2. Proposed Idea

In this study, we explore how selecting appropriate languages for pretraining can enhance speech recognition systems, especially for languages with limited resources. Our hypothesis suggests that using languages similar to the target language can significantly boost system performance, as shared linguistic traits facilitate smoother knowledge transfer during training, enhancing accuracy and efficiency. We utilize tools measuring linguistic similarities to investigate this hypothesis. Our research aims to improve speech recognition for underrepresented languages, addressing challenges faced by speakers of such languages and promoting system accessibility and inclusivity. By examining the relationship between language similarity and system performance, we seek to contribute to developing more accurate and inclusive speech recognition technologies. Despite the complexity of this task, we approach it with enthusiasm and commitment to uncovering insights that benefit individuals worldwide, regardless of linguistic backgrounds.

3. Experimental Setup

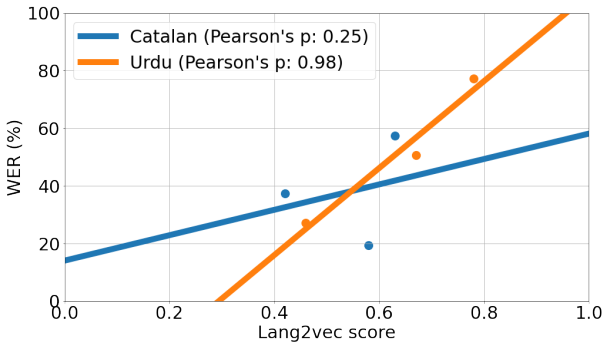
The experiments fine-tune pretraining models using limited data from the CommonVoice dataset for Urdu and Catalan. Three monolingual pre-trained models (English, Hindi, and French) are chosen, alongside the multilingual XLSR model for comparison. The lang2vec tool quantifies language relationships based on syntactical, phonological, and geographical similarities. Limiting the dataset to 50 hours ensures a fair comparison, simulating scenarios with constrained data availability typical in low-resource language settings. While Catalan has ample data, Urdu is notably under-resourced, with only 60 hours available.

4. Results

The experimental findings demonstrate a positive correlation between language similarity and ASR performance, with lower Word Error Rates (WER) observed when pretraining with closely related languages. For instance, pretraining with Hindi notably reduces WER for Urdu, while French yields improved results for Catalan. Monolingual pretraining models consistently outperform the large multilingual XLSR model. Notably, using the wav2vec English model results in a 37.4% error rate for Catalan, dropping to 19.31% with the wav2vec French model. Similarly, pretraining with the wav2vec Hindi model significantly reduces the error rate to 27.1% for Urdu, underscoring the significance of selecting a closely aligned base language, as shown in the accompanying table and figure.

	Catalan		Urdu	
	Distance	WER	Distance	WER
Hindi	0.63	57.3%	0.46	27.1%
English	0.42	37.4%	0.67	50.7%
French	0.58	19.3%	0.78	33.2%
XLSR	-	8.3%	-	33.2%

Table 1: *Distances and WERs between language pairs*



Correlation between lang2vec scores ($x-axis$) vs. WER results ($y-axis$) for Catalan and Urdu languages

Bootstrapping Spoken Information Retrieval for Unwritten Languages

Ondřej Klejch[1], *Electra Wallington*[1], *Thomas Reitmaier*[2], *Emily Nielsen*[2], *Dani Kalarikalayil Raju*[3], *Nina Markl*[4], *Gavin Bailey*[2], *Jennifer Pearson*[2], *Matt Jones*[2], *Simon Robinson*[2], *Peter Bell*[1]

[1]The Centre for Speech Technology Research, University of Edinburgh, UK
[2]Swansea University, UK [3]Studio Hasi, India [4]University of Essex, UK

o.klejch@ed.ac.uk

Speech technologies for 100 – 200 of the most well-resourced languages have achieved great accuracy in recent years. This progress was possible thanks to the hundreds of thousands of hours of training data available for these languages. However, these amounts of training data are not available for the majority of 7000 spoken languages. In the EPSRC UnMute project, one of our goals is to build technologies that would enable speakers of minority languages to train and deploy their own spoken interaction systems with no manually transcribed data and no AI specialist knowledge.

Here, we describe our approach [1] for bootstrapping a spoken information retrieval system that allows speakers of unwritten languages to upload media items, such as images and videos, together with their spoken descriptions. These media items can be later retrieved with spoken queries. Since we want our voice-search system to work from day zero, we cannot rely on any training data in the target language. Therefore, we split the system into two components: *a multilingual phone recogniser* trained on data from well-resourced languages and *a ranker* trained on spoken descriptions.

The multilingual phone recogniser is a hybrid model trained on data from well-resourced languages as described in [2]. In particular, we used 20 hours of English data from LibriSpeech [3] and 20 hours of German, French, Spanish, Polish and Russian from GlobalPhone [4]. We replaced the standard MFCC features with features extracted from self-supervised models to improve the robustness of the multilingual phone recogniser. We evaluated several self-supervised models for cross-lingual phone recognition on Bulgarian, Czech, Hausa, Portuguese, Swahili, Swedish and Ukrainian. We found that features extracted from the 18th layer of XLS-R 300M [5] achieve the best performance in cross-lingual phone recognition as can be seen in Figure 1. During decoding we used a bi-gram phone language model trained on the multilingual phone recogniser's phonetic transcripts training dataset.

The ranker is then trained to predict which media item belongs to the phonetic transcript of the query. It is implemented as a multi-class support vector machine [6] classifier using term-frequency inverse document frequency [7] weighted uni-grams, bi-grams and tri-grams extracted from the automatic phonetic transcripts of the queries. Since the ranker is a lightweight model, it can be retrained as more items and spoken descriptions are added to the system. It can also leverage spoken queries and subsequent user interactions/feedback to improve its performance.

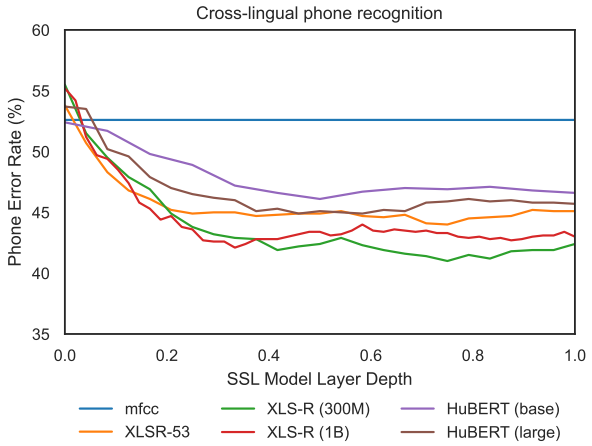


Figure 1: *Average phone error rate of cross-lingual phone recognition with a multilingual phone recogniser trained on English, German, French, Spanish, Polish and Russian when transcribing data from Bulgarian, Czech, Hausa, Portuguese, Swahili, Swedish and Ukrainian.*

1. References

[1] T. Reitmaier, D. K. Raju, O. Klejch, E. Wallington, N. Markl, J. Pearson, M. Jones, P. Bell, and S. Robinson, "Cultivating Spoken Language Technologies for Unwritten Languages," in *CHI*, 2024.

[2] O. Klejch, E. Wallington, and P. Bell, "Deciphering Speech: a Zero-Resource Approach to Cross-Lingual Transfer in ASR," in *Interspeech*, 2022.

[3] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *ICASSP*, 2015.

[4] T. Schultz, N. T. Vu, and T. Schlippe, "Globalphone: A multilingual text & speech database in 20 languages," in *ICASSP*, 2013.

[5] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale," in *Interspeech*, 2022.

[6] C. Cortes and V. Vapnik, "Support vector machine," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[7] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to information retrieval*. Cambridge University Press Cambridge, 2008, vol. 39.

Investigation of PEVD-based speech enhancement in wireless acoustic sensor networks

Emilie d'Olne, Patrick A. Naylor

Department of Electrical and Electronic Engineering, Imperial College London, UK
{emilie.dolne16, p.naylor}@imperial.ac.uk

Speech enhancement in wireless acoustic sensor networks (WASNs) has been a growing topic of interest in the literature. The ever-increasing number of devices equipped with microphones has made it possible to exploit microphones spread randomly in a room rather than from a single location as per traditional compact arrays. While this improved flexibility in microphone positioning means some nodes are likely to have a higher signal-to-noise ratio (SNR), network processing also introduces challenges: different devices operate with different clocks and may experience clock offset and skew; they also might have different frequency-dependent gains; and the wireless link between microphones in the network and the fusion centre introduces latency. These issues are particularly relevant for traditional enhancement techniques such as beamforming which typically rely on having calibrated arrays with known geometries. A popular solution is to apply a calibration step to remove non-idealities and perform geometry estimation, and then treat the network as single array and use standard enhancement techniques [1]. However, the calibration task is non-trivial and often ignores the effect of latency in the network.

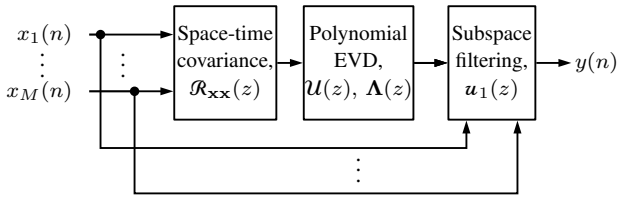


Figure 1: *PEVD-based enhancement in networks.*

Instead, this work explores the use of polynomial eigenvalue decomposition (PEVD)-based speech enhancement as a blind latency agnostic method for enhancement in WASNs. In particular, the effect of latency on the enhancement performance is studied. PEVD has been used previously with compact arrays of various geometries and enhancement is obtained through strong decorrelation of microphone signals in space, time, and frequency [2]. The method is blind as it does not require knowledge of the network geometry and does not use prior information about the scene, e.g. source location or noise characteristics. Furthermore, PEVD-based enhancement has the potential to perform simultaneous enhancement and dereverberation without introducing distortions in the target signal. A block diagram of the approach is given in Fig. 1, and readers are invited to consult [2] for an in-depth description of the method.

Simulations were created for a $5 \times 4 \times 3$ m shoe-box room with reverberation time $T_{60} = 0.4$ s, with 5 microphones placed at random locations capturing a single source also placed randomly and corrupted by diffuse babble noise. Latency was simulated by delaying microphones signals by a random time in the range $[0, 50]$ ms. Input SNR was calculated as the average SNR across all microphones. Preliminary results in Fig. 2 show the performance of PEVD compared to an oracle multichannel Wiener filter (MWF) (*omwf*, upper bound on performance), a practical implementation of the MWF using a voice activity detector (*vadmwf*), and KLT-based enhancement (*ogevd*). While enhancement of *vadmwf* and *ogevd* is consistently worse than the best microphone in the network, PEVD combines all microphone signals to obtain an enhanced signal with SNR comparable to the best microphone in the network, and improves predicted intelligibility (STOI) and signal quality (PESQ) across all input SNRs. Informal listening suggests that the fact that PEVD increases predicted intelligibility without providing noise reduction comes from dereverberation and whitening of the residual noise.

1. References

[1] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *18th IEEE Symp. on Commun. and Veh. Technol. in the Benelux (SCVT)*, 2011, pp. 1–6.

[2] V. W. Neo, S. Redif, J. G. McWhirter, J. Pestana, I. K. Proudler, S. Weiss, and P. A. Naylor, "Polynomial eigenvalue decomposition for multichannel broadband signal processing: A mathematical technique offering new insights and solutions," *IEEE Signal Process. Mag.*, vol. 40, no. 7, pp. 18–37, 2023.

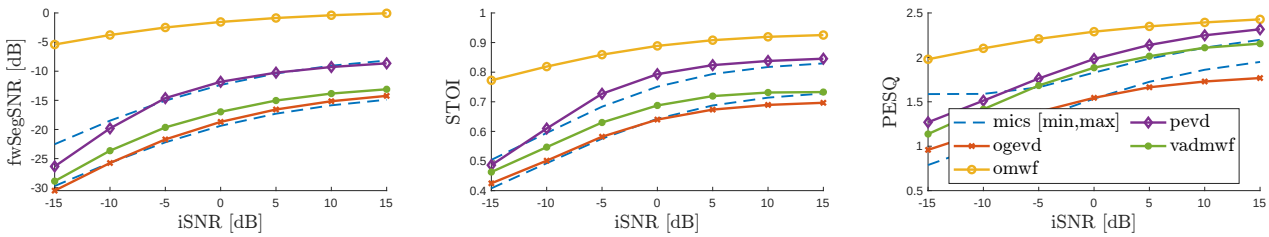


Figure 2: *Preliminary results for enhancement of a single source in diffuse babble noise, using 5 microphones experiencing random latency in the range* $[0, 50]$ *ms in a simulated* $5 \times 4 \times 3$ *m shoe-box room with* $T_{60} = 0.4$ *s. Results averaged over 500 sentences.*

Topical Emphasis synthesis with Voice Puppetry

Emelie Van De Vreken[1], *Korin Richmond*[1], *Catherine Lai*[1]

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

emelie.vandevreken@ed.ac.uk, korin.richmond@ed.ac.uk, c.lai@ed.ac.uk

With increased use of synthesised speech in a larger number of contexts, there is an increased need to create human-like voices that go beyond neutral, read speech. Here the concern is not with further improving comprehension of the lexical content, but with improving its expressiveness through prosody, thus making it more appropriate for use as part of conversational agents or to create audiobooks. Expressive speech refers to any speech which conveys information such as emotion, or speaker attitudes. For example, information can be emphasised by increased pitch movement [1].

We introduce voice puppetry with FastPitch [2] as a method of generating expressive speech. The aim is to keep the identity of the training speaker of a Text-to-Speech (TTS) model intact and only change *how* the utterance is said, based on a reference speech recording (see [3]). This can be considered as a case of style transfer or many-to-one voice conversion. Relevant prosodic characteristics are extracted from a recording, and the synthesised utterance is created such that it takes on the extracted characteristics without otherwise changing the training speaker identity. In FastPitch, these characteristics are phone durations, pitch, and/or energy. Using puppetry gives control through the use of voice, which can be more intuitive and accessible without the need for expert knowledge or explicit controls.

One possible application of voice puppetry is its usage in Wizard-of-Oz experiments. In this context, "live" puppetry can be combined with Automatic Speech Recognition to give researchers easy and fast control over both the content and style of synthesised speech during experiments [4]. A TTS voice can then be used in experiments as a conversational agent while retaining control and conversational flow.

One well-established technique for expressive synthesised speech is the use of Global Style Tokens (GSTs, [5]). Global Style Tokens are embeddings trained on unlabelled non-neutral data. They can encapsulate a variety of attributes such as the intonation, speaking rate, or emotion, separate from speaker identity and text content. TTS models conditioned on GSTs can produce more variety in prosody, however GSTs are not easily controllable or interpretable. The embeddings represent a latent "style" variable in high-dimensional space, with no guarantee which attribute(s) are encoded. By comparison, voice puppetry is more intuitive and transparent, by not relying on opaque prosody embeddings. It does however rely on recordings of the exact phonetic content with the exact desired prosody.

Voice puppetry has been implemented using FastPitch [2], which has explicitly modelled adaptors for phone durations, pitch and energy. These features are important for producing varying prosody, and thus expressiveness. Voice puppetry uses speech to control the adaptors at inference time, with fine-grained control. As control is on the phone level, the implementation of voice puppetry with FastPitch requires phone alignments of the reference speech to calculate per-phone targets.

Hand-corrected alignments of reference speech are used to establish the best possible puppetry of prosody. Mean Absolute Error (MAE) of phone durations, pitch, and energy between reference speech and puppeteered synthesised speech indicate that as implemented, only durations and pitch are good candidates for control. Evaluation uses the topical emphasis subset of the ADEPT dataset [6].

The automated alignment methods compared are: forced alignments, dynamic time warping (DTW) with MFCC features between TTS output and reference speech, and dynamic time warping with phonetic posteriorgram (PPG) features, extracted from a pretrained phoneme recogniser. Forced alignments with a pretrained aligner result in the lowest MAE of durations and pitch. DTW performs better with PPG features.

A listening test confirms that the topical emphasis from the reference audio is conveyed accurately in the synthesised speech across all alignment methods. Dynamic Time Warping with PPG features produced higher emphasis recognition accuracy, and higher quality MOS scores.

In conclusion, voice puppetry can be used to create expressive TTS utterances with with topical emphasis. In FastPitch, energy cannot be controlled adequately. Despite only controlling pitch and durations and imperfect phone alignment, participants achieve high emphasis recognition accuracy in a listening test.

1. References

[1] A. M. C. Sluijter, *Phonetic correlates of stress and accent*. Holland Academic Graphics, 1995.

[2] A. Łańcucki, "Fastpitch: Parallel text-to-speech with pitch prediction," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6588–6592.

[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] M. P. Aylett and Y. Vazquez-Alvarez, "Voice puppetry: Towards conversational hri woz experiments with synthesised voices," in *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 69–69.

[5] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International conference on machine learning*. PMLR, 2018, pp. 5180–5189.

[6] A. Torresquintero, T. H. Teh, C. G. R. Wallis, M. Staib, D. S. R. Mohan, V. Hu, L. Foglianti, J. Gao, and S. King, "Adept: A dataset for evaluating prosody transfer," 2021. [Online]. Available: https://arxiv.org/abs/2106.08321

SOT Triggered Neural Clustering for Speaker Attributed ASR

Xianrui Zheng[*1], *Guangzhi Sun*[1], *Chao Zhang*[2], *Philip C. Woodland*[1]

[1]Cambridge University Engineering Dept., Trumpington St., Cambridge, CB2 1PZ U.K
[2]Department of Electronic Engineering, Tsinghua University

{xz396, gs534}@cam.ac.uk, cz277@tsinghua.edu.cn, pw117@cam.ac.uk

The task of identifying 'who spoke what' is commonly achieved through cascaded systems. Initially, a voice activity detection (VAD) system is applied to filter out non-speech regions from the entire recording, with the speech regions being the VAD segments between non-speech regions. The segmented output is then processed by a speaker recognition system that determines the content spoken by each speaker. In situations where speech from different speakers overlap, a source separation system can be employed before the recognition system to disentangle overlapping speech into separate streams.

To avoid the transmission of errors from the diarisation stage to the automatic speech recognition (ASR) stage, a framework known as serialised output training (SOT) was introduced in [1]. SOT processes entire VAD segments, rather than taking the speaker homogeneous segments from the diarisation system. It generates transcriptions, starting with the first speaker in the segment, followed by the second speaker, and continues this pattern for any additional speakers.

Here, we propose a novel neural clustering approach, referred to as Segment-Level Discriminative Neural Clustering (SDNC), to assign cluster labels for each VAD segment. This method assigns a single cluster label to segments with a single speaker, while assigning multiple labels to segments with overlapping speech. In addition, SDNC can be integrated with SOT to provide speaker attributed transcriptions (shown in Fig. 1). This approach effectively mitigates errors that often arise in cascaded systems when data is passed from diarisation to ASR.

The input to the SDNC encoder is a sequence of window level speaker embeddings. The decoder takes the encoder outputs, together with external information regarding the number of speakers in each segment to decide the length of the output sequence. For each output label, the transformer decoder cross attention will attend to the encoder output features corresponding to the current segment only, while all encoder features corresponding to other segments will be masked out.

Fig. 2 shows an example, where the eight input speaker embeddings are from three VAD segments. The external information tells the SDNC decoder that the first segment (in blue) has one speaker the second segment (in yellow) has two speakers and the third segment (in grey) has one speaker. The SDNC decoder will then predict four outputs in total. When outputting $\boldsymbol{c}_1$, the cross attention of the SDNC decoder will attend to the first two encoder features, and attend to the next three encoder features when outputting $\boldsymbol{c}_2$ and $\boldsymbol{c}_3$.

Experiments used the augmented multi-party (AMI) dataset, where AMI multiple distant microphone data was the only supervised training data. Table 1 compares a cascaded system using spectral clustering (SC) to a parallel system using SDNC. Windows for extracting speaker embeddings for both setups had a size of 1.5 seconds. Results indicate that the proposed parallel system with SDNC outperforms the cascaded system with SC on both the Dev and Eval sets. Specifically, Table 1 shows a 24% relative DER-H reduction and a 4% relative cpWER reduction on the AMI Eval set.

Table 1: *DER (DER-H) and cpWER results for overlap-aware systems. DER-H only scores VAD segments with one speaker. SDNC only has DER-H since it does not have times for each cluster label for VAD segments that include overlap.*

System	Method	DER (DER-H)		cpWER	
		Dev	Eval	Dev	Eval
Cascaded	SC	8.6 (3.8)	5.6 (2.1)	37.6	35.9
Parallel	SDNC	(1.9)	(1.6)	34.8	34.6

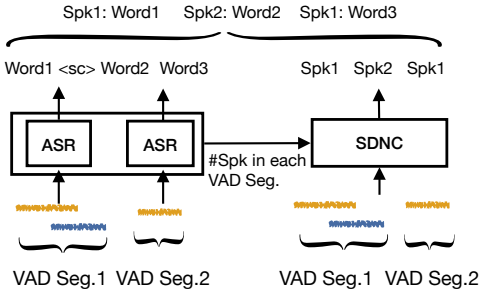


Figure 1: *Parallel system using SDNC and SOT. Colours represent different speakers. '<sc>' represents speaker change.*

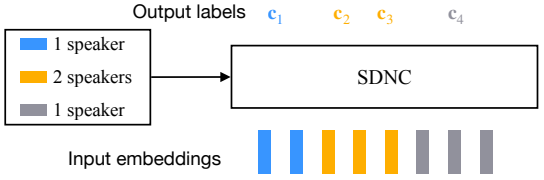


Figure 2: *A toy example of the input and output of the SDNC. Different colours represent different VAD segments and multiple speaker clusters can be output in one VAD segment with SDNC.*

*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

1. References

[1] N. Kanda, Y. Gaur, X. Wang, Z. Meng, and T. Yoshioka, "Serialized output training for end-to-end overlapped speech recognition," in *Proc. Interspeech*, Shanghai, 2020.

Fast and high-quality open-source speech synthesis with Matcha-TTS

Shivam Mehta, Ruibo Tu, Jonas Beskow, Éva Székely, Gustav Eje Henter

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

smehta@kth.se, ghe@kth.se

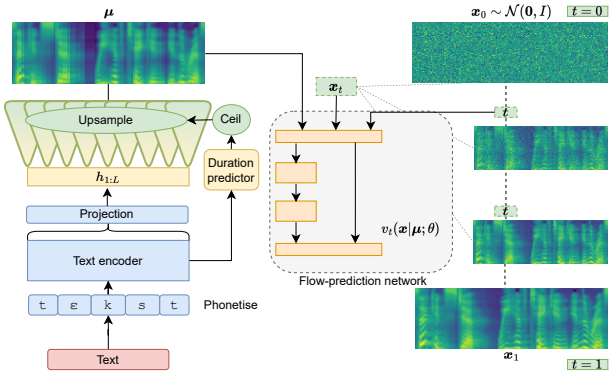


Figure 1: *Overview of synthesis using Matcha-TTS.*

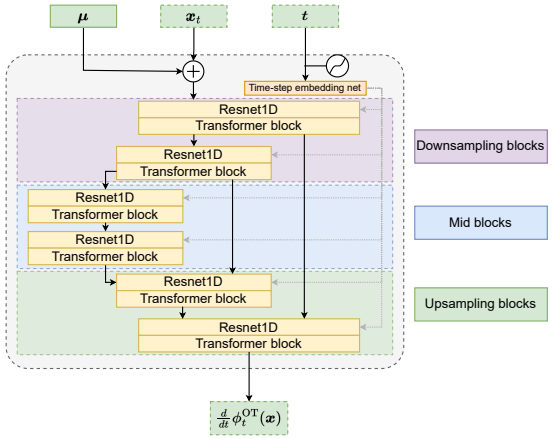


Figure 2: *Detailed architecture of the decoder (labelled "flow-prediction network" in Figure 1). See [2] for more information.*

Diffusion probabilistic models (DPMs) [1] have recently set new benchmarks in the quality of generated synthetic speech by effectively learning to reverse a diffusion process that transforms a simple distribution into complex data distributions typical of speech. Recent developments in TTS have employed these DPMs since they are strong probabilistic models. However, they suffer from slow synthesis speeds, as they rely on iteratively stochastic equations (SDEs or ODEs) with many steps to accurately sample from the learnt model distribution. Our work – Matcha-TTS [2] – addresses these limitations by employing a dual-approach: First we use Optimal-Transport Conditional Flow Matching (OT-CFM) [3] to learn straighter, more direct paths (ODE vector fields) from source to target distributions, enabling us to sample high-quality acoustic features from the learnt model in much fewer ODE-solver steps. Second, inspired by [4], we use a combination of 1D CNNs and Transformers within a U-net decoder ("flow-prediction network") to reduce memory usage and allow for faster training and synthesis.

Our experiments show that Matcha-TTS (label **MAT**) outperforms pre-trained baseline FastSpeech 2 (**FS2**), **VITS**, and, Grad-TTS (**GRAD**) with 2, 4, and 10 NFE (i.e., Euler-forward ODE solver steps) in terms of synthesis speed and intelligibility. It achieves higher mean opinion score (MOS) for naturalness than the baselines in our listening test. The model also demonstrates its efficiency on long utterances, where it rivals the speed of the fastest model (FS2) without sacrificing output quality. Please see our demo page at `https://shivammehta25.github.io/Matcha-TTS/` for synthesised audio, an interactive demo, pre-trained models, and open-source code.

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

Table 1: *Evaluation results (using several different NFE for ODE-based methods) and their number of parameters, minimum GPU RAM needed to train (GiB), real-time factor (including vocoding time) on the test set, ASR WER in percent, and mean opinion score with 95%-confidence interval. The parameter count and RTF for VOC are for the HiFi-GAN vocoder.*

Condition	Params.	RAM	RTF	WER	MOS
VOC	13.9M	-	0.001	1.97	4.13±0.09
FS2	41.2M	6.0	**0.010**	4.18	3.29±0.09
VITS	36.3M	12.4	0.074	2.52	3.71±0.08
GRAD-10	**14.8M**	7.8	0.049	3.44	3.49±0.08
GRAD-4	"	"	0.019	3.69	3.20±0.09
GCFM-4	"	"	0.019	2.70	3.57±0.08
MAT-10	18.2M	**4.8**	0.038	**2.09**	**3.84**±0.08
MAT-4	"	"	0.019	2.15	3.77±0.07
MAT-2	"	"	0.015	2.34	3.65±0.08

1. References

[1] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Proc. NeurIPS*, 2019.

[2] S. Mehta, R. Tu, J. Beskow, É. Székely, and G. E. Henter, "Matcha-TTS: A fast TTS architecture with conditional flow matching," in *Proc. ICASSP*, 2024, pp. 11 341–11 345.

[3] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," in *Proc. ICLR*, 2023.

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. CVPR*, 2022, pp. 10 684–10 695.

A Study of Continual Test-Time Adaptation for Speech Recognition

Xinying Wei, Mingjie Chen, Thomas Hain

Speech and Hearing Research Group, University of Sheffield, United Kindom
{xwei45, mingjie.chen, t.hain}@sheffield.ac.uk

Automatic speech recognition (ASR) systems often face a decline in performance when there is a significant difference between training and test data. Therefore numerous methods have been presented to adapt ASR systems to a variety of situations. Test-time adaptation specifically aims to adapt a trained model on unlabelled target test data, without accessing source data for training. Test-time adaptation can be described as finding new model parameters from single samples, i.e.

$$\theta_i = g(f_{\theta_s}, \mathbf{x}_i), \tag{1}$$

where $i \in \mathbb{N}_+$ refers to the target utterance index along adaptation, $\mathbf{x}_i$ is the target utterance, f_{θ_s} is the model function with the source parameter θ_s that transforms the target utterance $\mathbf{x}_i$ to its predicted probabilities of the transcript, and g is the adaptation algorithm that transforms the source model f_{θ_s} and the target utterance $\mathbf{x}_i$ to the new parameters θ_i. It aims to find the θ_i that yields the minimal word error rate (WER) on $\mathbf{x}_i$.

Only recently has there been renewed interest in test time adaptation for ASR on a per utterance level. One common approach is to use fine-tuning with additional sharpening of the target probability distributions through entropy minimisation [1, 2]. One method [1] adds minimum class confusion alongside entropy minimisation to reduce the dependence between classes. Another method recently proposed [2] applies negative sampling together with entropy minimisation to penalise predictions with low confidence. Existing test-time adaptation methods always take the original model parameters θ_0 as input.

$$\theta_i = g(f_{\theta_0}, \mathbf{x}_i), \tag{2}$$

The study presented here aims to explore how future adaptation performance could benefit from knowledge gained from past utterance samples. When test-time adaptation makes use of the previously obtained model parameters θ_t during the utterance-level adaptation process, it is referred to as *continual test-time adaptation*.

$$\theta_i = g(f_{\theta_{i-1}}, \mathbf{x}_i), \tag{3}$$

Test-time adaptation methods employed in current ASR test-time adaptation work [1, 2] include entropy minimisation, minimum class confusion and negative sampling.

Rénni Entropy Minimisation. Given an utterance of L frames with C classes, Rénni entropy minimisation in [2] is defined as

$$L_{\text{gem}} = \frac{1}{L}\sum_{i=1}^{L}\frac{1}{1-\alpha}\log\left(\sum_{j=1}^{C} p_{ij}^{\alpha}\right) \tag{4}$$

where $\alpha \in (1, \infty)$ is a hyperparameter that emphasises less likely classes, and p_{ij} is the probability of the j-th class on the i-th frame.

Shannon Entropy Minimisation. It is a special case of Rényi entropy when $\alpha \to 1$. Given an utterance with L frames with C classes, Shannon entropy minimisation in [1] is defined as

$$L_{\text{sem}} = -\frac{1}{L}\sum_{i=1}^{L}\sum_{j=1}^{C} p_{ij}\log p_{ij} \tag{5}$$

where p_{ij} is the probability of the j-th class on the i-th frame.

Minimum Class Confusion. Given an utterance of L frames with C classes, minimum class confusion in [1] is defined as

$$L_{\text{mcc}} = \sum_{j=1}^{C}\sum_{j'\neq j}^{C}\mathbf{p}_{\cdot j}^{\top}\mathbf{p}_{\cdot j'} \tag{6}$$

where $\mathbf{p}_{\cdot j}$ is the probabilities of the j-th class of the L frames.

Negative Sampling. Negative sampling penalises the classes with smaller probabilities than a specified threshold. Given an utterance of L frames, negative sampling in [2] is defined as

$$L_{\text{ns}} = -\frac{1}{L}\sum_{i=1}^{L}\log\left(1-\sum_{j=1}^{C} I_{ij}p_{ij}\right) \tag{7}$$

where $p'_{ij} = \frac{\exp(o_{ij})}{\sum_{j'=1}^{C}\exp(o_{ij'})}$, $p_{ij} = \frac{\exp(o_{ij}/T)}{\sum_{j'=1}^{C}\exp(o_{ij'}/T)}$ is p'_{ij} with temperature T, and o_{ij} is the logit value of the j-th class on the i-th frame. $I_{ij} = 1$ when p'_{ij} is smaller than a threshold τ, and $I_{ij} = 0$ for $p'_{ij} >= \tau$.

The objective functions used by [1] and [2] are defined as

$$L_{\text{SUTA}} = \lambda_{\text{SUTA}}L_{\text{sem}} + (1-\lambda_{\text{SUTA}})L_{\text{mcc}} \tag{8}$$

$$L_{\text{SGEM}} = \lambda_{\text{SGEM}}L_{\text{gem}} + (1-\lambda_{\text{SGEM}})L_{\text{ns}} \tag{9}$$

respectively, where $\lambda_{\text{SUTA}}, \lambda_{\text{SGEM}} \in [0, 1]$ are hyperparameters that balance component objectives.

This study evaluates existing test-time adaptation methods using both a continual testing regime as well as instantaneous performance, by showing WER as a function of historic and future data. WER change trends are compared between methods over adaptation time, and the stability of each method is explored by showing the development of per utterance WER distributions.

1. References

[1] G.-T. Lin, S.-W. Li, and H. yi Lee, "Listen, Adapt, Better WER: Source-free Single-utterance Test-time Adaptation for Automatic Speech Recognition," in *Proc. Interspeech 2022*, 2022, pp. 2198–2202.

[2] C. Kim, J. Park, H. Shim, and E. Yang, "SGEM: Test-Time Adaptation for Automatic Speech Recognition via Sequential-Level Generalized Entropy Minimization," in *Proc. Interspeech 2023*, 2023, pp. 3367–3371.

Prosodic cues and processing of Bulgarian-English code-switched speech

Yoana I. Dancheva[1], *Margreet Vogelzang*[12], *Ianthi M. Tsimpli*[1]

[1]University of Cambridge, UK
[2]Newcastle University, UK

yid20@cam.ac.uk, mv498@cam.ac.uk, imt20@cam.ac.uk

Code-switching (CS), the use of lexical and grammatical elements of more than one language, is a common phenomenon in bilingual speech. Yet despite the vast research on the syntax of CS sentences, little is known about the relationship between syntax and prosody. Recent studies have delved into the processing of CS speech by bilinguals [1, 2, 3] but we still have much to examine to understand what makes the processing of CS speech different to the processing of unilingual speech. Bulgarian-English CS is under-researched, and the prosody and processing of Bulgarian-English utterances have never been examined. We, therefore, strive to fill this gap in the literature by investigating the prosodic patterns and the processing of Bulgarian-English CS speech.

In the present research, we aim to answer the following questions:

1. What prosodic cues exist in the speech of Bulgarian-English bilinguals, which can potentially be used by the listener to predict an upcoming switch?
2. How does the processing of CS utterances differ from the processing of unilingual utterances in Bulgarian-English bilinguals living in a country of the L2 (namely, the UK)?

In the first study, we examined spoken data from fifteen Bulgarian-English bilinguals extracted from a pre-existing CS corpus [4]. We analysed syntactically similar CS and unilingual Bulgarian utterances in *Praat* for fundamental frequency (F0), speech rate, and voice onset time (VOT) [following 3]. The CS sentences included either an insertion or an alternation, and the matrix language was Bulgarian (English being the embedded language). The findings show a lower F0 before the CS and a higher F0 before the unilingual equivalent. A significant correlation between speech rate and CS shows that speech rate is slower before the CS, with the effect being strongest before CS nouns. The VOT also appears shorter in CS utterances. These findings confirm the presence of prosodic cues in CS speech, and facilitate further research on the role of prosodic cues in the processing of CS speech.

In the second study, we aim to expand on the findings from the first study, and additionally aim to understand how the processing of CS sentences is different from the processing of unilingual sentences. We are conducting a web-based sentence repetition task which contains CS utterances (insertional and alternational) and unilingual Bulgarian and unilingual English utterances. Firstly, participants are asked to listen to audio of a CS or unilingual sentence, which they then have to judge for "naturalness" by selecting a number on a Likert scale of 1 to 6 (1 = very unnatural, 6 = very natural). Upon selection of the score, the participants are asked to repeat the sentence they have heard aloud. The sentences are syntactically (and in some cases semantically) based on those found in the CS corpus used in the previous study. In the first stage of this study, we look at the repetition accuracy scores and the reaction times (extracted through the judgement of the audios). These findings will help us learn more about the processing of unilingual and CS utterances. In the second stage of the study, we will examine the prosodic properties of the different language conditions.

Preliminary data from thirty-three Bulgarian-English bilingual participants shows that there are no significant differences between the reaction times of the unilingual and the CS conditions. However, the repetition accuracy of the unilingual English condition is significantly lower than the accuracy of the CS conditions and the unilingual Bulgarian condition. This higher error rate points towards a higher processing cost in the L2 condition, which will be examined further by looking at the participants' language background and language use.

1. References

[1] J. Hofweber and T. Marinis, "What sentence repetition tasks can reveal about the processing effort associated with different types of code-switching," *Languages*, vol. 8, no. 1, 2023.

[2] M. A. Johns and J. Steuck, "Is codeswitching easy or difficult? testing processing cost through the prosodic structure of bilingual speech," *Cognition*, vol. 211, 2021.

[3] M. Fricke, J. F. Kroll, and P. E. Dussias, "Phonetic variation in bilingual speech: A lens for studying the production–comprehension link," *Journal of Memory and Language*, vol. 89, p. 110–137, 2016.

[4] Y. I. Dancheva, "Code-switching in Bulgarian-English bilinguals: Morphosyntactic properties and the role of attitudes and exposure," Master's thesis, University of Cambridge, 2022.

On the Optimization Aspects of Variational Appoaches to Speech Representation Learning

Sung-Lin Yeh, Hao Tang

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

sunglin.yeh@ed.ac.uk, hao.tang@ed.ac.uk

Abstract

Previous work on self-supervised speech representation learning shares striking similarities and calls for a unifying formalism. In this paper, we propose a flexible variational objective that generalizes recent approaches, such as VQ-APC and HuBERT, and connects to other approaches, such as VQ-CPC and wav2vec 2.0. Our main contribution is a discrete latent variable model for self-supervised learning in speech, providing a framework to instantiate and connect to prior work. The learned representations of our proposed approach obtain sizable improvements on phonetic classification, speaker verification, and automatic speech recognition over prior work. Our approach provides an optimization advantage compared to HuBERT and WavLM as several individually optimized terms in prior work can be optimized jointly in ours. Empirically, we show that the proposed loss function achieves lower loss values. Moreover, the optimization advantage correlates with the downstream performance.

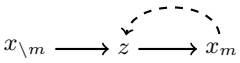


Figure 1: *The proposed graphical model in masked prediction. The input is partitioned to unmasked frames $x_{\setminus m}$ and the masked ones x_m. The dashed line indicates the variational distribution $q(z|x_m)$.*

1. Methods

Prior work on self-supervised speech representation learning is based on the prediction of one part of the input given the rest. The prediction process consists of two critical components: the partition of the input and the prediction of the partitioned contexts. To generalize prior work on speech representation learning, we consider widely adopted partition schemes such as future prediction and masked prediction. Under a partition scheme, we formulate a graphical model with discrete latent variables to model the prediction of the partitioned contexts.

Taking masked prediction as an example. The input x is first partitioned into unmasked frames $x_{\setminus m}$ and masked frames x_m. Then a graphical model is used model the generative process from $x_{\setminus m}$ to x_m. The variational lower bound of the non-tractable likelihood is

$$\mathcal{L}_{\text{MLM-VLB}} = \sum_{i \in m} \big[\mathrm{KL}\big(q(z_i|x_i) || p(z_i|x_{\setminus m})\big) + \mathbb{E}_{z_i \sim q}[-\log p(x_i|z_i)]\big] \quad (1)$$

$$\geq \sum_{i \in m} -\log p(x_i|x_{\setminus m}) = -\log p(x_m|x_{\setminus m}), \quad (2)$$

where we make two assumptions. The first is that the variational distribution q only depends the current frame. The second is that the masked frames are independent of each other given the unmasked ones. The distribution $p(z_i|x_{\setminus m})$ is usually parameterized with Transformer encoders.

2. Optimization

With the variational formalism of the prediction process, we show that HuBERT and WavLM are in fact optimizing the proposed variational loss function.

In HuBERT, k-means is applied to acoustic frames, and a Transformer encoder is used to predict cluster assignments of masked frames. The HuBERT loss function is defined as

$$\mathcal{L}_{\text{HuBERT}} = \sum_{i \in m} \mathbb{E}_{z_i \sim q}\big[\log p(z_i|x_{\setminus m})\big], \quad (3)$$

where q is a point-mass distribution on the cluster assignment of k-means and z_i is the assignment, $p(z_i|x_{\setminus m})$ is modeled by a Transformer encoder. The HuBERT loss is equivalent to the KL term of the variational lower bound (Eq (2)), in which the entropy of point-mass q is 0.

Moreover, by assuming $p(x_i|z_i)$ to be Gaussian, we can instantiate the k-means loss with the reconstruction term in Eq (2):

$$\mathbb{E}_{z_i \sim q}\big[p(x_i|z_i)\big] = \mathbb{E}_{z_i \sim q}\big[-\frac{1}{2}||x_i - V\mathbf{1}_{z_i}||^2\big],$$

where V is the centroids, also know as codebook. We have omitted constant terms for clarity. Again, q is hard in k-means, $q = \arg\max_q ||x_i - V\mathbf{1}_{z_i}||^2$. The two-step optimization in HuBERT is summarized below

$$q = \arg\max_q \mathbb{E}_{z_i \sim q}\big[||x_i - V\mathbf{1}_{z_i}||^2\big],$$

$$\theta = \arg\max_\theta \mathcal{L}_{\text{MLM-VLB}},$$

where θ is the parameters of the Transformer encoder. We compare it to the proposed approach:

$$\theta, q = \arg\max_{\theta, q} \mathcal{L}_{\text{MLM-VLB}},$$

where q is parameterized with V.

Factors influencing vowel categorisation flexibility

Stephanie Cooper, Brechtje Post

Phonetics Lab, University of Cambridge, United Kingdom

sc2346@cam.ac.uk, bmbp2@cam.ac.uk

It is generally assumed that listeners use phonetic categories to break incoming speech into smaller segments. These categories consist of peripheral boundaries and a central prototype or "best exemplar", which are determined by a range of factors including the listener's native dialect [1]. Phonetic category boundaries are flexible, making listeners remarkably capable of interpreting substantial variation in the speech signal. This is particularly evident in vowel categorisation, which has been found to be more flexible than consonant categorisation [2, 3, 4]. Recent work [5] shows that listeners can accept variants of a native vowel that are novel (not existing in their dialect) or even familiar (existing as another vowel in their dialect) when they are presented in a nonword context. Some of the factors which have been suggested to facilitate or impede this flexibility are spectral distance from the listener's native prototype of the vowel category, listener dialect, and lexical bias, which exerts a top-down effect onto phonetic categorisation [6, 7].

The current project aims to develop a method by which vowel categorisation flexibility can be investigated for different populations, but also to explore the relative weight a number of factors - some not explored before - may have on flexibility of category boundaries, including position of the incoming vowel along an F1 continuum, frontness of the vowel, listener dialect, and word status.

To investigate perceptual boundaries of vowels, continua of VC and CVC-structured sequences spoken by a male Standard Southern British English speaker are used in a forced choice identification task (Figure 1). We are interested in dialectal variation and flexibility around the DRESS and THOUGHT vowels, so each continuum contains vowels resynthesised in steps along either the /iː/ - /aː/ or the /uː/ - /ɒ/ continuum, but with the same consonant context, e.g. ranging from *heed* through *hid* and *head* to *had*. The stimuli also vary in their lexical status: some continua contain only words, some only nonwords, and some contain both. Listeners from three different dialect groups (Standard Southern British, Geordie, New Zealand) will take part in the identification task.

Preliminary results will be processed by the time of UK Speech. Generalised additive mixed models will examine differences in categorisation curves (% acceptance) for each point along the F1 continuum between groups, vowels, and lexical contexts. Peaks of acceptance will indicate where participant category prototypes sit, and gradients of slopes from these prototypes will indicate how wide the category boundaries are.

This study will provide further insight into the factors that need to be taken into account when examining vowel categorisation. Provided it is successful, the method developed here is highly adaptable and can be manipulated to examine a range of different vowels and factors. It can therefore be used as a baseline test for future work when researchers need to understand the boundaries of a given group's native categories.

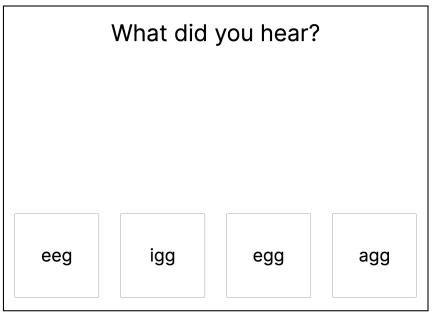


Figure 1: *Example of forced choice identification task: responding to a stimulus with a vowel along the /i:/ - /a:/ continuum.*

This project is funded by the Harding Distinguished Postgraduate Scholarship.

1. References

[1] X. Xie, R. M. Theodore, and E. B. Myers, "More than a boundary shift: Perceptual adaptation to foreign-accented speech reshapes the internal structure of phonetic categories," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 43, no. 1, pp. 206–217, Jan. 2017.

[2] S. Cooper and S. Cooper, "Exposure-independent comprehension of Greek-accented speech: Evidence from New Zealand listeners," in *Proceedings of the 20th International Congress of Phonetic Sciences*, 2023, pp. 6–10.

[3] J. Shaw, C. Best, K. Mulak, G. Docherty, B. Evans, P. Foulkes, J. Hay, J. Al-Tamimi, K. Mair, M. Peek, and S. Wood, "Effects of short-term exposure to unfamiliar regional accents: Australians' categorization of London and Yorkshire English consonants," in *Proceedings of the 15th Australasian International Speech Science and Technology Conference*, Dec. 2014, pp. 71–74.

[4] J. Shaw, C. T. Best, G. Docherty, B. G. Evans, P. Foulkes, J. Hay, and K. E. Mulak, "Resilience of English vowel perception across regional accent variation," *Laboratory Phonology*, vol. 9, no. 1, pp. 1–36, Jul. 2018.

[5] S. Cooper and B. Post, "Flexibility of vowel categorisation in newly learned words," Conference presentation at Colloquium of British Association of Academic Phoneticians, Cardiff, United Kingdom, 2024.

[6] R. A. Fox, "Effect of lexical status on phonetic categorization," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 10, no. 4, pp. 526–540, 1984.

[7] M. G. Gaskell and W. D. Marslen-Wilson, "Mechanisms of phonological inference in speech perception," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 24, no. 2, pp. 380–396, 1998.

Effects of Voice Similarity on the Quality of Experience of Voice Assistants

Crisron Rudolf Lucas[1], *Andrew Hines*[2]

[1]UCD School of Computer Science, Ireland
[2]UCD School of Computer Science, Ireland
crisron.lucas@ucdconnect.ie, andrew.hines@ucd.ie

Recent state-of-the-art Text-To-Speech (TTS) voices are able to achieve high Mean-Opinion-Scores (MOS) in terms of naturalness and intelligibility. However, less has been studied on the Quality of Experience (QoE) of TTS particularly on voice preference. For natural speech, related studies on the aspects of voice such as loudness, pitch, and speech rate show varied results on voice preference based on reputation, trust, reliance, dominance and inter-gender attractiveness. The concept of 'self-voice' and voice similarity as criteria for preferability has also been studied in the literature. With these in mind, we aim to systematically test whether matching the prosodic characteristics of TTS voice with the listeners' voice will have a positive correlation in terms of voice preference.

1. Introduction

1.1. Text-To-Speech Systems

Neural TTS and other state-of-the-art TTS systems are mostly evaluated in terms of naturalness and intelligibility [1]. Among the notable state-of-the-art systems yielding relatively high Mean-Opinion-Scores (MOS) are Tacotron2, Fastpitch, FastSpeech, Matcha TTS, and Speech T5 as shown in table 1.

Table 1: *List of recent TTS systems and their performance*

Year	Author	TTS	Dataset used	MOS Performance
2018	Shen et. al	Tacotron2	LJSpeech	4.526 (Excellent)
2021	Lancucki	FASTPITCH	LJSpeech	4.01 (Good)
2021	Ren et al	FastSpeech2	LJSpeech	3.83 (Fair)
2022	Ao et al	SpeechT5	LibriSpeech, LibriTTS (training) NISQA-TTS (evaluation)	3.65 (Fair)
2023	Mehta et al	MatchaTTS	LJSpeech	3.84 (Fair)

For reference, a MOS of 4.3-4.5 is already considered as excellent or human-level in terms of speech quality. While the systems mentioned have high intelligibility, most of them can still improve on naturalness and controllability [2]. Since most of the TTS systems are trained with entangled speaker-dependent and environmental features like prosody, accent, and background noise, the synthesized speech outputs have limited flexibility and variations in their prosodic features. The work done in [3] tried to address the problem using unsupervised style modeling with global style tokens (GST) which are controllable speech parameters. A variation of the previous implementation using sequence to sequence neural network was also done by [4]. In their work, they used fundamental frequency (pitch), pitch range, phone duration, speech energy, and spectral tilt as main parameters for controllability. Google TTS, one of the commercially available state-of-the-art TTS, has a decoder capable of changing the pitch (in terms of semitones), speaking rate, and also has support for Speech Synthesis Markup Language (SSML).

1.2. Voice preference

Voice preference is a Quality-of-Experience (QoE) concept involving different aspects of the voice and depends on different use cases or applications. It should be differentiated from the standard speech quality metrics such as naturalness and intelligibility which deal with how "human-like", clear, or audible the speech is perceived by the listener. Voice preference is defined by many qualitative descriptors such as attractiveness, dominance, and trustworthiness. These qualitative descriptors have already been linked to pitch and speech rate which are also metrics for naturalness. However, the direct link of these speech parameters to voice preference in the context of self-voice has not yet been thoroughly investigated for speech synthesis applications.

2. Objective

In this paper, we aim to explore the influence factors affecting voice preference particularly the individual and combined effects of varying speech rate and pitch in the context of matching the speech characteristics with the listener for improved quality of experience. In conducting this study, we are trying to answer the following research question: **Does matching pitch and speech rate improve the QoE and voice preference of synthetic speech?**

3. References

[1] Y.-Y. Chang. [Online]. Available: https://aclanthology.org/O11-1004.pdf

[2] X. Tan, T. Qin, F. Soong, and T.-Y. Liu. [Online]. Available: https://arxiv.org/pdf/2106.15561.pdf

[3] Y. Wang. [Online]. Available: https://arxiv.org/pdf/1803.09017.pdf

[4] T. Ratio. [Online]. Available: https://arxiv.org/pdf/2009.06775.pdf

HAFFORMER: A Hierarchical Attention-Free Framework for Alzheimer's Disease Detection From Spontaneous Speech

Zhongren Dong[1], *Zixing Zhang*[1], *Weixiang Xu*[1], *Jing Han*[2], *Jianjun Ou*[3], *Björn W Schuller*[4]

[1]College of Computer Science and Electronic Engineering, Hunan University, China
[2]Department of Computer Science and Technology, University of Cambridge, UK
[3]Department of Psychiatry, The Second Xiangya Hospital, Central South University, China
[4]GLAM, Department of Computing, Imperial College London, UK

zixingzhang@hnu.edu.cn, jh2298@cam.ac.uk

Automatically detecting Alzheimer's Disease (AD) from spontaneous speech plays an important role in its early diagnosis. Recent approaches highly rely on the Transformer architectures due to its efficiency in modelling long-range context dependencies. However, the quadratic increase in computational complexity associated with self-attention and the length of audio poses a challenge when deploying such models on edge devices. In this context, we construct a novel framework, namely Hierarchical Attention-Free Transformer (HAFFormer) (cf Fig. 1), to better deal with long speech for AD detection. Specifically, we employ an attention-free module of Multi-Scale Depthwise Convolution (MSDW) to replace the self-attention and thus avoid the expensive computation (cf Fig. 2 left panel), and a GELU-based Gated Linear Unit to replace the feedforward layer, aiming to automatically filter out the redundant information. Moreover, we design a hierarchical structure to force it to learn a variety of information grains, from the frame level to the dialogue level. By conducting extensive experiments on the ADReSS-M dataset, the introduced HAFFormer can achieve competitive results (82.6% accuracy) with other recent work, but with significant computational complexity and model size reduction compared to the standard Transformer. This shows the efficiency of HAFFormer in dealing with long audio for AD detection. An ablation study on the number of layers in the hierarchical model was also conducted, suggesting that a moderate number of layers are effective, while increasing the model complexity beyond a certain point may lead to overfitting on the dataset (cf Fig. 2 bottom panel).

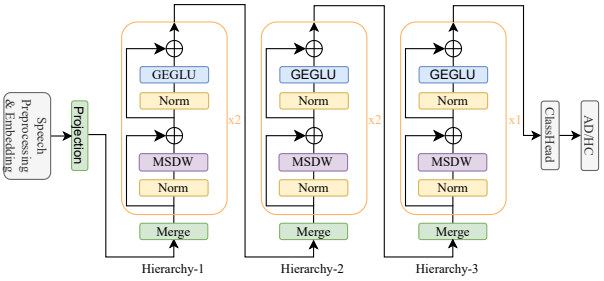


Figure 1: *An overview of the proposed Hierarchical Attention-Free Transformer (HAFFormer) framework for Alzheimer's disease detection. MSDW: multi-scale depth-wise convolution; GEGLU: GELU-based Gated Linear Units; AD/HC: Alzheimer's Disease or Healthy Control.*

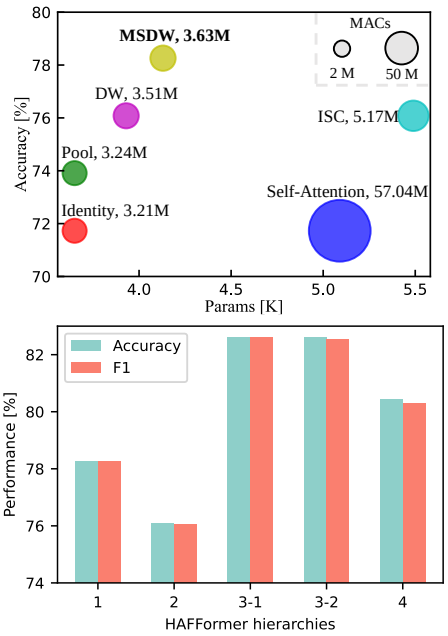


Figure 2: *Top: Model performance (accuracy) vs model size (number of parameters) when taking different types of token mixers under the same channel mixer (i. e., FFN). MACs: Multiply–accumulate operations. Bottom: Performance of the proposed HAFFormer when taking different numbers of hierarchy. 3-1/2 indicates the three hierarchies but the last one contains one or two blocks.*

Liquid+sonorant epenthesis in Connemara Irish English: bilinguals vs. monolinguals

Kate Tallon

Phonetics and Speech Laboratory, School of Linguistic, Speech and Communication Sciences, Trinity College Dublin, Ireland
tallonka@tcd.ie

Introduction & Background

Epenthesis of a schwa vowel within liquid+sonorant coda clusters (e.g., [ˈfiləm] 'film') is a widely reported feature of Irish English (IrE). Frequent in /lm/ clusters, it can also occur in /ln/ clusters (e.g. 'kiln'), and in /rm/, /rl/ and /rn/ (e.g., *farm, girl, barn*) [1]. Although widely viewed as a transfer from Irish, where epenthesis in sonorant+consonant clusters is prevalent (e.g., [borəb] *borb* 'abrupt', [dorəxə] *dorcha* 'dark' [2]), an alternative source, suggested in [3], would be the older forms of Scots and English brought to Ireland in past centuries. This paper systematically investigates liquid+sonorant clusters in the English of (i) Irish-English bilinguals and (ii) English monolinguals in the Connemara Gaeltacht. Additional sociolinguistic factors considered include age, gender and Linguistic Profile Score (LPS) – a measure of language dominance.

Methods

Table 1 summarises the occurrences of liquid+sonorant clusters in read and spontaneous speech, from 12 bilingual speakers (5M, 7F) and 12 monolingual speakers (6M, 6F) (age range 20-79, mean=44, SD=18).

Table 1: *Token count of liquid+sonorant clusters.*

	lm	rl	rm	rn	Total
Bilingual	41	38	37	48	164
Monolingual	37	39	36	38	150

Generalised Linear Mixed Models (GLMM) were run using lme4 [4] in R [5], to (i) compare bilinguals vs. monolinguals, and (ii) investigate sociolinguistic predictors of epenthesis.

Results

The GLMM results revealed that bilinguals had significantly higher rates of epenthesis than monolinguals ($p < 0.05$). Furthermore, while monolinguals only showed evidence of epenthesis in /lm/ clusters, bilinguals had epenthesis in all clusters and additionally displayed cases of final syllabic /l/ and /n/ in words such as 'girl' and 'barn'. For bilinguals, age was a significant predictor ($p < 0.001$), whereas no effect of age was found for monolinguals. Additionally, there was a significant positive correlation between LPS (dominance in Irish) and epenthesis for bilinguals (p <0.001). Gender was not found to be significant for either group of speakers. Figure 1 displays the rates of epenthesis for bilinguals, by age group and gender.

Discussion & Conclusion

A clear bilingual-monolingual difference emerges, and in the case of bilinguals, there are clear effects of age, and LPS (language dominance). Do these data shed light on the possible origins of IrE epenthesis? They do seem to support an Irish transfer source – perhaps not too surprising, given the prevalence of epenthesis in Irish. However, one cannot definitively rule out the possibility of an English or Scots source for this feature, as suggested by [2], but even so, it seems reasonable to assume Irish would be, at the very least, a reinforcing factor. The monolingual data, on the other hand, shows levelling towards a supraregional variety of IrE, with relatively few occurrences of epenthesis. This study adds to our knowledge of Connemara English spoken by both Irish-English bilinguals and English monolinguals, adding further to the discussion of IrE features with multiple possible origins.

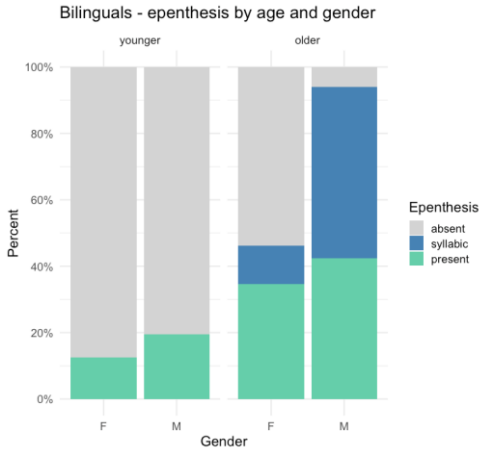


Figure 1: *Epenthesis rates in bilinguals by age and gender.*

Acknowledgments

This research was funded by the Irish Research Council, award no. GOIPG/2019/2448.

1. References

[1] Hickey, R. (2007). Irish English: History and present day forms. Cambridge University Press.
[2] Ní Chiosáin, M. (1999). Syllables and phonotactics in Irish. *The syllable: views and facts*, *45*, 551-576.
[3] Maguire, W. (2018). The Origins of Epenthesis in Liquid+Sonorant Clusters in Mid-Ulster English. *Transactions of the Philological Society*, *116*(3), 484-508.
[4] Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1 - 48.
[5] R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. In R Foundation for Statistical Computing.

Investigating the Emergent Audio Classification Ability of ASR Foundation Models

Rao Ma, Adian Liusie*, Mark J. F. Gales, Kate M. Knill*

ALTA Institute, Department of Engineering, University of Cambridge

rm2114@cam.ac.uk, al826@cam.ac.uk, mjfg@eng.cam.ac.uk, kmk1001@cam.ac.uk

1. Introduction

Text and vision foundation models can perform many tasks in a zero-shot setting, a desirable property that enables these systems to be applied in general and low-resource settings. There has been far less work, however, on the zero-shot abilities of ASR foundation models, with these systems typically fine-tuned to specific tasks or constrained to applications that match their training criterion and data annotation. In this work, we investigate the ability of Whisper, an ASR foundation model trained primarily for speech recognition, to perform zero-shot audio classification. We use simple template-based text prompts at the decoder and use the resulting decoding probabilities to generate zero-shot predictions. Without training the model on extra data or adding any new parameters, we demonstrate that Whisper shows promising zero-shot classification performance on a range of 8 audio-classification datasets. One important step to unlock the emergent ability is debiasing, where a simple unsupervised reweighting method of the class probabilities yields consistent significant performance gains. We further show that performance increases with model size, implying that as ASR foundation models scale up, they may exhibit improved zero-shot performance.

2. Methods

2.1. Zero-shot Prompting

In this work, we use a simple template filling prompting strategy, where given an input audio sample, we assess the probability of decoding a label sequence associated with each classification class. We leverage various 'prompts' by considering different templates to represent the label sequences. To convert likelihoods to class probabilities, we treat the ASR system as a generative classifier:

Let $P_\theta(x|s)$ represent the likelihood associated with ASR decoding the word sequence $x \in \mathcal{X}$ given an input audio s. Let $y \in \{\omega_1, \omega_2, ...\omega_K\}$ be one of K possible output classes, and $t(\omega_k) \in \mathcal{X}$ represent a particular mapping of a class to a word sequence representing the class. We assume that the zero-shot ASR classification probability $\tilde{P}_\theta$ for a particular class is proportional to the likelihood of generating each respective class label sequence given the input audio:

$$\tilde{P}_\theta(y = \omega_k|s) = \frac{P_\theta(t(\omega_k)|s)}{\sum_{\omega_j} P_\theta(t(\omega_j)|s)} \tag{1}$$

The model's prediction is then the class with the highest associated probability:

$$\hat{y} = \underset{\omega}{\text{argmax}} \tilde{P}_\theta(\omega|s) \tag{2}$$

* Equal Contribution.

2.2. Task Calibration

A concern with the zero-shot prompting approach described above is the potential presence of implicit biases. One can reweight the outputs of the classifier by introducing weights $\alpha_{1:K}$ to rescale the probabilities.

$$\hat{P}_\theta(\omega_k|s, \alpha_{1:K}) = \frac{\alpha_k \tilde{P}_\theta(\omega_k|s)}{\sum_j \alpha_j \tilde{P}_\theta(\omega_j|s)} \tag{3}$$

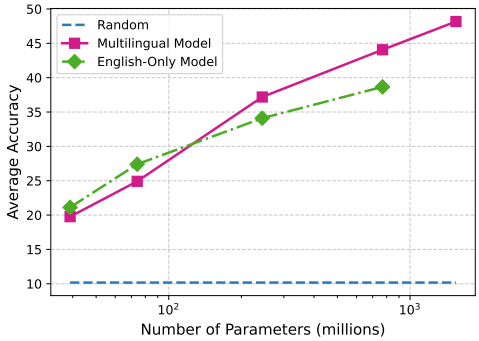
Assuming that unsupervised data is available for a particular task (or if all the evaluation is available as an unsupervised set), the output probabilities can be reweighted to ensure that the corresponding output prior matches the expected true prior, done by finding the weights $\bar{\alpha}_{1:K}$ that ensure such a prior,

$$\hat{P}_\theta(\omega_k|\alpha_{1:K}) = \mathbb{E}_s\{\hat{P}_\theta(\omega_k|s, \alpha_{1:K})\} \tag{4}$$

$$\bar{\alpha}_{1:K} = \underset{\alpha_{1:K}}{\text{argmin}} \sum_{\forall \omega} |\hat{P}_\theta(\omega|\alpha_{1:K}) - P(\omega)| \tag{5}$$

Where $P(\omega)$ is the true prior for the considered task. In cases where the underlying class distribution is not known, the prior can be assumed to be uniform, $P(\omega) = \frac{1}{K}$, which is the assumption made throughout this paper. The solution has a single free variable, but by constraining $\alpha_1 = 1$ one can find an exact solution that perfectly matches the prior, a search which can be done efficiently.

3. Experiments



We assess our systems across 8 diverse audio classification datasets, encompassing 6 distinct tasks. Results on the average classification accuracy show that Whisper performs zero-shot audio classification much better than random. Using template prompts and task calibration, Whisper large-v2 achieves an average zero-shot accuracy of 48.2%, considerably better than the average random performance (10.4%). Further, increasing parameter size yields a performance boost and the multilingual Whisper performs better than the English-only model.

Oral Session B

SummaryMixing: A Linear-Complexity Alternative to Self-Attention for Speech Recognition and Understanding

Titouan Parcollet, Rogier van Dalen, Shucong Zhang, Sourav Bhattacharya

Samsung AI Centre, Cambridge, United-Kingdom

{t.parcollet|r.vandalen|s1.zhang|sourav.b1}@samsung.com

Abstract

Modern speech processing systems rely on self-attention. Unfortunately, token mixing with self-attention takes quadratic time in the length of the speech utterance, slowing down inference and training and increasing memory consumption. Cheaper alternatives to self-attention for ASR have been developed, but they fail to consistently reach the same level of accuracy. This paper, therefore, proposes a novel linear-time alternative to self-attention. It summarises an utterance with the mean over vectors for all time steps. This single summary is then combined with time-specific information. We call this method "SummaryMixing". SummaryMixing in SOTA ASR models makes it feasible to preserve or exceed previous speech recognition performance while making training and inference up to 28 % faster and reducing memory use by half.

1. Introduction

Automatic speech recognition (ASR) has greatly benefitted from deep learning. However, in a push to improve recognition accuracy, ASR systems have steadily increased in model size. Modern industry-scale ASR models often contain hundreds of millions or even billions of neural parameters. Training these models requires many GPU hours and results in large carbon footprints. At the same time, it has become more important for models to be deployed on edge devices, where resources are scarce. Thus, this paper focuses on improving the efficiency of speech processing models.

At the core of current state-of-the-art (SOTA) speech systems are multi-head self-attention (MHSA) cells. MHSA learns interactions between pairs of frames originating from the speech signal, and the interaction is also referred to as token mixing. Most state-of-the-art speech models use MHSA. However, considering each pair of frames takes quadratic time in the input sequence length, making MHSA costly.

Recent works have pointed out that under some conditions, pair-wise self-attention operations in practice behave like linear operations. For example, [1] first showed that the upper encoder layers in trained Transformer-based ASR models behave like feed-forward layers, which is also verified by [2] for Conformer models. Furthermore, [3] demonstrated that the attention weights of trained Branchformer models tend to all have the same value, reducing the attention mechanism to computing an average.

Therefore, this work introduces an alternative to self-attention that takes only linear time in the sequence length. Instead of computing pair-wise interactions, it summarises a whole utterance as a mean over a contribution for each time step. The obtained summary is then fed back to each time step. We call this method "SummaryMixing".

Table 1: *Speech recognition results on encoder-decoder models with CTC plus Transformer decoding on the Librispeech dataset. "GPU hours" is the total training time. "VRAM" reports the peak amount of VRAM during training.*

Encoder	**Variant**	**WER %**			**GPU**	**VRAM**
		dev-clean	*test-clean*	*test-other*	hours	GB
ContextNet	—	3.3	2.3	5.9	160	25
Conformer	Self-attention	**2.8**	2.3	5.4	137	46
Branchformer	Self-attention	2.9	2.2	**5.1**	132	45
	CNN Only	3.1	2.4	5.7	**83**	22
	HyperMixer	3.1	2.3	5.6	126	30
	FastFormer	3.0	2.2	5.4	96	23
Proposed						
Conformer	SummaryMix.	**2.8**	**2.1**	**5.1**	98	**21**
Branchformer	SummaryMix.	2.9	2.2	**5.1**	105	26
+ SummaryMixing decoder		3.1	2.3	5.3	104	26

Our proposed SummaryMixing[1] achieves a training time reduction of up to 28% compared to MHSA. In decoding, its real-time factor does not increase with utterance length. SummaryMixing also halves the memory consumption in training and decoding, and reaches the performance of SOTA ASR systems on five datasets of different languages and acoustic conditions. These findings extend to other speech understanding tasks including spoken language understanding (SLU) and keyword spotting (KWS). To the best of our knowledge, it is the first time a linear-time method matches or surpasses the performance of MHSA for speech tasks across various scenarios.

2. References

[1] S. Zhang, E. Loweimi, P. Bell, and S. Renals, "On the usefulness of self-attention for automatic speech recognition with transformers," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 89–96.

[2] K. Shim, J. Choi, and W. Sung, "Understanding the role of self attention for efficient speech recognition," in *International Conference on Learning Representations*, 2022.

[3] Y. Peng, S. Dalmia, I. Lane, and S. Watanabe, "Branchformer: Parallel MLP-attention architectures to capture local and global context for speech recognition and understanding," in *International Conference on Machine Learning*. PMLR, 2022, pp. 17 627–17 643.

[1]The source code is available at https://github.com/SamsungLabs/SummaryMixing.

Can We Trust Explainable AI Methods on ASR? An Evaluation on Phoneme Recognition

Xiaoliang Wu, Peter Bell, Ajitha Rajan

School of Informatics, University of Edinburgh
x.wu-53@sms.ed.ac.uk, peter.bell@ed.ac.uk, arajan@ed.ac.uk

This paper[1] has been accepted and presented at ICASSP 2024. It focuses on the reliability of explainable AI (XAI) technologies applied to automatic speech recognition (ASR) systems. We outline our approach by addressing the motivation, existing challenges, methodologies, and experimental results.

Motivation: XAI has become critically important in fields such as image classification, boosting confidence among users and developers. However, the application of XAI within the speech domain remains limited. Assessing the reliability of XAI techniques in speech AI prior to their broader implementation is essential.

Challenge: A significant challenge in XAI is the absence of concrete ground truth, complicating the validation of explanations. In image classification, while certain features like a dog's eyes might be identified by XAI methods as key to classifying an image as a dog, verifying these explanations against an objective ground truth is often not possible. This challenge similarly affects the verification of explanations in speech recognition.

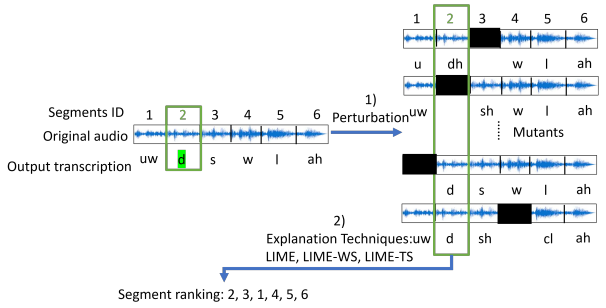


Figure 1: *An outline of generating an explanation for a phoneme appearing in the output transcription.*

Methodology: We address this challenge using the TIMIT Phoneme Recognition (PR) task, employing the standard recipe from the Kaldi toolkit. This task is predictable and controllable, with manual labeling and segmentation at the phoneme level providing a form of ground truth.

We then adapt the Local Interpretable Model-Agnostic Explanations (LIME[2]) and two variants—LIME-WS and LIME-TS for this audio-based task. We choose to adapt LIME owing to its popularity and effectiveness in the vision and text domain[3]. Our approach involves defining what is an explanation for a phoneme, generating explanations by perturbing audio segments and then evaluating the effectiveness of the explanation by comparing with the groundtruth.

- **What is an Explanation for a Phoneme?** In image tasks, models often highlight the areas of eyes, to explain why this image is labeled as a face. Similarly, for PR, an explanation should pinpoint key frames of speech for recognizing a phoneme. For example, as shown in the Figure 1, we have an input audio, alongside its original transcription. For the phoneme 'd' which appears in the transcription, the explanation is the importance ranking of segments in the input audio.
- **Framework of generating an Explanation:** Figure 1 presents an overview of generating an explanation for a phoneme in the transcription output. We start with an input audio and its segments. We then perturb the audio by masking out segments randomly. In Figure 1, segments shaded in black within each mutant are the ones that have been masked. For the phoneme of interest, 'd', those perturbations and the corresponding transcriptions are fed into three explanation techniques to compute the importance ranking of segments in the audio. Details of using LIME, LIME-WS and LIME-TS can be found in our paper[1].
- **Evaluate the effectiveness of an explanation:** Finally, we evaluate the validity of generated explanations using a new metric. The metric is defined as $validity = \frac{N_1}{N}$, where N is the total number of phonemes, N_1 is phonemes with the top-ranked segment matching the ground truth. For this metric, higher is better.

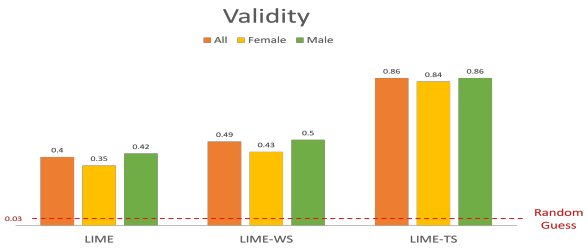


Figure 2: *validity of explanations generated by three explanation methods against a baseline of random guessing of segment ranking as an explanation using Kaldi PR.*

Results: The LIME-TS method is the most reliable and all techniques are significantly better than random guess, confirming the applicability of XAI in the speech domain. All three LIME based XAI techniques are more reliable for male over female speakers, reflecting the existing biases in the TIMIT dataset.

1. References

[1] **Xiaoliang Wu**, P. Bell, and A. Rajan, "Can we trust explainable ai methods on asr? an evaluation on phoneme recognition," 2023.

[2] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," *CoRR*, vol. abs/1602.04938, 2016. [Online]. Available: http://arxiv.org/abs/1602.04938

[3] N. Gill, M. Kurka, and W. Phan, "Machine learning interpretability with h2o driverless ai," 2019.

Poster Session C

Linear-Complexity Self-Supervised Learning for Speech Processing

Shucong Zhang, Titouan Parcollet, Rogier van Dalen, Sourav Bhattacharya

Samsung AI Centre, Cambridge, United-Kingdom
{s1.zhang|t.parcollet|r.vandalen|sourav.b1}@samsung.com

Abstract

Self-supervised learning (SSL) models usually require weeks of pre-training with dozens of high-end GPUs. These models typically have a multi-headed self-attention (MHSA) context encoder. However, MHSA takes quadratic time and space in the input length, contributing to the high pre-training cost. Linear-complexity alternatives to MHSA have been proposed. For instance, in supervised training, the SummaryMixing model is the first to outperform MHSA across multiple speech processing tasks. However, these cheaper alternatives have not been explored for SSL yet. This paper studies a linear-complexity context encoder for SSL for the first time. With better or equivalent performance for the automatic speech recognition, intent classification, emotion recognition, and automatic speaker verification downstream tasks of the MP3S benchmark, SummaryMixing reduces the pre-training time and peak VRAM of wav2vec 2.0 model by 18% and by 23%, respectively, leading to the pre-training of a 155M wav2vec 2.0 model finished within one week with 4 Tesla A100 GPUs.

1. Introduction

Self-supervised learning (SSL) models have demonstrated state-of-the-art (SOTA) performance for speech processing tasks. However, the pre-training of modern SSL models usually leads to extremely big training cost, resulting in an exceedingly high barrier for the research of this area. Thus, this paper improves the pre-training efficiency from the aspect of the context encoder of SSL models.

The context encoder of the SOTA SSL models is usually a multi-headed self-attention (MHSA) Transformer or Conformer encoder. MHSA has a quadratic time and space complexity in the input sequence length, slowing down pre-training and increasing VRAM consumption. However, sub-quadratic complexity alternatives to MHSA usually lead to inferior results for speech processing tasks [1].

A recently-developed linear-complexity model, SummaryMixing [1], is promising for developing linear-complexity SSL models, since it is the first linear-complexity model that surpasses SOTA MHSA models for automatic speech recognition and spoken language understanding under supervised training. SummaryMixing has two branches: a local branch uses a point-wise feed-forward network to capture the local information, and a summary branch which uses the average vector of the input frames to capture the global information. The output of the two branches are merged to form the hidden representations of the input. In this paper, we equip wav2vec 2.0 using a Conformer context encoder with SummaryMixing. As shown in Table 1, compared to MHSA Conformer wav2vec 2.0 model, our proposed model gives better or equivalent results for all the downstream tasks of the MP3S benchmark [2]. For the efficiency aspect, SummaryMixing reduces the pre-training time and the peak VRAM by 18% and 23%, respectively, making the pre-training of a 155M wav2vec 2.0 model finished within 7 days with 4 Tesla A100 GPUs.

Table 1: *The speech recognition word error rates (WER), the accuracy (Acc.) for Intent Classification (IC), Emotion Recognition (ER), the equal error rate (EER) for Automatic Speaker Verification (ASV), and the pre-training GPU hours and peak VRAM of wav2vec 2.0 models*

Pre-train data Librilight 4.3k hours	Context Encoder	
Downstream Task	MHSA 165M	SummaryMixing 155M
LibriSpeech 100h d.c/t.c/t.o WER ↓	6.6/6.9/21.3	**6.2/6.6/17.2**
Welsh 15.8h dev/test WER ↓	48.8/50.8	**45.7/48.3**
Basque 11h dev/test WER ↓	43.6/44.0	**41.4/42.1**
SLURP IC ACC. ↑	78.1	**80.5**
IEMOCAP ER ACC. ↑	**64.8**	64.3
VoxCeleb1 ASV EER ↓	2.6	**2.4**
GPU Hours ↓	207	**164**
VRAM GB ↓	76	**54**

To the best of our knowledge, this paper is the first to present a linear-complexity SSL model with no performance drop on downstream tasks. We release the recipes of SummaryMixing wav2vec 2.0 with SpeechBrain [3] at `https://github.com/SamsungLabs/SummaryMixing`

2. References

[1] T. Parcollet, R. van Dalen, S. Zhang, and S. Bhattacharya, "SummaryMixing: A linear-complexity alternative to self-attention for speech recognition and understanding," *arXiv preprint arXiv:2307.07421*, 2024.

[2] S. Zaiem, Y. Kemiche, T. Parcollet, S. Essid, and M. Ravanelli, "Speech self-supervised representation benchmarking: Are we doing it right?" in *INTERSPEECH*, 2023.

[3] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

Linear Complexity Unified Streaming and Non-streaming Conformers for Transducer-based Speech Recognition

Titouan Parcollet, Rogier van Dalen, Shucong Zhang, Sourav Bhattacharya

Samsung AI Centre, Cambridge, United-Kingdom

{t.parcollet|r.vandalen|s1.zhang|sourav.b1}@samsung.com

Abstract

Automatic speech recognition (ASR) with an encoder equipped with self-attention, whether streaming or non-streaming, takes quadratic time in the length of the speech utterance. This slows down training and decoding, increase their cost, and limit the deployment of the ASR in constrained devices. SummaryMixing is a promising linear-time complexity alternative to self-attention for non-streaming speech recognition that, for the first time, preserves or outperforms the accuracy of self-attention models. Unfortunately, the original definition of SummaryMixing is not suited to streaming speech recognition. Hence, this work extends SummaryMixing to a Conformer Transducer that works in both a streaming and an offline mode. It shows that this new linear-time complexity speech encoder outperforms self-attention in both scenarios while requiring less compute and memory during training and decoding.

1. Introduction

Streaming and non-streaming automatic speech recognition (ASR) have followed the general trend in deep learning and increased steadily both in model size and architecture complexity. A few speech encoders have reached billions of neural parameters while improving the recognition accuracies. Such an increase in required computing and memory resources, however, conflicts with the surging demand for constrained-resources deployment of on-device ASR models. Streaming ASR is a great example as the perceived latency at decoding time is almost as important as the word error rate (WER).

Multi-head self-attention (MHSA) is one of the core components of most ASR systems. Their success lies in the ability to capture fine-grained global interactions between pairs of acoustic frames. In an offline, or non-streaming, ASR scenario, MHSA will be computed over the whole speech utterance. For streaming, the MHSA output of a given time step often depends only on all or a subset of the past frames. Unfortunately, MHSA complexity grows quadratically with the input sequence length.

Typical efficient changes to self-attention outside of the speech modality are low-rank approximation, linearization, or sparsification. For ASR more specifically, the Squeezeformer, the Efficient Conformer, and the Emformer aim to reduce the impact of the quadratic complexity by reducing the sequence length. Using these methods, the memory and training time are indeed reduced, but the quadratic time complexity remains. Instead, we aim to remove the quadratic time complexity. In that direction, the Fastformer proposed to linearise self-attention. It was successfully applied to non-streaming ASR exhibiting faster training times but at the cost of degraded accuracies.

More recently, a few articles reported that under certain conditions, pair-wise MHSA operations in trained ASR models including transformers and conformers, behave like simple feed-forward layers [1, 2]. This observation led to the development of SummaryMixing for ASR [3], an alternative to MHSA that takes linear time in the sequence length. In short, SummaryMixing summarises the whole sequence in a single vector that is concatenated back to each frame before further processing. This vector contains global relations, while each transformed frame models local interactions. Unfortunately, the definition of SummaryMixing is not compatible with streaming ASR.

Table 1: *Subset of speech recognition results from the full paper on the VoxPopuli dataset with a non-streaming and streaming conformer transducer trained with and without dynamic chunk training (DCT). The left context is infinite.*

Models	DCT	Non-streaming WER % Dev.	Non-streaming WER % Test	640ms WER % Dev.	640ms WER % Test
MHSA	—	11.5	11.7	—	—
Sum.Mix.	—	11.0	11.2	—	—
MHSA	✓	11.6	11.7	13.0	13.4
Sum.Mix.	✓	11.2	11.3	12.8	13.0

This article extends SummaryMixing to streaming ASR and evaluates its performance in a unified streaming and non-streaming training and decoding framework using dynamic chunk training on two ASR datasets. The behavior of SummaryMixing is then compared to MHSA with varying sequence lengths in terms of inference real-time factor, WERs, and memory consumption. Experiments conducted with a state-of-the-art streaming and non-streaming conformer Transducer show that SummaryMixing achieves equivalent or better WER than MHSA while reducing the training time by 15.5% as well as the memory footprint both during training and inference by 16%. We also release in the SpeechBrain toolkit a half-precision Transducer loss halving the memory required during training.

2. References

[1] S. Zhang, E. Loweimi, P. Bell, and S. Renals, "On the usefulness of self-attention for automatic speech recognition with transformers," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 89–96.

[2] K. Shim, J. Choi, and W. Sung, "Understanding the role of self attention for efficient speech recognition," in *International Conference on Learning Representations*, 2022.

[3] T. Parcollet, R. van Dalen, S. Zhang, and S. Bhattacharya, "Summarymixing: A linear-complexity alternative to self-attention for speech recognition and understanding," 2024.

Fine-tuning of Self-supervised Models Jointly with Autoencoders

Wenjie Peng, Thomas Hain

Speech and Hearing Research Group
Department of Computer Science, The University of Sheffield, United Kingdom
wpeng16@sheffield.ac.uk, t.hain@sheffield.ac.uk

Abstract

Recently, self-supervised learning (SSL) models have received great attention due to the promising results they have achieved on almost all downstream tasks. A common approach to make use of SSL models for downstream tasks is fine-tuning with task specific supervision. However, SSL models have been shown to be prone to over-fitting in this case, as typically the data size available for fine-tuning is much smaller than that in pre-training. To address this issue, this work makes use of the reconstruction loss in autoencoders. Specifically, MLP-based autoencoders were employed on top of wav2vec2.0 model, where speech representations from wav2vec2.0 are used as input to the autoencoders. The bottleneck output of the autoencoders is jointly used for an ASR task. Experimental results on AMI and BTS datasets showed that the performance can be further improved after combining CTC loss with the reconstruction loss, where the WER can be reduced from 14.28% to 13.97% and 24.45% to 24.17% respectively.

Index Terms: Self-supervised models, autoencoder, reconstruction loss

1. Method

Figure 1 illustrates the methods used in the current experiments of fine-tuning SSL models on ASR tasks. As shown in Figure 1(a), the baseline method adds a feedforward neural network on top of the output from the pre-trained SSL model, where only CTC loss is used for fine-tuning. In the proposed method as shown in Figure 1(b), an MLP-based autoencoder is adopted on top of the SSL model, where the self-supervised speech representations are sent to the encoder, and the bottleneck features from the autoencoder are used as input for the feedforward neural network for ASR task. In addition to CTC loss, an auxiliary reconstruction loss is also calculated, where MSE loss is adopted.

In the proposed method, the final loss is defined as follows:

$$L = L_{CTC} + \alpha L_{MSE} \quad (1)$$

where α is a hyper-parameter for the reconstruction loss.

2. Experiments

2.1. Data and implementation details

AMI [1] and British Telephone Speech (BTS) datasets are used in the current experiments, where AMI dataset consists of 100-hour meeting data while BTS dataset consists of 370 hours of noisy speech. Speechbrain toolkit is used for model training and evaluation. Wav2vec2.0 [2] pre-trained model with large size is adopted and only the transformer blocks are learnable during fine-tuning, which is common in fine-tuning SSL models. The autoencoders used in this experiment consist of an encoder and decoder, both of which are 3-layer MLP-based neural networks. The layer size for the encoder is 1024-1024-512-256, which is symmetric for the decoder.

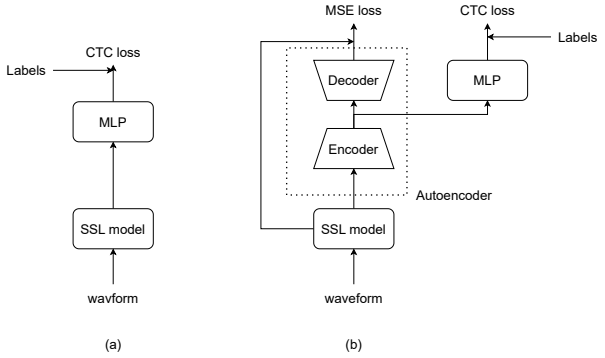


Figure 1: *Overview of the methods. (a) baseline method with CTC loss only. (b)the proposed method with CTC and MSE loss*

2.2. Results

The fine-tuning results are shown in Table 1. It can be seen from Table 1 that the WER can be further slightly reduced after introducing the MSE reconstruction loss. The α values for AMI and BTS datasets is 0.15 and 0.1 respectively.

Loss	AMI	BTS
L_{CTC}	14.28	24.45
$L_{CTC} + \alpha L_{MSE}$	13.97	24.17

Table 1: *WER of fine-tuning results on AMI and BTS dataset*

3. References

[1] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal *et al.*, "The ami meeting corpus: A pre-announcement," in *International workshop on machine learning for multimodal interaction*. Springer, 2005, pp. 28–39.

[2] 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.

Analysis of self-supervised speech representations for tone languages

Opeyemi Osakuade, Simon King

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

O.M.Osakuade@sms.ed.ac.uk, Simon.King@ed.ac.uk

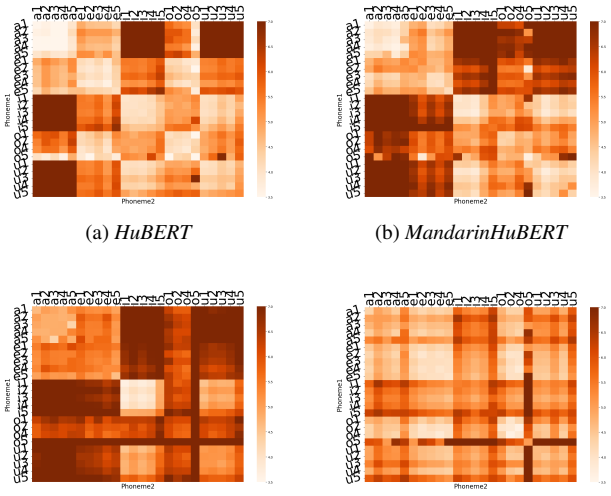
Self-supervised speech representations offer an effective means of learning from unlabelled speech data. This approach has been particularly effective for English, demonstrating that these representations capture the phonetic details required for recognising or generating speech. The representations of speech learned by such models (henceforth "SSL speech representations") have led to improvements in tasks including Automatic Speech Recognition [1] and speech-to-speech translation [2]. Typical approaches incorporate quantisation (via clustering) to convert continuous representations into discrete units. These discrete units are particularly useful for capturing linguistic information. Research [3] has shown a strong correlation between these discrete units and phonemes, and with broad phoneme classes. Further investigations [4] have revealed that these discrete units correspond to sub-phonetic events, including the detailed dynamics of speech, such as the distinct closure and release phases of plosives. However, tone languages – in which pitch variations convey meaning – remain relatively unexplored.

If we were only interested in *classifying* tone, we could have passed the vector-space representation from a selected layer of an SSL model into a simple classifier, trained in a supervised manner to perform tone classification. However, this would not answer our research question "Do *discrete* representations of speech from an SSL model capture tone distinctions as well as phonetic distinctions?"

We compared three state-of-the-art speech representations, wav2vec 2.0, HuBERT, and XLSR, for two tone languages, Mandarin and Yoruba. We followed a standard approach of embeddeding speech using an SSL model then clustering the resulting vector-space representations into symbols. This results in a string of symbols for each spoken utterance. For every pair of vowel phones in a corpus, we computed the edit distance between their symbol representations. A small edit distance indicates that the two SSL symbol sequences are very similar. If a pair of vowels with a small edit distance have the same phone label, but *different* tone labels, then we can say that the SSL representation has captured the phonetic distinction but failed to capture tone.

We visualise the average pairwise edit distances, broken down by vowel-with-tone, by vowel-without-tone, or only by tone, and these reveal the capability of SSL symbolic representations to capture phonetic vs. tonal distinctions. For example, in Figure 1a, we observe a clear block diagonal pattern, revealing HuBERT's ability to distinguish between vowel phones for Mandarin; but, in stark contrast, the edit distances *within* each block (e.g a1–a5) are all relatively small and there is no diagonal pattern: tone distinctions are not captured. These patterns are similar for wav2vec 2.0 in Figure 1c. Surprisingly, in Figures 1b and 1d, even the MandarinHuBERT and XLS-R models – which include data from tone languages in their training or fine-tuning – are not much better.

We conclude that symbolic SSL speech representations fail to capture tone, and hypothesise that this stems from the feature extraction employed when training the model.



(a) *HuBERT* (b) *MandarinHuBERT*

(c) *wav2vec 2.0* (d) *XLS-R*

Figure 1: *Pairwise edit distance matrices for Mandarin vowel-with-tone, in which letters indicate vowel and numbers indicate tone*

1. References

[1] B. Irvin, M. Stamenovic, M. Kegler, and L. C. Yang, "Self-supervised learning for speech enhancement through synthesis," in *Proc. ICASSP*, Rhodes Island, Greece, June. 2023, pp. 1–5.

[2] A. Lee, P.-J. Chen, C. Wang, J. Gu, S. Popuri, X. Ma, A. Polyak, Y. Adi, Q. He, Y. Tang *et al.*, "Direct speech-to-speech translation with discrete units," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022, pp. 3327–3339.

[3] A. Sicherman and Y. Adi, "Analysing discrete self-supervised speech representation for spoken language modeling," in *Proc. ICASSP*, Rhodes Island, Greece, June. 2023, pp. 1–5.

[4] D. Wells, H. Tang, and K. Richmond, "Phonetic analysis of self-supervised representations of english speech," in *Proc. Interspeech*, Incheon, Korea, Sep. 2022, pp. 3583–3587.

The UKIS 2024 Committee would like to thank ISCA and Interspeech 2023 organisers for kindly providing this template.

Investigation of Spanish-accented English Pronunciation Patterns in ASR

Margot Masson, Julie Carson-Berndsen

School of Computer Science, University College Dublin, Ireland
Science Foundation Ireland Centre for Research Training in Digitally Enhanced Reality (d-real)[1]
margot.masson@ucdconnect.ie, julie.berndsen@ucd.ie

1. Introduction

Previous studies [1] have highlighted the recurrence of specific error patterns across several speakers of the same accent, emphasizing the need for accent-specific evaluation and explainability in ASR systems. In our prior work [2], we leveraged phonemic differences between first language (L1) French and second language (L2) English to generate non-native artificial speech variants. Here, our focus shifts to investigating the Spanish accent in English, aiming to investigate how the information captured in synthesised speech compares with existing phonetic knowledge of Spanish-accented English.

Spanish phonology differs considerably from English phonology in several aspects [3], including phoneme sets and syllable structures (phonotactics). As a result of these differences, Spanish speakers might have difficulties to pronounce some English-but-not-Spanish phonemes; additionally, they might struggle with English consonant clusters and add a vowel to the beginning of English words with an initial /s/ consonant cluster, for example pronouncing "estreet" instead of "street".

In this paper, we propose to investigate the pronunciation patterns of Spanish-accented English with a focus on phonotactics. We use artificially accented speech to retrieve non-native phonemic similarity patterns of Spanish, before analysing the confusions patterns in light of phonotactic differences between Spanish and English, with a focus on the epenthesis of /e/ before initial consonant clusters beginning with /s/ (sC clusters).

2. Phonemic patterns

Similarly to what we did for French [2], we generate synthetic speech with two methods (*artificial voice* and *artificial phonemes*), and compare them with *natural* speech. In *artificial voice*, English utterances are produced by synthesising English with a text-to-speech (TTS) system configured to pronounce Spanish. This corresponds to the worst case scenario, where a Spanish native speaker encounters English for the first time. In *artificial phonemes*, we replace only the English-not-Spanish phonemes by similar ones in the sense of a similarity matrix built from the *artificial voice* confusions. Here, we are building on the generation bias to gain insight into the information captured by the TTS and whether it relates to expected Spanish patterns in English. Primary results, similar to those obtained for French, show that synthesised speech accurately captured the pronunciation patterns of Spanish, and that the ASR is sensitive to phoneme-level variations.

[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.

3. Phonotactic patterns

In order to investigate the /e/ epenthesis before initial sC clusters, we filter the sentences from the above mentioned *natural* and *artificial voice* speech data sets that contain sC clusters. In addition, similar to *artificial phonemes*, we produce an *artificial phonotactic* set, which consists of speech synthesised after manually adding a /e/ before sC clusters in the phoneme sequences. We run the speech and phoneme recognition systems on these sets and analyse the confusions obtained.

We analyse the ASR confusions, and specifically the confusions for the epenthesis itself and its close context (i.e. the last phoneme of the preceding word and the first phoneme of the word with sC cluster). This analysis reveals that, as expected, the epenthesis is present more often in the case of preceding consonant than preceding vowel. It also highlights that *artificial voice* and *artificial phonotactic* do not follow the same tendencies, suggesting that the /e/-to-speech from the text-to-speech system also depends on the system's modelling of Spanish pronunciation.

4. Future work

The analysis of the /e/ epenthesis phonotactic phenomenon reveals that this phenomena is only partly captured by artificial speech. We observe interesting trends across different /s/-preceding phonemes but the reasons behind them are still unclear. Thus, future work will focus on investigating further these contexts across different TTS and ASR, as well as speaker gender. In addition, we plan to human-evaluate the samples, and compare human versus ASR recognition of the epenthesis in the different contexts. This, along with the use of other TTSs, will give us an insight into the actual recognition rate of epenthesis, irrespective of the TTS. Also, other phonotactic and prosodic patterns are still to be investigated.

5. References

[1] E. O'Neill and J. Carson-Berndsen, "Investigating the sensitivity of automatic speech recognition systems to phonetic variation in l2 englishes," *University of Pennsylvania Working Papers in Linguistics*, vol. 29.2, pp. 109–118, 2023.

[2] M. Masson and J. Carson-Berndsen, "Investigating phoneme similarity with artificially accented speech," in *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics, 2023, pp. 49–57.

[3] J. C. Silva Valencia, "A comparative linguistic analysis of english and spanish phonological system," *GIST - Education and Learning Research Journal*, vol. 25, p. 140–156, 2022.

CognoSpeak: an automatic, remote assessment of early cognitive decline in real-world conversational speech

Madhurananda Pahar[1], *Fuxiang Tao*[1], *Bahman Mirheidari*[1], *Nathan Pevy*[1], *Rebecca Bright*[2], *Swapnil Gadgil*[2], *Lise Sproson* [3], *Dorota Braun*[4], *Caitlin Illingworth*[4], *Daniel Blackburn*[4], *Heidi Christensen*[1]

{[1]Department of Computer Science, [4]Sheffield Institute for Translational Neuroscience } University of Sheffield, Sheffield; [2]Therapy Box, London; [3]NIHR Devices for Dignity HTC, Sheffield

m.pahar@sheffield.ac.uk, f.tao@sheffield.ac.uk

Abstract

The early signs of cognitive decline are often noticeable in conversational speech and identifying those signs is crucial in dealing with later and more serious stages of neurodegenerative diseases. Clinical detection is costly and time-consuming and the existing automated systems lack a large database, rich metadata and demographic information. CognoSpeak asks memory-probing questions used in outpatient consultations using a virtual agent on a mobile or web platform and conducts cognitive tests such as short and long-term memory tests, verbal fluency tests, and picture descriptions and gathers multimodal data such as audio and video along with large metadata. The highest F1-score of 0.939 is achieved from a foundation model (RoBERTa) to discriminate dementia and mild cognitive impairment (MCI) from healthy subjects while using the audio of long-term memory tests from 126 subjects. CognoSpeak is an automatic, remote, low-cost, repeatable, non-invasive and less stressful alternative to existing clinical cognitive assessments.

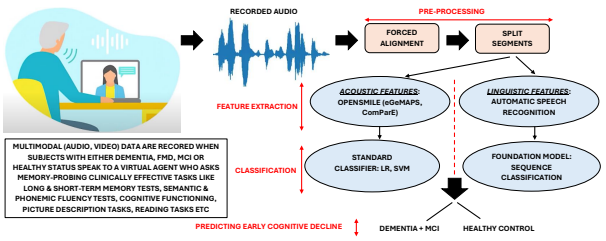


Figure 1: ***CognoSpeak*** *collects real-world audio/video recordings when a virtual agent prompts clinically proven effective tasks. The subjects' cognitive status was grouped into three categories which are dementia, MCI, and healthy. The audio recordings of 126 subjects were pre-processed by forced alignment and segmentation before acoustic and linguistic features were extracted and used by standard and foundation model classifiers respectively.*

Figure 1 explains the data collection system. It is ongoing, and to date, over 1000 participants have completed the assessment. Of these, manual transcripts are so far available for 126 individuals, and these have subsequently been selected for the experiments conducted in this work (Table 1). We have focused on only four sets of tasks, namely the short-term memory test, the long-term memory test, the semantic fluency test and the cookie theft picture description test and preprocessed the audio by forced alignment and segmentation before extracting both acoustic and linguistic features, as shown in Figure 1.

Figure 2 shows RoBERTa has outperformed RoBERTa and DistilBERT by a narrow margin by achieving the highest F_1-score of 0.939 while using the long-term memory test whereas

Table 1: *Demographic information of the participants.*

Diagnosis	Frequency	Age	Female	Male
Dementia	8	65.4±18.5	2	6
MCI	39	65.3±11.4	22	17
HC	79	50.7±17.4	39	39
Total	126	56.2±17.3	63	62

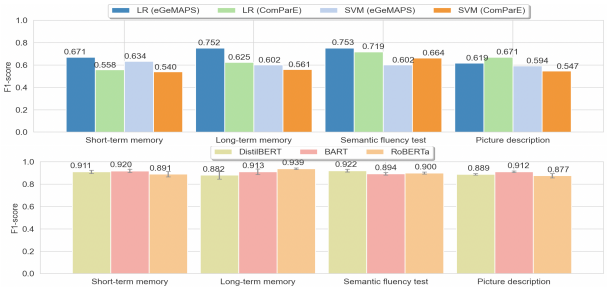


Figure 2: *Classification based on individual question type*

short-term memory test, semantic fluency test and picture description task also demonstrate promising results.

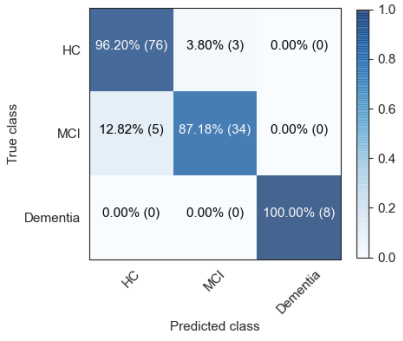


Figure 3: *Confusion matrix generated by RoBERTa model.*

CognoSpeak will be an ideal tool for clinical trials as it can be quick, done from home, and can track mental health issues like mood, anxiety and cognition, thus it provides a low-cost, repeatable, non-invasive and less stressful alternative to the current cognitive assessment methods to detect early signs of cognitive decline. As for our future work, rigorous classifier training and fine-tuning will be carried out to improve the classification performance, as well as for non-native English speakers. The performance of CognoSpeak as an automatic remote long-term monitoring tool will also be investigated using the follow-up patients alongside their MoCA, MCE, RUDAS, PHQ-9 and GAD-7 scores. Finally, the multimodal data along with the rich metadata will be released to the community soon.

video-SALMONN: Speech-Enhanced Audio-Visual Large Language Models

Guangzhi Sun[1]

[1]University of Cambridge

gs534@cam.ac.uk

1. Introduction

Speech understanding as an element of the more generic video understanding using audio-visual large language models (av-LLMs) is a crucial yet understudied aspect. This paper proposes video-SALMONN, a single end-to-end av-LLM for video processing, which can understand not only visual frame sequences, audio events and music, but speech as well. To obtain fine-grained temporal information required by speech understanding, while keeping efficient for other video elements, this paper proposes a novel multi-resolution causal Q-Former (MRC Q-Former) structure to connect pre-trained audio-visual encoders and the backbone large language model. Moreover, dedicated training approaches including the diversity loss and the unpaired audio-visual mixed training scheme are proposed to avoid frames or modality dominance. On the introduced speech-audio-visual evaluation benchmark, video-SALMONN achieves more than 25% absolute accuracy improvements on the video-QA task and over 30% absolute accuracy improvements on audio-visual QA tasks with human speech. In addition, video-SALMONN demonstrates remarkable video comprehension and reasoning abilities on tasks that are unprecedented by other av-LLMs.

2. video-SALMONN Architecture

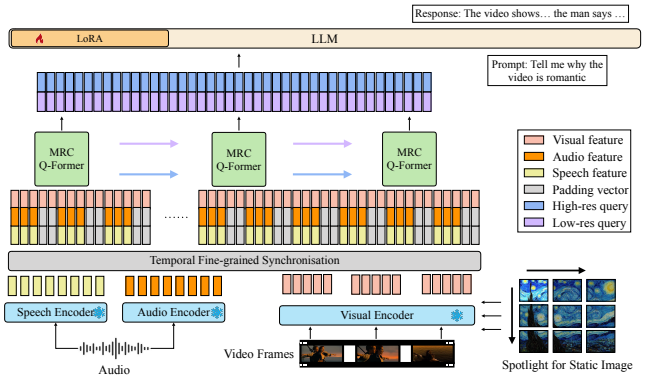


Figure 1: *The model architecture of video-SALMONN.*

As shown in Fig. 1, key components include the synchronisation module and the MRC Q-Former. First, visual (image or video), speech and non-speech audio are encoded using corresponding pre-trained encoders. The visual encoder converts the input image into a certain number of vectors via the image encoder from InstructBLIP. When video input is given, the visual encoder encodes each video frame separately as a sequence of images at a 2 Hz frame rate, and the output image features are concatenated along the temporal dimension to form a sequence of visual frames. Following SALMONN, Whisper encoder and BEATs encoder are adopted to encode speech and non-speech audio respectively from the same audio stream at a 50 Hz spectrogram frame rate.

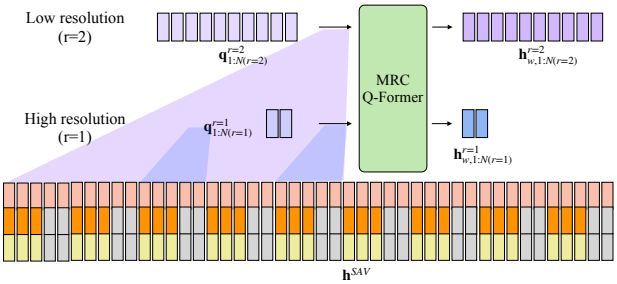


Figure 2: *Illustration of the MRC Q-Former structure with two levels of temporal resolutions. The high-resolution sliding window covers $k = 5$ input features with two query vectors and the low-resolution Q-Former covers $k = 25$ with 10 query vectors.*

The video-SALMONN system is trained with cross-entropy loss and diversity loss as shown below.

$$\mathcal{L}_{\text{diverse}} = \sum_{r=2}^{R} \sum_{w=1}^{W(r)} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \text{sim}(\mathbf{h}_{w,i}^{(r)}, \mathbf{h}_{w,j}^{(r)}) \quad (1)$$

3. Results

Table 1: *The audio benchmark and visual benchmark results.*

Systems	ASR ↓	Video QA ↑	Image Cap. ↑
Whisper large-v2	2.9%	-	-
InstructBLIP 13B	-	21.0%	84.5
Video-LLaMA 7B	100.0%	22.5%	22.0
video-SALMONN 13B (ours)	**2.6%**	**49.6%**	**89.6%**

Table 2: *Audio-visual benchmark results.*

Systems	AVSR ↓	AVQA ↑	AVSSD ↑
Whisper large-v2	8.3%	-	-
InstructBLIP 13B	-	-	1.1%
Video-LLaMA 7B	-	21.3%	41.9%
video-SALMONN 13B (ours)	**7.7%**	**70.5%**	**47.6%**

Using Structured Conversational Prompts in the Diagnosis of Dementia

Fritz Peters[1], *Heidi Christensen*[1]

[1]Department of Computer Science, University of Sheffield, United Kingdom

fpeters3@sheffield.ac.uk, heidi.christensen@sheffield.ac.uk

Advances in speech and language technologies have brought us closer to healthcare applications that provide a truly meaningful benefit for individuals with various health conditions. Existing applications have focused on either translating standard assessments into an automated tool, or on using conversation as a window into a person's cognitive health. Holding a conversation draws onto different cognitive processes associated with attention, memory, and executive functioning. In addition, it reveals a person's ability to produce and understand speech/language. Abnormalities in these aspects of a person's cognition can be elicited by conversational prompts and have been proven to be a useful feature in the automatic diagnosis of cognitive conditions such as dementia [1].

Dementia

The term dementia describes a progressive decline in a person's memory and cognitive functioning, which is severe enough to interfere with their everyday life. It is not a disease but rather a collection of symptoms that can be caused by various underlying diseases (e.g., Alzheimer's Disease and Frontotemporal Dementia). Traditionally, the routine assessment for dementia includes an initial consultation with a GP during which they take the patient's history. This conversation-like interaction allows the clinician to ask specific questions, interpret the patient's responses, and through that to start to get an impression of their cognitive condition. To further assess the patient's functioning, clinicians usually administer paper-based cognitive assessments such as the General Practitioner Assessment of Cognition (GPCOG) [2].

Existing automated screening tools

Previous work in the automatic assessment of dementia has often focused on translating routine assessments into a computer-based tool. For example, Chien et al. [3] created a set of online tests including three different types of standard tasks: fluency, picture description, and logical memory. Similar tasks were used by O'Malley et al. [4] which enabled them to identify patients with Mild Cognitive Impairment (MCI) with a sensitivity of 80%. These tasks provide some surface-level insight into different aspects of a patient's cognitive functioning. However, they fail to provide more condition-specific information, such as patient history, which could be assessed through, for example, a semi-structured conversation. Automated approaches that focus on patient conversations face the challenge of recreating a meaningful patient-clinician interaction. This includes asking the patient clarifying questions and prompting them to further elaborate on potentially meaningful topics.

Proposed work

Our proposed work aims to develop a set of conversational prompts that reliably elicits a patient's speech containing dementia-specific speech and language markers for a more robust automatic detection of the condition. We aim to address the following two shortcomings of existing automatic assessment frameworks:

1. Limited number of standard tasks used
2. Disrupted conversational flow

(1) Rather than simply translating existing cognitive assessments, we are planning to introduce a set of conversational prompts that target different aspects of a patient's cognitive functioning. To do so, we will scope out existing speech- and language-based prompts across a number of conditions and existing assessments. This will allow us to identify any links between different types of prompts and cognitive domains. Based on these links, we will propose and evaluate a set of new conversational prompts. One possible direction of this work is to utilise recent advances in the field of Artificial Intelligence (AI) by exploring different options of including generative AI in the proposed prompts.

(2) To fully harness the power of conversation as a window into a patient's cognitive health, we aim to approximate a more realistic flow of conversation. This will include the provision of additional prompts when further clarification or elaboration by the patient is needed. For example, during the description of a significant memory, we might want to ask the patient to give a more detailed description of certain aspects.

Overall, the aim of this work is to generate a set of meaningful conversational prompts that can be used in the automatic assessment of dementia.

1. References

[1] M. Shi, G. Cheung, and S. R. Shahamiri, "Speech and language processing with deep learning for dementia diagnosis: A systematic review," *Psychiatry Research*, vol. 329, p. 115538, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165178123004882

[2] H. Brodaty, D. Pond, N. M. Kemp, G. Luscombe, L. Harding, K. Berman, and F. A. Huppert, "The gpcog: A new screening test for dementia designed for general practice," *Journal of the American Geriatrics Society (JAGS)*, vol. 50, no. 3, pp. 530–534, 2002.

[3] Y. W. Chien, S. Y. Hong, W. T. Cheah, L. H. Yao, Y. L. Chang, and L. C. Fu, "An automatic assessment system for alzheimer's disease based on speech using feature sequence generator and recurrent neural network," *Scientific reports*, vol. 9, no. 1, pp. 19 597–10, 2019.

[4] R. P. D. O'Malley, B. Mirheidari, K. Harkness, M. Reuber, A. Venneri, T. Walker, H. Christensen, and D. Blackburn, "Fully automated cognitive screening tool based on assessment of speech and language," *Journal of neurology, neurosurgery and psychiatry*, vol. 92, no. 1, pp. 12–15, 2021.

Crossmodal ASR Error Correction with Discrete Speech Units

Yuanchao Li, Pinzhen Chen, Peter Bell, Catherine Lai

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

yuanchao.li@ed.ac.uk

Abstract

ASR remains unsatisfactory in scenarios where the speaking style diverges from that used to train ASR systems, resulting in erroneous transcripts. To address this, ASR Error Correction (AEC), a post-ASR processing approach, is required. In this work, we tackle an understudied issue: the LOW-RESOURCE OUT-OF-DOMAIN (LROOD) problem, by investigating crossmodal AEC on very limited downstream data with 1-best hypothesis transcription. We explore pre-training and fine-tuning strategies and uncover an ASR domain discrepancy phenomenon, shedding light on appropriate training schemes for LROOD data. Moreover, we propose the incorporation of discrete speech units to align with and enhance the word embeddings for improving AEC quality. Results from multiple corpora and several evaluation metrics demonstrate the feasibility and efficacy of our proposed AEC approach on LROOD data as well as its generalizability and superiority on large-scale data. Finally, a study on speech emotion recognition confirms that our model produces ASR error-robust transcripts suitable for downstream applications.

1. Introduction

In some scenarios of ASR applications, the speech data does not always have much in common with the data used to train the ASR systems, resulting in an Out-of-Domain (OOD) problem. For example, the acoustic characteristics of emotional and pathological speech could be very different from those of audio books, as they may contain irregular, disfluent, and ambiguous patterns, which are unexpected in read speech [1, 2]. As a result, when ASR is applied to these downstream tasks, the transcripts produced may be of poor quality and difficult to use.

In this report, we propose the use of an S2S model with self-supervised AWEs to address the challenge of low-resource OOD speech in AEC. Our results demonstrate the necessity of pre-training and fine-tuning, as well as discover the existence of the ASR domain discrepancy problem. We compare different acoustic features and investigate the contextual distribution of lexical and acoustic word embeddings. Several comparison experiments confirm the effectiveness of our approach and pave the way for future research on AEC for low-resource OOD speech. Moreover, a downstream task of SER further demonstrates the improved quality of the corrected transcript, showing the applicability of our approach.

2. Architecture and Results

The proposed architecture is illustrated in Figure 1. We utilize the Montreal Forced Aligner to delineate word boundaries and pre-generate AWEs. These AWEs are then combined with word embeddings from input word tokens and aligned via a cross-attention mechanism to yield acoustic-enhanced word embeddings. These enhanced embeddings are fed into the Transformer decoder to generate corrected word tokens. We explore the impact of Pre-Training (PT) and Fine-Tuning (FT) on OOD data, as well as their combination with AWEs. The summarized results are presented in Table 1, indicating their respective effectiveness. Utilizing the corrected transcripts, we conduct speech emotion recognition, achieving a +0.90 improvement with a -1.11 reduction in WER (Table 2). This represents a more significant improvement compared to the existing literature, which achieved a +1.70 accuracy improvement with a -5.00 reduction in WER.

More details and results will be presented in the poster.

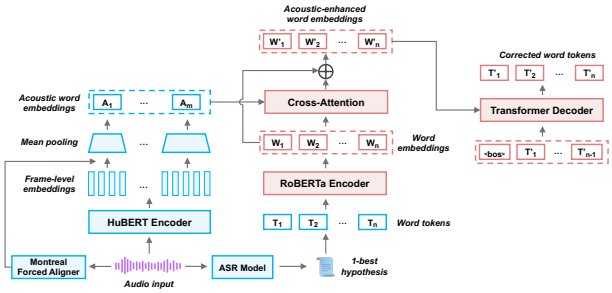


Figure 1: *Architecture of our crossmodal AEC.*

Table 1: *Result summary of our proposed approach.*

Model	WER↓	BLEU↑	GLEU↑
Original ASR transcript	17.18	76.56	75.29
PT	17.14	76.61	75.34
PT+FT	16.40	78.00	76.58
PT+FT+AWEs	**16.07**	**78.22**	**76.96**

Table 2: *Comparison results of SER performance.*

Transcript	WER↓	BLEU↑	GLEU↑	UA↑
Original	17.18	76.56	75.29	60.92
Corrected	16.07	78.22	76.96	**61.82**

3. References

[1] Y. Li, Z. Zhao, O. Klejch, P. Bell, and C. Lai, "ASR and emotional speech: A word-level investigation of the mutual impact of speech and emotion recognition," in *Interspeech 2023*, 2023.

[2] A. Potamianos, S. Narayanan, and S. Lee, "Automatic speech recognition for children," in *Fifth European Conference on Speech Communication and Technology*, 1997.

Exploring Acoustic Features for Challenging Voice Anonymisation Speakers

Henry Card and Jennifer Williams

School of Electronics and Computer Science, University of Southampton, UK

hc3g21@soton.ac.uk, j.williams@soton.ac.uk

Abstract

Voice privacy is crucial for protecting personal information and ensuring confidentiality. Both voice recognition and voice anonymisation are vital processes for voice privacy systems. Ensuring these are effective is a key challenge within this field. This project explores the features that cause certain speakers to be challenging to anonymise, in order to improve both voice anonymisation and recognition systems. Composed of three experiments on the Voice Privacy Challenge 2020 dataset, this project classifies speakers into challenging and non-challenging groups, extracts acoustic features from each utterance, and trains a series of machine learning models to explore which features are the most impactful when determining the challenging speakers. The models achieved up to 90.1% F1-score on cross-validation, and up to 66.6% F1-score on a held-out speaker set. From this classifier performance, we identified a set of features corresponding to what makes a speaker challenging or not. Our features were found to be statistically significant, and results were confirmed using human listeners. These features can be used in future work to optimise voice privacy systems.

1. Motivation

This project explores extracting acoustic features from a range of speaker voices in the anonymised VCTK dataset to discover the underlying factors that contribute to the difficulties of recognising a speaker's identity. A previous study has shown it is possible to identify challenging speakers for voice anonymisation [1], however, the factors that distinguish these speakers are still unknown. Uncovering these features will enhance reliability and effectiveness of both speaker recognition and anonymisation systems, thus, refining their real-world utilisation. Furthermore, this project uses voice privacy data rather than conventional speaker recognition datasets, allowing traversal of the topic of concealment of a speaker's vocal characteristics, a key factor of voice privacy systems [2].

2. Approach

We extracted speaker x-vector embeddings for each utterance in original and anonymised speech, and used paired enrolment/trial testing to identify speakers who were difficult to recognise, grouping them into 2 categories: *challenging* and *non-challenging*. Next, we extracted 6,373 functional OpenSMILE features using ComParE 2016 and performed PCA to identify key features (PCA=86 for anonymised speech, PCA=88 for original speech). We then trained five classifiers (Nearest Centroid, Gaussian Naive Bayes, Decision Trees, Linear SVM, and Radial-Basis Function SVM) using this PCA feature set to predict challenging and non-challenging speaker classes.

Using the best-performing classifiers, we examined SHAP values [3] for key features to see which features contributed most to the classification decision between challenging and non-challenging speakers. We also explored KL-divergence (in both directions, as KL is mathematically asymmetrical) to characterise how different values of the key features correspond to challenging versus non-challenging speakers. In our listening test, we asked 23 human listeners from Qualtrics/Prolific two types of questions probing whether speakers were difficult to understand alongside a word presence task, and used this to further verify our results from experiments above in the context of voice anonymisation utility.

3. Results and Discussion

We have introduced an approach for finding features that correlate to challenging and non-challenging speakers for voice anonymisation. We were able to identify several key features that correlated with challenging speakers across our testing conditions, across speakers, and across machine learning models.

In future work, we seek to better understand the interplay between speaker voice characteristics and voice anonymisation system performance. By altering these features to increase the difficulty of identification, these findings could vastly improve current systems used for anonymisation of speaker voices, as well as provide further understanding into possible challenges within speaker identification software. The results obtained from this project act as a foundation, enabling the possibility to develop better speaker identification and anonymisation systems, by applying these discoveries to understand why speakers are hard to identify. Our work may also translate across speech technology disciplines and have direct application to speech synthesis, speech recognition, and others.

This work was supported by the UK EPSRC through the Trustworthy Autonomous Systems (TAS) Hub (EP/V00784X/1)

4. References

[1] J. Williams, K. Pizzi, N. Tomashenko, and S. Das, "Anonymizing speaker voices: Easy to imitate, difficult to recognize?" in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12 491–12 495.

[2] A. Nautsch, A. Jiménez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delgado, M. Todisco, M. A. Hmani, A. Mtibaa *et al.*, "Preserving privacy in speaker and speech characterisation," *Computer Speech & Language*, vol. 58, pp. 441–480, 2019.

[3] D. Fryer, I. Strümke, and H. Nguyen, "Shapley values for feature selection: The good, the bad, and the axioms," *IEEE Access*, vol. 9, pp. 144 352–144 360, 2021.

English for Academic Purposes (EAP) Tutors' Knowledge, Beliefs and Practices in Relation to the Use of Artificial Intelligence (AI)

Zoe Handley

[1]The University of York
zoe.handley@york.ac.uk

Abstract

Since its launch in November 2022, ChatGPT and other generative AI tools have attracted a lot of attention from the university sector. Capable of automatically generating text, generative AI tools could be used by students to produce written academic assignments without engaging with the learning materials. The debate in the university sector has therefore tended to focus on whether students should be allowed to complete their assignments with the assistance of AI. Many institutions currently prohibit the use of AI and are encouraging tutors to AI-proof assessments. Generative AI tools can, however, also be used to provide a range of other AI-powered writing support, including spelling and grammar checking, Automated Writing Evaluation (AWE), and paraphrasing functions, some of which have long been available in word processors and learners are actively encouraged to use. Moreover, some claim that using generative AI can raise students' awareness of the features of different genres among other writing conventions, and that being able to effectively use generative AI will be important to students' future employability. Other institutions are therefore explicitly raising students' awareness of their strengths and limitations and supporting students to develop strategies for working with them. This paper reports the preliminary results of a survey of, and follow-up interviews with, English for Academic Purposes (EAP) tutors in the UK. Exploring tutors' understanding of AI, their attitudes towards its' use in teaching and assessment, as well as how they are using it to support their teaching, the results of this study have the potential to help shape evolving policies on the use of AI in Higher Education (HE).

Index Terms: Teacher knowledge, language learning, generative AI, Automated Writing Evaluation (AWE), Machine Translation (MT)

1. Method

This paper reports preliminary reflections on the findings of interviews with University English for Academic Purposes (EAP) tutors in the UK about the impact of AI on their profession. Interviews explored tutors' understanding of AI as well as their insights into its impact on their profession, including the impact of AI on learning as well as assessment. The data were explored using thematic coding.

2. Participants

Eleven EAP tutors participated in the interviews, nine men and two women. Their knowledge of (generative) AI varied from an awareness of some of the tools available to an in-depth understanding of the technology, including experience developing mega prompts. Most were curious about AI and had experimented with a range of different tools. Those who had experimented were actively using AI tools with their students and/or actively involved in policy development concerning the use of AI for assessment at their institutions.

3. Results and discussion

My research tends to explore how technology, including wikis, can be harnessed to facilitate language learning. Reflecting use of generative AI in wider society, I believe students ought to be taught strategies for using these tools to support the writing process once they have developed adequate mastery of the language, a position many of the participants appeared to share.

Policy on the use of AI for assessments, however, naturally drove much of the discussion. Tutors often noted that their institution did not have a policy or, if there was a policy, that it was vague, with some wanting guidance on which specific tools are (not) permitted. Others questioned the value of such policies because they believed they were unenforceable.

Another driver was students' observed use of AI, with some tutors reporting students openly using AI tools during lectures and seminars to translate learning materials from English to their first language and vice versa. This raised questions about what it means to study in the UK and what future employers' expectations might be of a graduate. While some tutors felt that future employers would assume a certain level of English, others believed students should primarily be assessed on their ideas and thinking, not the accuracy of their language.

Relatedly, reflecting universities' emerging policies on the use of AI in assessments, tutors reflected on what it means to author a text and the extent to which they felt it was acceptable to use AI at different stages in the writing process. It was generally agreed that it was permissible to use AI to support brainstorming and proofreading. Some, however, advocated for greater integration of AI tools throughout the process.

Beyond generative AI, several tutors reflected on the potential for AI to perform the role of a personal tutor to, for example, support engagement with readings and provide interactive speaking practice, among other uses.

4. Conclusions

Tutors need time to explore the capacities of AI and reflect on the fundamental questions it raises about the purpose of studying for a degree in the UK. The examples elicited from tutors in this study will provide a useful stimulus for reflection for working groups as well as for use in training workshops.

Authentic Speaker Recognition

Qiang Huang

University of Sunderland

qiang.huang@sunderland.ac.uk

Abstract

Authentic speaker recognition aims to recognise source speakers from the voices converted by using the acoustic characteristics from target speakers. In this work we attempt to explore the feasibility of recognising authentic speakers from converted voices. To improve the robustness against converted voices, our recognition model is constructed by using multiple locally aggregated descriptors (VLAD) in ResNet architecture. The obtained results show promising performances in recognising authentic speakers from converted voices.

1. Introduction

During last ten years, the use of deep neural networks has significantly boosted speech synthesis technologies and enables voice conversion to be used in some fields, such as entertainment industry and healthcare. However, there exist potential risks that converted voices could be used for deceit, which enables people to believe the fake information conveyed by the converted voices that people could be familiar with. Therefore, preventing the incorrect use of one's voice with voice conversion technologies becomes more and more important.

To tackle some potential issues caused by fake voices, there have been some studies in detecting synthesized voices in order to distinguish real human voices from synthesized ones. These studies mainly rely on the use of deep neural networks to build a binary classifier to identify whether input voices are synthesized or not. Our work aims to go deeper to find who is the authentic source speaker given converted voices as it is assumed that the information from source speakers cannot be completely eliminated in the converted voices.

We followed the work in [1] to build a baseline speaker recognition system, where a dictionary-based NetVlad embedded within a ResNet trunk architecture was used to process variable-length of transformed utterances by aggregating features across time. Figure 1(a) shows its architecture including encoder and classifier. To effectively learn source speaker's features from converted voices, we propose a hierarchical structure, as shown in Figure 1(b), by expanding the last convolution block in the baseline system. In the proposed structure, the output of each sub-block will be connected to a VLAD layer, and then goes to a shared FC layer.

2. Data and Result

To conduct voice conversion, paired source and target utterances were selected from CSTR VCTK [2]. This Corpus includes speech data from 108 English speakers (excluding p280 and p315). In our experiments, 100 utterances from each speaker were randomly selected and used as source utterances. Their corresponding target utterances were also randomly selected from any specific speaker except the source speaker. FragmentVC[3] was used to implement voice conversion. We finally had 10800 (100×108) converted utterances in total, 9000 utterances for training and 1800 utterances for testing.

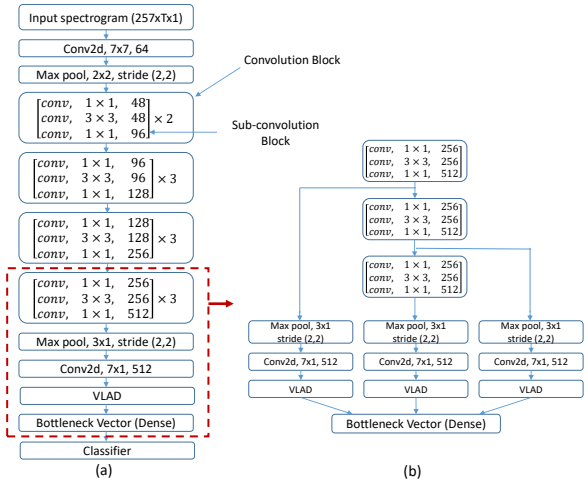


Figure 1: *Hierarchical VLAD for authentic speaker recognition.*

Table 1: *Comparison of recognition performances*

Model	Accuracy on test (%)
Baseline1	13.50 ± 0.28
Baseline2	14.16 ± 0.27
HVLAD	15.38 ± 0.25

Baseline1: uses Fig 1(a), but replaces the VLAD layer with a Flatten layer

Baseline2: uses Fig 1(a) only

HVLAD: replaces the last block in Fig 1(a) with Fig 1(b)

Table 1 shows comparisons of recognition accuracy obtained on the test data by using two baseline methods and our proposed approach. The results show that the use of our proposed approach can yield higher accuracy than the two baseline methods

3. References

[1] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *ICASSP*, May 2019, pp. 5791–5795.

[2] J. Yamagishi, C. Veaux, and K. MacDonald, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," University of Edinburgh. The Centre for Speech Technology Research (CSTR).

[3] Y. Y. Lin, C.-M. Chien, J. hao Lin, H. yi Lee, and L.-S. Lee, "Fragmentvc: Any-to-any voice conversion by end-to-end extracting and fusing fine-grained voice fragments with attention," in *ICASSP*, 2021, pp. 5939–5943.

Learn and Don't Forget: Adding a New Language to ASR Foundation Models

Mengjie Qian, Siyuan Tang, Rao Ma, Kate M. Knill, Mark J. F. Gales

ALTA Institute, Machine Intelligence Lab, Department of Engineering, Cambridge University, UK

{mq227,st941,rm2114,kmk1001,mjfg100}@cam.ac.uk

Foundation Automatic Speech Recognition (ASR) models often support many languages, e.g. 100 languages in Whisper [1]. However, there has been limited work on integrating an additional, typically low-resource, language, while maintaining performance on the original language set. Commonly used fine-tuning (FT) method, while simple, may degrade performance on previously supported languages – a problem known as catastrophic forgetting. We compare three approaches that exploit adaptation parameters: Soft Language Code Tuning (SLCT), which trains only the language code; Soft Prompt Tuning (SPT), which trains prepended tokens; and Low Rank Adaptation (LoRA) where a small set of additional parameters are optimised [2]. Each method aims to incorporate new language features with minimal interference to existing model parameters. Figure 1 illustrates the decoding process for these methods. Furthermore, we introduce Elastic Weight Consolidation (EWC) [3] to the fine-tuning process, this approach offers an alternative compromise with the potential to maintain performance in specific target languages.

To demonstrate the effectiveness of the methods, we experimented on the Whisper large-v3 model and the FLEURS benchmark dataset, which comprises 102 languages, of which 19 languages are not supported by the Whisper large-v3 model. After initial experiments, we focused on three languages (Irish, Asturian, and Sorani Kurdish) for further investigation, selected for their diverse range of zero-shot ASR performance in terms of Character Error Rates (CER). Table 4 presents a comparison of different fine-tuning approaches across three target languages. Baseline performance is obtained using the same Whisper large-v3 model for all three languages, while other models (FT, SLCT, SPT, and LoRA) are specifically trained for each language using the respective FLEURS training sets. Experiments indicate that while FT achieves the best performance improvement, it also leads to significant forgetting in existing languages. SLCT, which only adjusts 0.00008% of the model's 1550M parameters, shows noticeable performance gains compared to the baseline on all of the three target languages. SPT, which tunes 0.001% of the model parameters, further enhances the performance. LoRA, tuning 0.13% of model parameters, demonstrates improved performance proportionate to the increased size of trained parameters.

To address the challenge of catastrophic forgetting, we adopt Fisher overlap as an analytic tool for assessing potential forgetting on different languages. Using EWC with Fisher parameters, we can constrain forgetting and preserve performance on existing languages. Our experiments also reveal that it's hard to learn a new task without forgetting a previous task if two tasks have a high Fisher overlap.

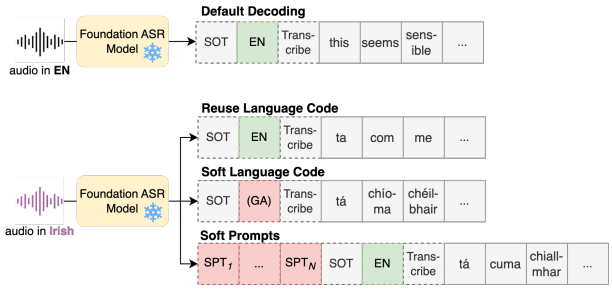


Figure 1: *Illustration of the decoding process for the proposed methods on an existing language and a new language. Three methods are demonstrated for a new language: (a) reusing an existing language code, (2) employing a trained soft language code, or (c) using prepended soft prompts to capture new domain knowledge. Reference text for Irish audio: "tá cuma chiallmhar air seo".*

Table 1: *ASR performance (CER: Character Error Rates, WER: Word Error Rates) on FLEURS Irish, Asturian, and Sorani Kurdish test sets with different parameter-tuning approaches.*

Model	Irish		Asturian		Sorani Kurdish	
	CER	WER	CER	WER	CER	WER
Baseline	78.8	110.4	13.5	47.5	38.6	110.8
FT	22.3	43.7	3.7	11.4	7.3	32.9
SLCT	41.5	81.4	7.4	29.8	35.8	95.2
SPT	38.1	74.9	5.1	19.4	13.6	61.4
LoRA	28.3	54.8	4.1	14.0	8.7	41.3

1. References

[1] A. Radford, J. W. Kim, T. Xu *et al.*, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[2] E. J. Hu, P. Wallis, Z. Allen-Zhu *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," in *International Conference on Learning Representations (ICLR)*, 2021.

[3] J. Kirkpatrick, R. Pascanu, N. Rabinowitz *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

This paper reports on research supported by EPSRC Project EP/V006223/1 (Multimodal Video Search by Examples) and Cambridge University Press & Assessment, a department of The Chancellor, Masters, and Scholars of the University of Cambridge.

Phonation based control system using laryngeal bio-impedance and machine learning.

Eugenio Donati[1], *Christos Chousidis*[2]

[1]University of West London
[2]University of Surrey

eugenio.donati@uwl.ac.uk, c.chousidis@surrey.ac.uk

The analysis and processing of human voice information plays a pivotal role in numerous applications ranging across several subject areas. Examples can be found in medicine, assistive technologies, music technology, education, and general human-computer interaction. This research approaches the extraction of voice information using laryngeal bioimpedance to efficiently convert singing voice into MIDI messages, whilst discarding non-singing phonation. The acquisition of human voice information is inherently tied to sound recordings and microphones delivering a complex signal that can result computationally taxing to analyse and susceptible to both noise and room acoustics [1]. Compared to microphones, laryngeal bioimpedance delivers a much simpler signal that can result in faster and more efficient processing, as well as bypassing acoustics interferences. The first stage of this research saw the design and construction of a dedicated system for the evaluation of vocal folds movement. This is based on the measurement of the variability of bioelectrical impedance across the vocal folds during phonation cycles achieved by means of pair of electrodes placed across the neck of a user. For improved usability and applicability, the system was then fitted with a novel self-calibrating sensor system which independently evaluates the optimal electrode placement. In the second stage, the measurement system was employed for the creation of a dataset dedicated to the distinction of speech and singing acts. The dataset comprises 7200 bioimpedance recordings with a 50:50 ratio of speech and singing recorded from 12 participants. A cepstral analysis of the data was then performed through MFCCs to extract representative features of speech and singing which were then used to train an artificial neural network for the binary classification of singing and speaking voices. The proposed system was designed for the purpose of acting as a control system to limit the conversion of voice into MIDI only for singing voices. The classifier delivered an accuracy of 94% (fig. 1) with a latency of about 200ms between predictions [2].

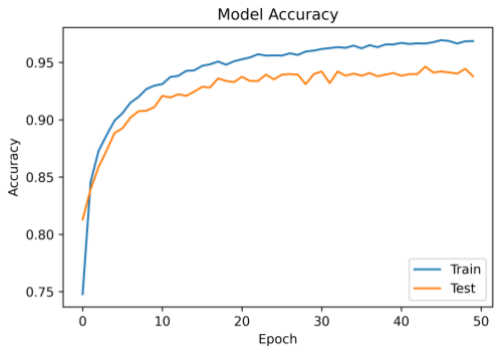


Figure 1: *Model accuracy in training and testing.*

In the third stage, a real-time conversion system was implemented for voice control of hardware and software electronic systems through the MIDI protocol. The use of laryngeal bioimpedance allowed for the pitch estimation to be based on Zero Crossing Rate, which in turn delivered a conversion system with a maximum latency of 12ms in between messages. In addition, MIDI pitch bend messages were implemented to ensure the ability of the system to emulate musical expression such as legato (fig. 2).

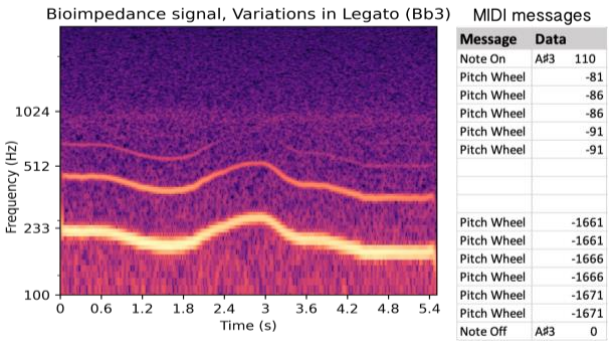


Figure 2: *MIDI conversion system performance.*

In the final stage, the measurement device, the classification network, and the conversion system were integrated. This resulted in a self-calibrating device that can distinguish speech and singing and isolate the latter for a fast and precise conversion of singing information into MIDI through laryngeal bioimpedance measurements. Whether the classifier latency can be considered real-time is significantly tied to its application. For voice-to-MIDI conversion, however, this latency remains too high for the system to be considered real-time suggesting further research is needed for the optimisation of the classifier. Overall, the discussed research delivered an improved sensor system for the measurement of laryngeal bioimpedance as well as a novel approach for the conversion of voice into MIDI, and contributes to the existing knowledge in the distinction of speaking and singing voices.

1. References

[1] Donati, E. and Chousidis, C., 2022. Electroglottography based real-time voice-to-MIDI controller. Neuroscience Informatics, 2(2), p.100041.

[2] Donati, E., Chousidis, C., Ribeiro, H.D.M. and Russo, N., 2023. Classification of Speaking and Singing Voices Using Bioimpedance Measurements and Deep Learning. *Journal of Voice*.

The Effect of Decoding Strategies on the Quality and Diversity of Speech Generated by Large Speech Models

Adaeze Adigwe[1], *Zehai Tu*[2], *Simon King*[1]

[1] The Centre for Speech Technology Research, University of Edinburgh, United Kingdom
[2] LightSpeed Studios, Tencent

adaeze.adigwe@ed.ac.uk, zehaitu@global.tencent.com, Simon.King@ed.ac.uk

The latest trend in Text-to-Speech (TTS) is to use Large Language Models configured for use with speech, i.e., Large Speech Models. The potential of these models lies in their *general purpose capabilities*: they can be applied to various downstream speech generation tasks: TTS, zero-shot voice conversion, style transfer, speech editing, prosody transfer, etc.

These generative models are capable of producing prosodically varied renditions with high quality. However, the impressive samples used to showcase these models are often carefully cherry-picked. These models are also known to generate poor outputs containing artefacts and hallucinations, such as saying something different to the input text.

In speech generation, quality and expressivity are crucial. For generative models, both of these are related to the decoding strategy and its hyper-parameters: e.g.temperature influences output diversity; cf. similar studies in text and image generation [1]. Both theory and anecdotal evidence suggest that fewer artefacts and hallucinations should be produced at lower temperatures, at the cost of less diverse speech.

In this work we are interested in investigating what fraction of renditions are both error-free and expressive, as the decoding strategy / hyper-parameters are varied. Is there a trade-off between quality and diversity? Does less diverse output correspond to less expressive speech? What are the failure modes of these models?

1. Methodology

Our evaluations use two Large Speech Models: **a) VoiceCraft** [2]and **b) Parler-TTS** [3]. Parler-TTS is a Huggingface implementation of "Natural Language Guidance of High-Fidelity Text-to-Speech with Synthetic Annotations" [4]. These models represent different speech generation prompting strategies: VoiceCraft employs an audio prompt to control the speaker and style, whereas ParlerTTS relies on a natural language prompt to describe them. We select a range of test sentences that vary from easy & short to longer and more complex with punctuation, foreign words, compound words, etc.

We used pre-trained checkpoints provided by model authors and conducted hyper-parameter sweeps for the following decoding/sampling strategies:

1. *Temperature:* controls the model 'creativity' by making the probability distribution of tokens closer to uniform.
2. *Top-K:* how many options are considered when sampling.
3. *Top-P:* Also known as nucleus sampling, this restricts sampling to the set of most probable tokens with cumulative probability up to P.

For each synthetic utterance, we estimate a range of speech quality metrics. All of them are *automatic* estimates using publicly-available pre-trained models. (Future work will also involve listening tests.) Using the open-ai-whisper medium ASR model, we estimate Word Error Rate (WER). Using the Torchaudio Speech Quality and Intelligibility Measures, and the Brouhaha library, we compute: predicted MOS (mean opinion score), SI-SDR, speech-to-non-speech ratio, PESQ, STOI, SNR, and C50 (a measure of reverberation).

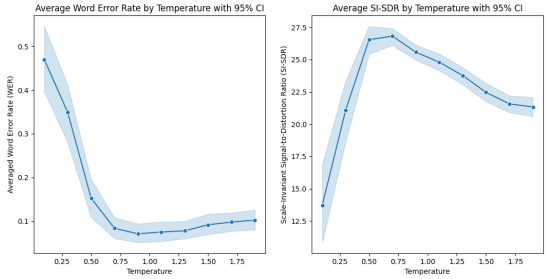


Figure 1: *ParlerTTS - Left: average WER vs. temperature range; Right: Average Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) vs. temperature*

Results We observe a high WER at low temperatures < 0.5 as in Figure 1. This is due to the model not generating all the words in the input. A similar behaviour at low temperatures can be seen in the plots for other metrics. Expert listening revealed patterns error types in these models: mid-sentence speaker changes, long pauses mid-utterance, noise/quality issues, and word skips/repetition. These errors tend to occur more frequently at extreme low and high temperatures. The complexity of the sentence adversely affects quality, but we have not fully explored this factor yet.

In future, we will explore whether combinations of these and other metrics can automatically filter out low quality speech, and thus be used to 'cherry pick' the best renditions.

Acknowledgement Funded by LightSpeed Studios, Tencent.

2. References

[1] H. Zhang, D. Duckworth, D. Ippolito, and A. Neelakantan, "Trading off diversity and quality in natural language generation," *arXiv preprint arXiv:2004.10450*, 2020.

[2] P. Peng, P.-Y. Huang, D. Li, A. Mohamed, and D. Harwath, "Voicecraft: Zero-shot speech editing and text-to-speech in the wild," *arXiv preprint arXiv:2403.16973*, 2024.

[3] Y. Lacombe, V. Srivastav, and S. Gandhi, "Parler-TTS," https://github.com/huggingface/parler-tts.

[4] D. Lyth and S. King, "Natural language guidance of high-fidelity text-to-speech with synthetic annotations," *arXiv preprint 2402.01912*, 2024.

Efficiency and (lack of) flexibility in an LSTM network model of spoken word recognition

Máté Aller, Matthew H. Davis

MRC Cognition and Brain Sciences Unit, University of Cambridge
mate.aller@mrc-cbu.cam.ac.uk, matt.davis@mrc-cbu.cam.ac.uk

Introduction

Recent advances in artificial neural networks have enabled the design of automatic speech recognition systems that identify spoken words with an accuracy approaching human listeners. By analysing the functional characteristics and internal representations of such systems, and comparing them to human listeners, we can gain novel insights into classic psycholinguistic findings and testable predictions for neuroimaging experiments exploring the neural computations for human speech perception.

Here we build on a recently published end-to-end model of human speech recognition ('EARSHOT', [1]). This is a recurrent neural network trained to map from acoustic representations of spoken words to representations of word meaning (semantics). It exhibits a human-like time course of word identification with parallel activation (due to phonological overlap) of onset-aligned 'cohort' neighbours (e.g. chain/change), and reduced, or delayed competition effects between rhyme neighbours (e.g. chain/gain). Our first steps reported here are to characterise EARSHOT's behaviour in recognising speech from varying talkers, at different rates and with reduced spectral detail. We discuss similarities and differences in behaviour and in the underlying computational mechanisms that support human listening.

Results and Discussion

Uniqueness point effects. Human listeners can recognize spoken words once sufficient speech is heard to uniquely identify a single word (i.e. at their uniqueness point, UP). Faster identification of early UP words depends on speech rate: slower speech increases the recognition advantage for early UP words, but UP effects disappear at fast speech rates for human listeners [2]. Similarly, EARSHOT shows a recognition time (RT) advantage for words with early compared to late UPs (main effect of UP position: $F(1,9)=5500.53$, $p<.0001$), which depends on speech rate (interaction between UP location and speech rate: $F(2,18)=2768.96$, $p<.0001$). Furthermore, EARSHOT shows accurate recognition of fast speech even when this is not included in training. However, unlike humans, EARSHOT continues to show an RT advantage for early UP items presented at a fast speech rate (paired t-test early vs. late UP locations for 2x rate: $t(9)=36.95$, $p<.0001$). Furthermore, EARSHOT exhibits much higher error rates (mean: 31.35% vs. 5.15% for humans in [2]). This is largely due to poor cross-talker generalization [1].

Talker change effects. Human listeners show a consistent RT cost when words from different talkers are mixed in a test task compared to when words are grouped by talker [3]. To test this in EARSHOT, we trained and tested models in which words are grouped by talker. Unlike humans, EARSHOT, shows no effect of talker changes on recognition time (main effect of grouping for excluded words: $F(1,9)=1.09$, $p=.3247$), or error rates ($F(1,9)=1.18$, $p=.3050$).

Effects of signal degradation. Compared to clear speech for normally-hearing (NH) listeners, cochlear implant users (CI) and NH listeners hearing spectrally-degraded speech show impaired word recognition [4]. Degraded words are less accurately and more slowly recognised; they exhibit reduced competition from onset neighbours (e.g., chain/change) and increased competition from rhyme competitors (e.g., chain/gain). Training EARSHOT on degraded speech replicates effects seen in CI users: models trained on 4 or 16 channel vocoded speech learned more slowly and were less accurate than models trained on clear speech ($F(2,28)=4361.15$, $p<.0001$). Recognition was slower for target words heard as vocoded speech compared to clear speech, in a way that varies as a function of phonological overlap (degradation by word type interaction: $F(6,84)=3902.83$, $p<.0001$). When presented with time compressed speech, recognition accuracy deteriorated more for degraded than for clear speech (degradation by speech rate interaction: $F(2,28)=13.71$, $p<.0001$). However, EARSHOT cannot simulate results for NH listeners hearing degraded speech since models do not generalize from clear to degraded speech or vice-versa (maximum generalization accuracy <5%).

Discussion. EARSHOT shows a human-like time-course of word identification, and can also recognise speech at different rates and after spectral degradation. However, it is not constrained by processing time limitations and hence can recognise faster speech more efficiently than humans. Furthermore, EARSHOT fails to generalize over talkers or benefit from listening conditions in which talker identity can predicted for successive words suggesting that it lacks human-like mechanisms for speaker normalisation. This motivates future explorations of models that include explicit talker-identity tasks. We will also consider how other forms of prior knowledge (e.g. priming) influence neural network representations and simulate effects of prediction and perceptual learning during human speech perception.

References

[1] J. S. Magnuson *et al.*, 'EARSHOT: A Minimal Neural Network Model of Incremental Human Speech Recognition', *Cognitive Science*, vol. 44, no. 4, p. e12823, 2020

[2] M. Radeau, J. Morais, P. Mousty, and P. Bertelson, 'The Effect of Speaking Rate on the Role of the Uniqueness Point in Spoken Word Recognition', *Journal of Memory and Language*, vol. 42, no. 3, pp. 406–422, Apr. 2000

[3] J. S. Magnuson, H. C. Nusbaum, R. Akahane-Yamada, and D. Saltzman, 'Talker familiarity and the accommodation of talker variability', *Atten Percept Psychophys*, vol. 83, no. 4, pp. 1842–1860, May 2021

[4] B. McMurray, A. Farris-Trimble, and H. Rigler, 'Waiting for lexical access: Cochlear implants or severely degraded input lead listeners to process speech less incrementally', *Cognition*, vol. 169, pp. 147–164, Dec. 2017

Designing Conversational Assistants to Discourage Abusive Language

Tanvi Dinkar[*], *Gavin Abercrombie*[*], *Chris Pidcock*[°], *Benedict Jones*[◇], *Matthew Aylett*[°*]

[◇]University of Strathclyde, [°]Cereproc Text-to-Speech, [*]Heriot Watt University

{T.Dinkar, G.Abercrombie, M.Aylett}@hw.ac.uk, chris@cereproc.com,
benedict.jones@strath.ac.uk

Research in text-to-speech (TTS) has found that conversational assistants with female gendered voices (such as those of many conversational assistants such as Amazon's Alexa or Apples's Siri) often give "deflecting, lacklustre or apologetic" responses to abusive, sexist comments, enciting further abuse and perpetuating gender stereotypes [1]. For example, [1] famously showed that one of Siri's responses to user abuse was "I'd blush if I could". According to [2], companies such as Amazon and Apple justify these deliberate design choices, citing 'user preference' as the reason for creating voices anthropomorphised as female. Additionally, simply creating synthesised voices perceived as gender neutral may not be an ideal solution, or even possible, as people still attribute binary gender to synthesised voices despite minimal or no gender markers [3]. The aim of this research is to focus on creating synthesised voices aimed at discouraging such responses.

Related work. Previous work has shown that voices perceived as male tend to be perceived as more dominant, aggressive and mean [4]. [5] found that acoustic factors such as pitch are correlated with perceived dominance, and therefore, altering the acoustic features can play a role in the way synthesised voices are perceived socially. Research has also explored adapting the synthesised voice to a given social situation, although in a different type of scenario. For example, [6] focused on altering the perceived personality of the synthesised voice (focusing on Big five personality traits), motivated by adapting the voice to the user's personality. They found that certain personality traits could be synthesised, and changed within a controlled range (e.g. low to high agreeableness and extroversion). Following a similar method, we will focus on synthesising voices perceived as friendly yet dominant to test whether such voices discourage abuse. This entails creating a synthesised voice than can be controlled for low/ high valence and dominance.

Data and voice selection. We use data from the LibriTTS-R corpus [7], which contains speech data from 2,456 individuals. We initially select speakers that only have > 5 minutes of audio, as using speakers with audio less than this may result in a synthesised voice that does not sound natural.

Method and participant study. A standard for rating human voices when assessing pitch and formant frequencies is to select words that contain vowels [8]. Given that listeners only require very short exposures to voices to form impressions [9], following [5] we extract instances of one word each per vowel, and join these instances together to form one audio clip for each speaker. We will recruit study participants on the crowdsourcing platform Prolific.[1].

1. We will collect ratings of the voices in order to synthesise a TTS voice that can be controlled for the two dimensions. Participants will be presented with the voice exerts on an in-house platform and asked to rate the voices on the two dimensions using a likert scale. Participants will additionally be asked to rate the voice clip on a scale of masculine/feminine.
2. Similar to the procedure described in [6], we plan to create a synthesised voice using using the ratings collected previously and CereProc's speech synthesis system CereWave. This will result in voices that can be controlled for the two dimensions. allowing us to assess their impact on user behaviour.

[1] https://www.prolific.com/

1. Acknowledgements

Tanvi Dinkar, Gavin Abercrombie, Matthew Aylett and Benedict Jones are supported by the EPSRC project 'Gender Bias in Conversational AI' (EP/T023767/1).

2. References

[1] M. West, R. Kraut, and H. Ei Chew, *I'd Blush if I Could: Closing Gender Divides in Digital Skills through Education*, 2019.

[2] J. Stern, "Alexa, Siri, Cortana: The problem with all-female digital assistants," , 2017, Wall Street Journal. Accessed: 2024-05-08.

[3] M. P. Aylett, S. J. Sutton, and Y. Vazquez-Alvarez, "The right kind of unnatural: designing a robot voice," in *Proceedings of CUI*. ACM, 2019.

[4] B. Jones and V. K. M. Shiramizu, "Do people perceive male and female artificial intelligence (AI) conversational agents' voices differently?" 2023. [Online]. Available: https://doi.org/10.31234/osf.io/9v8e4

[5] V. K. Shiramizu, A. J. Lee, D. Altenburg, D. R. Feinberg, and B. C. Jones, "The role of valence, dominance, and pitch in perceptions of artificial intelligence (AI) conversational agents' voices," *Scientific Reports*, vol. 12, 2022.

[6] S. Gao, M. P. Aylett, D. A. Braude, and C. Lai, "Synthesising personality with neural speech synthesis," in *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. ACL, 2023, pp. 393–399.

[7] Y. Koizumi, H. Zen, S. Karita, Y. Ding, K. Yatabe, N. Morioka, M. Bacchiani, Y. Zhang, W. Han, and A. Bapna, "LibriTTS-R: A restored multi-speaker text-to-speech corpus," *arXiv e-prints*, pp. arXiv–2305, 2023.

[8] M. Armstrong, A. Lee, and D. Feinberg, "A house of cards: bias in perception of body size mediates the relationship between voice pitch and perceptions of dominance," *Animal Behaviour*, vol. 147, pp. 43–51, 2019.

[9] N. Lavan, "The time course of person perception from voices: A behavioral study," *Psychological Science*, vol. 34, no. 7, pp. 771–783, 2023, pMID: 37227791. [Online]. Available: https://doi.org/10.1177/09567976231161565

MULTICHANNEL BINAURAL SPEECH ENHANCEMENT USING DEEP COMPLEX CONVOLUTIONAL TRANSFORMER NETWORKS

Vikas Tokala[1], *Emilie d'Olne* [1], *Mike Brookes*[1], *Simon Doclo*[2], *Jesper Jensen*[3,4], *Patrick A. Naylor* [1]

[1]Dept. of Electrical and Electronic Engineering, Imperial College London, UK
[2] Dept. of Medical Physics and Acoustics, Carl von Ossietzky Universität Oldenburg, Germany.
[3] Demant A/S, Smørum, Denmark.
[4] Dept. of Electronic Systems, Aalborg University, Denmark

v.tokala@imperial.ac.uk

Modern assistive listening devices such as binaural hearing aids and virtual reality headsets are equipped with multiple microphones allowing speech enhancement methods to utilize this multichannel input for binaural speech enhancement. In this work, we propose two neural multichannel binaural speech enhancement methods that aim to preserve the spatial information of speech while improving intelligibility and suppressing an additive noise field. They incorporate a complex convolutional neural network guided by a beamformer (Fig. 2) and an end-to-end neural network based on a convolutional encoder-decoder architecture (Fig. 1). The networks estimate complex ratio masks in the time-frequency domain for the left and right channels. In terms of objective metrics, simulation results show that the proposed methods significantly improve binaural speech intelligibility and reduce noise while preserving the binaural spatial cues of the speech signal in acoustic situations with a single target speaker in the frontal azimuthal plane and isotropic noise of various types.

In this study, we introduce two types of multichannel binaural speech enhancement systems. The first utilizes a beamformer to guide the enhancement model(Guided OMVDR (GOMVDR) and Guided FCIM (GFCIM)), while the second employs an end-to-end neural network architecture (Multichannel Binaural Complex Convolutional Transformer Network (MBCCTN)) based on the Convolutional Encoder-Decoder (CED) architecture for multichannel binaural speech enhancement. Figure 1 shows the block diagram of the end-to-end neural model and Fig. 2 shows the modified encoder stage for the beamformer-guided model. For the beamformer stage, binaural versions of the Oracle MVDR (OMVDR) and Fixed Cylindrical Isotropic MVDR (FCIM) beamformers are considered. An OMVDR beamformer with access to the clean speech statistics is used for comparison as an approximate upper bound on performance. The loss function $\mathcal{L}$ is given by

$$\mathcal{L}=\alpha\mathcal{L}_{SNR}+\beta\mathcal{L}_{STOI}+\gamma\mathcal{L}_{ILD}+\kappa\mathcal{L}_{IPD}, \quad (1)$$

where $\mathcal{L}_{SNR}$ is the Signal-to-Noise Ratio (SNR) loss, $\mathcal{L}_{STOI}$ is the STOI loss, and $\mathcal{L}_{ILD}$ and $\mathcal{L}_{IPD}$ are the Interaural Level Difference (ILD) and Interaural Phase Difference (IPD) error losses. The parameters α, β, γ, and κ are weights applied to each term.

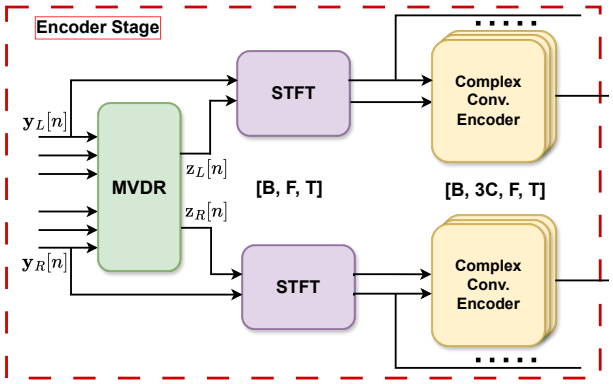


Figure 2: *Beamformer-based encoder stage for the GOMVDR and GFCIM enhancers.*

The proposed methods provided a maximum Frequency-weighted Segmental SNR (fwSegSNR) improvement of 15 dB and a Modified Binaural STOI (MBSTOI) score improvement of 0.2 over the noisy signal. The ILD errors were calculated for all the models and the errors were under 0.5 dB for all input SNR (iSNR)s. The Interaural Signal Phase Difference (ISPD) is calculated by evaluating the models using clean speech signals and measuring the errors in IPD before and after processing and the proposed methods had an ISPD error below 6°. Listening examples can be found at https://vikastokala.github.io/mbcctn/.

This work was supported by funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 956369 and the UK Engineering and Physical Sciences Research Council [grant number EP/S035842/1].

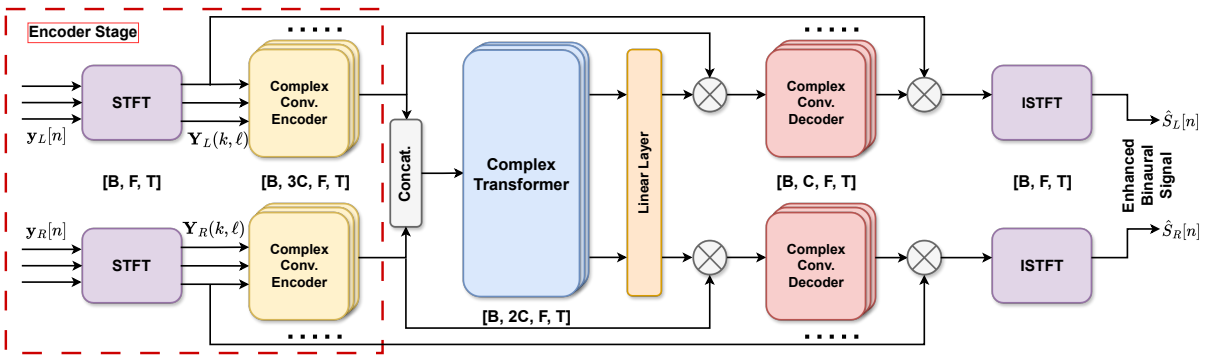


Figure 1: *Block diagram of the model architecture of the MBCCTN enhancer.*

Low-dimensional Style Token Control for Hyperarticulated Speech Synthesis

Dan Wells,[1] *Miku Nishihara,*[2] *Korin Richmond,*[1] *Aidan Pine*[3]

[1]The Centre for Speech Technology Research, University of Edinburgh, United Kingdom
[2]Department of Computer Science, Nagoya Institute of Technology, Japan
[3]National Research Council Canada, Canada

{dan.wells, korin.richmond}@ed.ac.uk, m.nishihara.966@stn.nitech.ac.jp,
aidan.pine@nrc-cnrc.gc.ca

Global style tokens (GSTs) allow for rich modelling of the variation in a speech corpus and subsequent control of text-to-speech synthesis (TTS) [1]. However, certain styles of speech may be marked by variation along multiple dimensions, complicating the interpretation and control of learned style tokens. One example is hyperarticulated or 'clear' speech, which in English is characterised by reduced speaking rate, increased F0, more complete articulation of vowels and plosive consonants, among other factors [2]. We are interested in synthesising hyperarticulated speech for computer-assisted language learning, where high-quality TTS would enable arbitrary generation of speech samples for students, and hyperarticulation in those samples may aid pronunciation learning for new words and unfamiliar speech sounds. Teachers or students might additionally want explicit control over the degree of hyperarticulation in generated samples, for example moving from very clear and careful speech in early language learning toward a more fluent and conversational style after gaining familiarity with the language.

We trained a FastSpeech 2 model [3] with a GST style encoder on data from the Alba speech corpus, comprising around 4 hours of normal read speech (*Plain*) plus 20 minutes of prompts read in a clear speech style *"as if talking to somebody with a hearing impairment"* (*Clear*) [4]. We then applied principal component analysis (PCA) to GST weights extracted from this model for parallel utterances in *Plain* and *Clear* styles, as shown in Figure 1. Note that *Plain* utterances appear to form two clusters (blue and red points), where only one was expected from the labels in the corpus; we excluded red utterances and reapplied PCA to focus on a single *Plain* style going forward. The resulting PCA1 dimension accounted for 50% of the variation in the data, so we decided to use this as our single control parameter for synthesis. To do this, we selected several test points along PCA1, where lower values correspond to more *Clear* articulation, set all other dimensions to their global mean values, and then applied the inverse PCA transformation to produce the appropriate GST weights for our FastSpeech 2 model.

We conducted a subjective listening test to assess how well we could control the two styles by presenting participants with pairs of utterances synthesised at 9 points along the PCA1 dimension, covering and extending beyond the observed range of the data, and asking them to select which *"sounds most like the person is trying to speak clearly or carefully."* The results in Figure 2 indicate that synthesising at points further toward the *Plain* end of PCA1 indeed produces samples which are judged progressively less *Clear*-styled. On the other hand, while our model is able to produce recognisably *Clear* speech, there appears to be less fine-grained control within this style. We similarly found slightly worse performance on *Clear* samples in a separate audio quality MOS evaluation and in objective measures of synthesised F0 and segment durations.

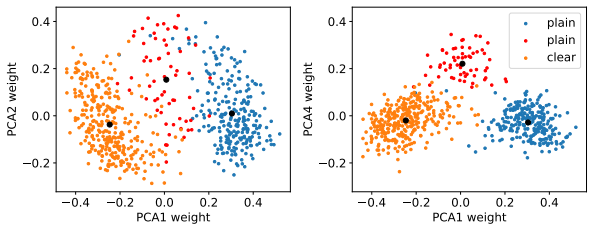


Figure 1: *Initial PCA projections of GST weights for parallel utterances, labelled by style. Black circles mark cluster means.*

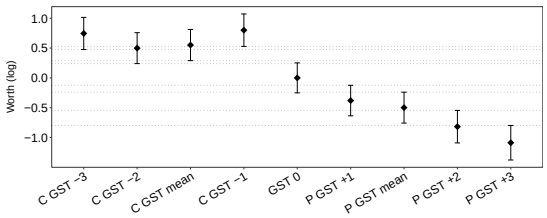


Figure 2: *Clearness AB results for utterances synthesised at different points along the PCA1 axis. Worth values and 95% confidence intervals from a Bradley-Terry model of system rankings.*

1. References

[1] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5180–5189.

[2] M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking Clearly for the Hard of Hearing II," *Journal of Speech, Language, and Hearing Research*, vol. 29, no. 4, pp. 434–446, 1986.

[3] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations*, 2021.

[4] C. Valentini-Botinhao, M. Wester, J. Yamagishi, M. Toman, M. Pucher, and D. Schabus, "Non linear time compression of clear and normal speech at high rates," 2019, arXiv:1901.07239 [eess].

Acknowledgements: We would like to thank Professor Keiichi Tokuda and Associate Professor Yoshihiko Nankaku of Nagoya Institute of Technology for participation in discussions of this research. 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), the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Sciences, and the National Research Council of Canada's Ideation Fund: 'Small Teams – Big Ideas'.

Spoken English Learner Corpus Transcription Validation Process

Mateus Miranda[1]

1 Cambridge Language Sciences Research Centre, University of Cambridge, UK.
mm2542@cam.ac.uk, mateusesm2@gmail.com

Abstract

The architecture of a learner spoken corpus involves several steps, such as design, data recording, the creation of transcription parameters, and the validation process. Speech has several elements that must be registered in the transposition to written form, making transcription by humans time consuming, but software promises automatic transcription. This paper, which reports on a broader oral corpus compilation project of more than 100 recordings of learner language calibrated to the Common European Framework of Reference (CEFR) for pragmatics studies, seeks to discuss the transcription procedures established to reduce disagreements related to the guidelines. The methodology combines the analyses from automatic speech transcription software and human transcribers that includes (a) transcription workshops and (b) transcription tasks. Finally, it presents the corpus transcription validation process to reduce disagreements related to the project criteria to prevent the lack of homogeneity in the final product.

1. Corpus Compilation

The recordings covered two different tasks:

(1) Free discussion: An informal conversation about students` lives, interests, hobbies, etc.; and

(2) Thought-provoking picture description: This part is based on a set of thought-provoking pictures. The interviewer shows the participants some pictures and asks them to choose two to set the scene and describe them.

Table 1: *Corpus Information*

Corpus Matrix	
Size	116 recordings (20 to 30 minutes)
Transcription	Orthographic
CEFR levels	A2 to C2
Gender	Balanced
Age group	Varied

2. Transcription Software *vs* Learner Language

Spoken language often necessitates a validation process for quality control and improvement [1] especially when not transcribed by a single person. Thus, the following quality control and improvement steps were taken:

STEP 1: Analyse the functionality and suitability of automatic speech transcription software for learner language. The procedures adopted in the first experiment were to:
(a) submit the recordings for transcription by software;
(b) analyse their performance to verify the suitability of offering a quality transcription; and
(c) to evaluate the software, three short extracts from a CEFR B1 level learner were chosen to be automatically transcribed by three paid services: Temi, Trint and Otter.

STEP 2: To reduce disagreements related to the corpus transcription, workshops and training sessions were provided to transcribers.
The workshops' main goals were to:
(a) give an overview on learner corpus research;
(b) introduce the corpus architecture and transcription criteria; present tools useful for transcription (Notepad++, Audacity, Praat and WinPitch); and
(c) submit the three CEFR B1 short extracts to be manually transcribed by them and check inter-rater agreement.

STEP 3: Measure inter-rater agreement between transcribers.

One transcriber was ranked first among the group and was selected to help in the revision process to keep the final corpus transcriptions homogeneous. After the inter-rater agreement analysis, a feedback section was held to focus on transcribers' difficulties and reduce disagreement in future transcriptions.

3. Conclusion

- The transcription of learners' speech requires the establishment of specific criteria and, consequently, requires the training of those who will transcribe for a homogeneous corpus;
- This can be done through training and monitoring of the transcription team and leads to an improvement in manual transcription;
- Software transcription can be provided to serve as a means of comparison or as a basis for manual transcription;
- Software programs still struggle to offer accurate transcription of native speakers or learner language, especially when two or more participants are involved.

4. References

[1] Schiel F. Baumann, A. Draxler, C. Ellbogen, T. Hoole, P, and Steffen, A. The validation of speech corpora. Munich: University of Munich, 2004.

Spontaneous and Scripted Speech Classification for Multilingual Audio

Shahar Elisha[1,2], *Emmanouil Benetos*[1], *Mariano Beguerisse-Díaz*[2]

[1]Queen Mary University of London, UK
[2]Spotify
s.elisha@qmul.ac.uk

Large scale audio-based classification of spontaneous and scripted speech in multiple languages can enable a range of applications. Firstly, increasingly large, multilingual audio catalogues require efficient, performant and generalisable content segmentation models. As user preferences range over stylistic variations [1], enriching catalogues with style labels allows improved recommendation systems and discovery experiences for listeners. Secondly, we want to address the need to understand the effects of speaking style and language on downstream speech processing models [2, 3, 4, 5]. This is particularly important today, given the explosion in popularity, accessibility and performance of modern audio models, and the increasing availability of large datasets of natural speech. While the differences between scripted and improvised speech have been investigated in the linguistics literature [6], only a handful of papers approach the problem from a computational perspective [7, 8, 9]. We pursue this line of research with an emphasis on large-scale and diverse data, and a comparison between large audio models and handcrafted acoustic features.

We model the task as a binary classification of scripted or spontaneous speech. First, we compile a dataset of 2,400 manually annotated podcast episodes, of which 1,000 are "scripted" and 1,400 are "spontaneous", totalling 1,900 hours of audio (see Fig. 1 for a breakdown by language). We then extract a range of audio-based features: handcrafted acoustic and prosodic features (e.g. pitch, loudness, MFCCs, and duration of voiced and unvoiced segments), YAMNet [10], Whisper [11], and a Whisper model fine-tuned on podcast audio. We train, evaluate, and compare the performance of the different audio features, with an emphasis on understanding performance by language. Lastly, we measure how well a model trained on podcasts translates to other domains and publicly accessible datasets. To our knowledge, this is the first attempt at building a spontaneous versus scripted speech classifier on a large and diverse multilingual dataset.

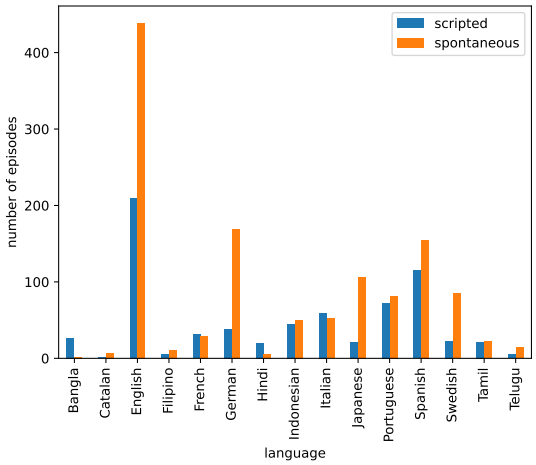


Figure 1: *Distribution of scripted and spontaneous episodes by language*

1. References

[1] K. Martikainen, J. Karlgren, and K. P. Truong, "Exploring audio-based stylistic variation in podcasts," in *Interspeech 2022*, 2022.

[2] P. Wagner, J. Trouvain, and F. Zimmerer, "In defense of stylistic diversity in speech research," *Journal of Phonetics*, vol. 48, pp. 1–12, 2015, the Impact of Stylistic Diversity on Phonetic and Phonological Evidence and Modeling.

[3] Y. Lee and J. Kreiman, "Acoustic voice variation in spontaneous speech," *The Journal of the Acoustical Society of America*, vol. 151, no. 5, pp. 3462–3472, 05 2022.

[4] J. Williams and S. King, "Disentangling Style Factors from Speaker Representations," in *Proc. Interspeech 2019*, 2019, pp. 3945–3949.

[5] Q. Li and T. Chaspari, "Exploring transfer learning between scripted and spontaneous speech for emotion recognition," in *2019 International Conference on Multimodal Interaction*, ser. ICMI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 435–439.

[6] B. V. Tucker and Y. Mukai, *Spontaneous Speech*, ser. Elements in Phonetics. Cambridge University Press, 2023.

[7] N. Ryant and M. Liberman, "Automatic Analysis of Phonetic Speech Style Dimensions," in *Proc. Interspeech 2016*, 2016, pp. 77–81.

[8] 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.

[9] G. Christodoulides, "Speaking Style Prosodic Variation and the Prosody-Syntax Interface: A Large-Scale Corpus Study," in *Proc. Speech Prosody 2020*, 2020, pp. 705–709.

[10] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[11] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 28 492–28 518.

Exploring Gender Disparities in Automatic Speech Recognition Technology

Hend ElGhazaly, Dr. Bahman Mirheidari, Dr. Nafise Sadat Moosavi, Professor Heidi Christensen

The University of Sheffield, UK

helghazaly1@sheffield.ac.uk, b.mirheidari@sheffield.ac.uk, n.s.moosavi@sheffield.ac.uk, heidi.christensen@sheffield.ac.uk

1. Introduction

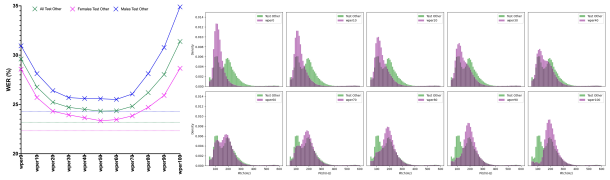
Speech technologies have demonstrated their effectiveness and efficiency across various domains, however, they can inadvertently perpetuate biases. Bias refers to a performance gap discriminating against certain groups or individuals. Despite extensive research on mitigating bias in AI, the concept of fairness in speech technologies remains relatively underexplored. In speech research, the most common types of bias relate to demographics such as gender, race, accents, and language variants [1]. Imbalanced datasets can lead to a mismatch between the speaker and the trained acoustic model. While it is impossible to de-bias speech systems, developing fair systems requires mitigating such biases and not exacerbating existing ones. Several studies showed a correlation between the gender distribution in the training set and the performance of speech models. However, different studies have come to different conclusions. For instance, [2] demonstrated that demographically balanced training sets improve the fairness of speaker recognition systems.[3] claim that increasing the number of women audio files in the dataset results in better performance for both women and overall speakers. On the other hand, [4] showed that training data does not need to be gender balanced to ensure the best performance. The contradictory results suggest other factors might affect the performance apart from the speakers' gender. This study aims to provide a preliminary exploration of gender bias in ASR moving beyond the traditional focus on gender ratios.[1] Our exploration includes an analysis of other potential influencing factors along with gender, such as the difficulty of the text in training and test data, the semantic similarity between the contents of training and test sets, and the pitch distribution of speakers.

2. Methodology and Experimental Design

The LibriSpeech Kaldi recipe and dataset were used to establish a baseline system. The LibriSpeech corpus was used to create 11 training subsets with varying gender distributions to investigate to what extent increasing the number of women audio files in the training set improved the ASR performance (WER). The women's percentage in the training set was changed incrementally from zero to 100. Only the number of audio files with female speakers changed in each training set. For the text analysis, Flesch-Kincaid Grade Level and cosine similarity scores of the datasets were computed to measure the text difficulty and semantic similarity respectively.

3. Results

While gender representation in training data is a significant factor, our study reveals that it should not be the sole focus. Instead, a combination of diverse factors must be considered to enhance the fairness and performance of ASRs. We explore text complexity within each chapter, semantic similarity between training and test data, and pitch distribution in both training and test data. Our findings suggest a complex interplay between the gender ratio in training data and ASR performance, challenging the notion that adjusting gender ratios alone can improve system accuracy. For example, scenarios where training data is exclusively composed of women's voices do not consistently result in lower WER for women compared to a more balanced gender representation. This indicates that while gender ratio in training data is a crucial factor, their impact on ASR performance is not singular, echoing the insights from researchers like [4], who highlight the importance of content and prosody. Moreover, our investigation into the text difficulty and semantic similarity suggest that these elements do not directly affect the WER in various settings. Additionally, our analysis of pitch distributions brings to light significant differences contingent upon the pitches included in the training data, emphasizing the need for a balanced representation to encompass a diverse range of pitches regardless of gender.



(a) *Mean WERs* (b) *Pitch distributions*

Figure 1: *Illustrations of the mean WERs and pitch distributions in the evaluation of the Test Other set*

4. References

[1] S. Feng, B. M. Halpern, O. Kudina, and O. Scharenborg, "Towards inclusive automatic speech recognition," *Computer Speech & Language*, vol. 84, 2024.

[2] G. Fenu, M. Marras, G. Medda, G. Meloni *et al.*, "Fair voice biometrics: Impact of demographic imbalance on group fairness in speaker recognition," in *Interspeech*, 2021.

[3] M. Garnerin, S. Rossato, and L. Besacier, "Investigating the impact of gender representation in asr training data: A case study on librispeech," in *3rd Workshop on Gender Bias in Natural Language Processing*, 2021.

[4] Y. Meng, Y.-H. Chou, A. T. Liu, and H.-y. Lee, "Don't speak too fast: The impact of data bias on self-supervised speech models," in *IEEE ICASSP*, 2022.

[1] We acknowledge that the gender spectrum is more diverse. Our focus on only 2 genders is driven by their representation in the dataset.

Word Time Prediction for Attention-Based Encoder-Decoder Models in Automatic Speech Recognition

Dongcheng Jiang & Phil Woodland

Department of Engineering, University of Cambridge, Trumpington St., Cambridge, CB2 1PZ UK
{dj346,pw117}@cam.ac.uk

Attention-based encoder-decoder (AED) models have been popular in automatic speech recognition (ASR) and include the widely-used Whisper model [1]. Apart from low word error rates (WERs), exact word timing can also be important, which requires identifying non-speech regions and locating words in speech regions. By default, current AED models do not detect local non-speech, and it has been found that using the cross-attention weight matrices for best-path alignment cannot guarantee good time accuracy, and external models can be needed for accurate timing [2]. To remove the need for an external model for time alignment, two types of add-on module for a pre-trained AED model are here proposed and compared, namely a label-synchronous aligner (L-sync) and a frame-synchronous aligner (F-sync). Both take the AED backbone encoder and decoder vectors as inputs, and during aligner training the AED backbone is frozen so that ASR performance is not affected.

Label-synchronous (L-sync) aligner

The L-sync aligner [3] tries to segment the original utterance for alignment. Each word or non-speech duration is normalised by the utterance length to form normalised durations. The aligner jointly models the duration of each word and possible non-speech before (or after) the word. For each utterance the cross-entropy loss is applied between the predicted normalised duration and the reference based on ground-truth timestamps. Hence the L-sync aligner is trained via supervised learning with external timestamps, which can be obtained using ASR systems based on hidden Markov Models (HMMs) since the HMM-based systems naturally model unit boundaries and the timestamps generated by Viterbi alignment are empirically accurate. During time alignment, given the utterance length, the time of each word or non-speech segment can be recovered by a forward pass through the model to find the normalised duration.

Frame-synchronous (F-sync) aligner

The F-sync aligner removes the need for external timestamps during training by comparing AED models with HMM-based systems, which use a separate acoustic model and language model (LM). Neural networks for acoustic modelling are trained using frame-level cross-entropy, and HMM state priors discounted to obtain acoustic (pseudo-)likelihoods for generative modelling and the LM integrated during decoding. AED models need to discount the internal LM and integrate another LM during decoding to obtain lower WERs. The key difference between the state prior and the internal LM is the duration of each unit, which can be estimated from the posterior probability. Therefore the F-sync aligner imitates single emitting state HMMs and uses full-sum training with acoustic-based probabilities by discounting the duration and internal LM from the original output (see Fig. 1). Voice activity detection (VAD) was trained using supervised learning and the VAD output is also used during full-sum training. The Viterbi algorithm is used for alignment with acoustic probabilities and a single emitting state HMM topology.

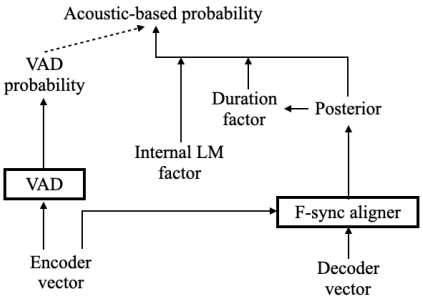


Figure 1: *Schematic for F-sync aligner*

Key results

The AED ASR model was trained using Librispeech train-clean-100 with WERs 10%/26% on test-clean/other. The forced alignment in [4] was adopted as the time reference. The average absolute mismatch of word start and end time were measured, together with the accuracy using a 200 ms collar. Using the Librispeech test sets for forced alignment, the L-sync aligner can obtain 95% accuracy and average mismatch about 80 ms, while the F-sync aligner can obtain 99% accuracy and an average mismatch 40-50 ms. Since F-sync uses the Viterbi algorithm and L-sync only requires a model forward pass, L-sync is much faster and better for batch processing. Similar experiments have also been performed for Whisper. Initial tests on Whisper's built-in function for word timestamps showed much poorer accuracy (average mismatch over 150 ms) compared to the L-sync and F-sync aligners.

1. References

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023.

[2] M. Bain, J. Huh, T. Han, and A. Zisserman, "WhisperX: Time-accurate speech transcription of long-form audio," in *Proc. Interspeech*, 2023.

[3] D. Jiang, C. Zhang, and P. C. Woodland, "A neural time alignment module for end-to-end automatic speech recognition," in *Proc. Interspeech*, 2023.

[4] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, "Speech model pre-training for end-to-end spoken language understanding," in *Proc. Interspeech*, 2019.

Dongcheng Jiang is funded by Trinity College, University of Cambridge. 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.

Semantic Map-based Generation of Navigation Instructions

Svetlana Stoyanchev[2], *Chengzu Li*[1,2]*, *Chao Zhang*[2], *Simone Teufel*[1], *Rama Sanand Doddipatla*[2],

[1]University of Cambridge, [2]Toshiba Europe Limited

{cl917, sht25}@cam.ac.uk
{svetlana.stoyanchev, chao.zhang, rama.doddipatla}@toshiba.eu

Summary and Motivation

Our longer-term goal is to build mobile robots with spatial awareness and reasoning capabilities which can follow natural language commands and verbally communicate with users. Vision and Language Navigation task (VLN) addresses robot navigation in a physical or virtual environment following natural language instructions [1]. The data annotation for this task is time-consuming and costly to scale up, and the development of models that address the task is severely limited by the availability of training data. Automatically generated instructions may be used to augment training data and improve the performance of VLN systems [2, 3]. Generating instructions automatically is also useful for the scenarios when a person navigates in an unknown space following system-generated instructions.

In this work, we explore the navigation instruction generation (VL-GEN), which is the inverse of the VLN task in that it generates natural language instructions for a path in the virtual (or physical) environment. VL-GEN model generates language instruction in the context of the physical environment, grounding object references and action instructions to the given space. Previous studies frame VL-GEN as the end-to-end task of generating text from a sequence of photo-realistic RGB images as the visual input [2, 3, 4]. One of the shortcomings of this approach is that using panoramic images is resource-intensive as it requires processing of multiple image inputs corresponding to different points on the path. In addition, panoramic images contain many details that are irrelevant for the task. The model has to learn to interpret the environments from RGB panoramas, such as object recognition, and generate instructions at the same time.

In this work, we frame the instruction generation task as image captioning and use semantic maps as the intermediate representation for spatial reasoning. We envision the separation of the VL-GEN task into two steps: 1) environment interpretation, which is addressed by semantic SLAM in physical robotic systems, and 2) spatial reasoning. We focus on the second step and explore the feasibility of using top-down semantic map for instruction generation. This approach is human-interpretable as it is natural for people to understand navigation instructions from a top-down map. The proposed approach is also light-weight as it encodes information necessary for the navigation in a single abstract image.

Approach Overview

We formalize the VL-GEN task as image captioning with the input of a semantic map with the path creating the dataset with top-down semantic maps derived from the Room2Room corpus [5], as illustrated on Figure 1. Our research question asks whether it is feasible to use the top-down semantic map (a single RGB image) as our main source of information. We also explore whether additional information (*panoramas, region names, and actions*) can further improve performance. Our automatic evaluation shows that the proposed method without the use panorama images performs on-par with the existing LSTM-based end-to-end approaches that use only panoramic images. Our human evaluation results indicate that incorporating more information in different modalities tends to improve the performance for the generation task, however, the absolute performance the models remains low, indicating that there is much room for improvement.

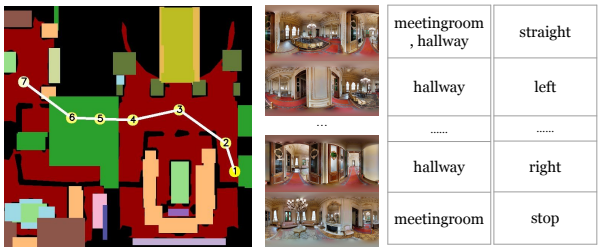


Figure 1: *An example navigation scenario from our new dataset for instruction generation, with the navigation path overlayed on the semantic map.*

1. References

[1] W. Wu, T. Chang, and X. Li, "Visual-and-language navigation: A survey and taxonomy," *arXiv preprint arXiv:2108.11544*, 2021.

[2] D. Fried, R. Hu, V. Cirik, A. Rohrbach, J. Andreas, L.-P. Morency, T. Berg-Kirkpatrick, K. Saenko, D. Klein, and T. Darrell, "Speaker-follower models for vision-and-language navigation," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[3] H. Tan, L. Yu, and M. Bansal, "Learning to navigate unseen environments: Back translation with environmental dropout," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

[4] S. Wang *et al.*, "Less is more: Generating grounded navigation instructions from landmarks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[5] P. Anderson *et al.*, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

*Work done on internship at Toshiba Europe Limited.

Leveraging data sources for foundation models in dialectal Arabic TTS

Simon Shelley, Omar Elsherief, Oscar Saz

Emotech Ltd, London, UK

simon@emotech.co, omar@emotech.co, oscar@emotech.co

Abstract

Speech synthesis is currently experiencing an era of rapid development as part of the wider push for generative AI. The latest architectures aim to use deep learning to learn multi–lingual, multi–speaker models as foundation models. However, these still favour languages where more data is available, especially English; while many other languages are still under–represented. Synthesis of dialectal Arabic speech remains largely a challenge due to multiple reasons, the main one being that little written content exists for them, as the majority of written Arabic is either Classic Arabic (CA) or Modern Standard Arabic (MSA). This paper presents a strategy to leverage sources of data like conversational speech and found data to build foundation models for Arabic dialect speech synthesis.

1. Introduction

Foundation models in speech and language are currently the state–of–the–art for many tasks. They are characterised by being trained on large quantities of multilingual audio and text and having billions of parameters that can learn to encode all the features of the different languages and speakers. Examples of this are Whisper [1] or wav2vec2.0 for Automatic Speech Recognition (ASR); GPT or Llama [2] for Natural Language Processing (NLP); and TorToiSe [3] for Text–to–Speech Synthesis (TTS).

Foundation speech synthesis models are trained with large amounts of multilingual and multispeaker data, so that they can learn to generate audio with a natural pronunciation, correct emotion and matching a given speaker voice with no or minimal data. With this foundation, they can be used to "clone" voices with little data. Although there are ethical connotations to such voice "cloning", there are also multiple applications in research and industry where these models can be helpful.

2. Experimental Setup

This work focuses on main dialects of the Arabian peninsula. These are Najdi Arabic, from Riyadh and Central Saudi Arabia; Hejazi Arabic, from Jeddah and the Red Sea coast; and Gulf Arabic, from the UAE and the Persian Gulf coast.

2.1. Speech recognition data

A first source of data used was training data for dialectal Automatic Speech Recognition (ASR). These are nearly 1,000 hours of dialectal Arabic data in Nadji, Hejazi and Gulf Arabic obtained from different sources. The data is similar to those used to train foundation speech models in that it contains several hundred speakers covering the variability of age, gender and background of the local population. Being datasets intended for ASR, the data was manually transcribed for training.

However, Utilizing ASR data for dialectal Arabic presents formidable challenges. Arabic TTS depends heavily on diacritics for precise phoneme generation. However, the current scarcity of diacritic data results in inaccuracies, leading to suboptimal TTS models, especially for dialects lacking sufficient diacritic information. Furthermore, the original data low sampling rate of 16 kHz falls short of the quality standards required for many TTS applications, further hindering model development. Additionally, the conversational nature of dialectal Arabic recordings introduces various complexities, including dysfluencies, vocalic and non-vocalic noises, and other conversational artifacts, which pose additional obstacles to building effective TTS models tailored to dialectal Arabic. These multifaceted challenges underscore the considerable difficulty in developing robust dialectal Arabic TTS systems.

2.2. Found data

The second source studied was found data. TTS development can usually use this type of resources, with audiobooks being the most common one, like LibriTTS [4]. Audiobooks are well suited for TTS because they contain carefully read speech and the full text is available. Once the audio and the text have been aligned, it is fairly simple to generate a large amount of training data from such sources.

Dialectal Arabic, however, cannot rely on audiobooks for TTS, since MSA is the main language used for Arabic literature globally. Hence, it is necessary to rely on other audio sources like lectures or podcasts. These resources are most complex to deal with for TTS. First, there is no text to fall back on, and ASR outputs are the closest to transcripts available. Second, these sources are less formal than audiobooks and can present background noises, dysfluencies, multiple speakers, and a casual speaking style. Thus, data selection is highly important; segments with the best ASR transcript have to be selected, background noise has to be removed, as well as laughter and other sounds. Using this approach, up to 1,000 hours of data were collected for this purpose from more than 15 different sources.

3. Conclusion

With several thousand hours of dialectal Arabic data, the proposed use of non–conventional TTS data is the most complete attempt to develop a truly dialectal Arabic TTS for the Nadji, Hejazi and Gulf dialects. Despite the complexities of dealing with such heterogeneous sources of data, the current model is able to generate natural sounding dialectal Arabic from multiple voices. In the future, this model can be used as a foundation model for few–shot synthesis of a new dialectal Arabic voice.

4. References

[1] A. Radford and J.–W. Kim and T. Xu and G. Brockman and C. McLeavey and I. Sutskever, "Robust Speech Recognition via Large–Scale Weak Supervision" , *arXiv:2212.04356*, 2022.

[2] H. Touvron et al., "LLaMA: Open and Efficient Foundation Language Models" *arXiv:2302.13971*, 2023.

[3] J. Betker et al., "Better speech synthesis through scaling" *arXiv:2305.07243v2*, 2023.

[4] H. Zen et al., "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech" *arXiv:1904.02882*, 2019.

Using speech graph analysis to differentiate people at clinical high risk of psychosis from healthy controls

Xinyi Liang[1]

[1]King's College London, UK
xinyi.liang@kcl.ac.uk

Background

Psychosis is a severe mental health disorder that affects 2.62% of the population [1]. Formal thought disorder is one of the core symptoms of psychosis, manifesting in the form of disorganised speech [2]. Various studies have indicated that disorganised speech might be a predictive factor of future transition to psychosis in clinical high-risk (CHR) populations [3]. Speech graph analysis is one of the automated approaches to quantify disorganised speech and has gained promising results [4]. However, most research exploring the relationship between speech markers and psychosis has been constrained by small sample sizes. Advances in online assessments and screening combined with automated transcription techniques have facilitated the inclusion and analyses of large-scale samples.

Objectives

The present study aims to investigate whether speech graph analysis can effectively discriminate individuals at CHR of psychosis from controls using a modest to large sample.

Methods

This ongoing study recruited participants from the general population via an online website. Participants were required to complete the prodromal questionnaire (PQ-16) at the screening phase. Participants who scored above threshold of PQ-16 were invited for The Comprehensive Assessment of At-Risk Mental States (CAARMS) interview which is currently the gold standard tool for identifying CHR populations. Conversely, participants who scored below threshold of PQ-16 were selectively invited to the interview to constitute a robust control group. Thus, the present study comprises 4 groups: CAARMS+ group (identified as "at risk" by interview + high PQ-16 score), CAARMS- group (identified as "no risk" + high PQ-16 sore), a robust control group (identified as "no risk" + low PQ-16 score) and a no-interview control group (low PQ-16 score but not invited to interview). At present, this study has 82 participants in CAARMS+ group, 181 in CAARMS- group, 65 in solid control group and 101 participants in no-interview control group. To generate excerpts of speech, all participants were asked to describe 8 ambiguous images from the thematic apperception test for one minute each. Then each excerpt was converted into a graph using SpeechGraphs software, such that each word corresponded to a node and the temporal links between words were represented by edges (see example Figure 1.). The number of nodes in the largest connected component (LCC) and the number of nodes in the largest strongly connected component (LSC) are two important speech graph connectedness measures which can represent how disorganised speech is.

Results

Preliminary results were done using automated transcriptions. Kruskal-Wallis tests indicate that there were no significant differences in LSC and LCC across the 4 groups. Group differences in speech graph connectedness were examined using analysis of covariance, while controlling for the effects of age, gender, ethnicity, education and whether English is participant' first language. However, results indicate that there were no significant differences between 4 groups in terms of LSC and LCC. Further analyses using manual transcriptions are still in progress.

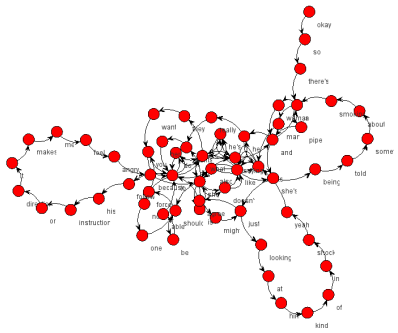


Figure 1: *Example speech graph from a participant describing one of the pictures from the speech task*

1. References

[1] T. Binbay, K. Alptekin, H. Elbi, N. Zağlı, M. Drukker, F. A. Tanık, F. Özkınay, H. Onay, and J. Van Os, "Lifetime prevalence and correlates of schizophrenia and disorders with psychotic symptoms in the general population of izmir, turkey." *Turk Psikiyatri Dergisi*, vol. 23, no. 3, 2012.

[2] A. Ayer, B. Yalınçetin, E. Aydınlı, Ş. Sevilmiş, H. Ulaş, T. Binbay, B. B. Akdede, and K. Alptekin, "Formal thought disorder in first-episode psychosis," *Comprehensive Psychiatry*, vol. 70, pp. 209–215, 2016.

[3] A. Demjaha, S. Weinstein, D. Stahl, F. Day, L. Valmaggia, G. Rutigliano, A. De Micheli, P. Fusar-Poli, and P. McGuire, "Formal thought disorder in people at ultra-high risk of psychosis," *BJPsych open*, vol. 3, no. 4, pp. 165–170, 2017.

[4] T. J. Spencer, B. Thompson, D. Oliver, K. Diederen, A. Demjaha, S. Weinstein, S. E. Morgan, F. Day, L. Valmaggia, G. Rutigliano *et al.*, "Lower speech connectedness linked to incidence of psychosis in people at clinical high risk," *Schizophrenia research*, vol. 228, pp. 493–501, 2021.

Improving Retrieval-Augmented Response Generation in Goal-Oriented Dialogue Question Answering

Norbert Braunschweiler[1], *Abigail Sticha*[2], *Rama Doddipatla*[1], *Kate Knill*[2]

[1]Toshiba Europe Ltd, UK
[2]Cambridge University, UK

norbert.braunschweiler@toshiba.eu, abbysticha@gmail.com, rama.doddipatla@toshiba.eu, kmk1001@cam.ac.uk

Goal-oriented dialogue systems, such as assistant chatbots and conversational AI systems, have gained prominence for their question-answering capabilities, often utilizing large language models (LLMs) as knowledge bases. However, these systems face limitations when knowledge outside their intrinsic scope is required. We address these limitations by designing more faithful systems that can accurately respond to users based on external information. Guided by reference-free evaluation metrics, we present two novel methods to prompt LLMs which surpass the baselines in accuracy, linguistic quality, and faithfulness. The first method employs a reranking technique using LLMs to rank document relevance without the need for fine-tuning. The second system builds upon the *ReAct* framework by incorporating a self-reflection mechanism, ensuring answers are grounded in retrieved content. Overall, our methods advance few-shot prompting as a way to learn to condition on external evidence, and significantly reduce hallucinations.

Benchmark corpus: ***MultiDoc2Dial*** [1] requires the system to both take into account *dialogue history* and read *multiple documents* to generate knowledge-grounded agent responses; introduces a *RAG* (Retrieval-Augmented Generation) benchmark model. This allows us to compare our prompt-tuned GPT-3 models to a fine-tuned BERT/BART architecture.

Method: We employ *LlamaIndex* [2], a framework for LLM applications to ingest, structure, and access domain-specific data. We boost retrieval performance using *LLM Rerank* which uses an LLM to re-order nodes by asking it to return relevant documents and a score of how relevant they are, and returns the top N-ranked nodes. We improve response generation by utilizing the **ReAct** [3] method which is an agent-based chat mode built on top of the *LlamaIndex* query engine. For each chat interaction in our implementation, the agent enters a **ReAct** loop where it has the option to *reason*, *act* which in our case entails querying the index, or produce a *final response*. The **ReAct** framework has the flexibility to coerce the system to take specific actions at different points in the chain. We use this to guide the chain to begin by using the Query Engine, but coerce the chain to reground its answer when the final answer is below a given heuristic. We call this model **ReAct+Reground** illustrated in Figure 1: when the *ReAct* chain decides on an answer, an evaluator module is used to evaluate the response based on a given heuristic. If the heuristic is above a chosen threshold X, the answer is used as final response, but if it is <X, the evaluator returns linguistic feedback to the LLM which is used by the *ReAct* agent to decide on the next reasoning/action step.

As heuristic we use the *groundedness* score of the multi-dimensional evaluation framework of [4] called **UniEval**. This framework provides a unified *multi-dimensional* evaluator outputting scores for *naturalness*, *coherence*, *engagingness*, *understandability* and *groundedness* as well as their single-dimensional evaluator for *factual consistency*. While the first four dimensions evaluate the linguistic quality of generated responses, *groundedness* and *factual consistency* are used as faithfulness metrics.

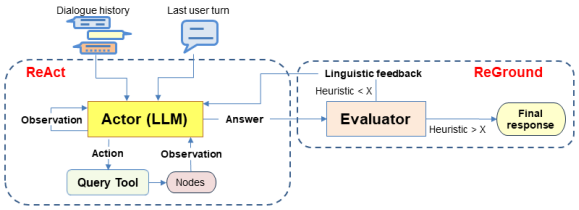


Figure 1: *ReAct+Reground framework.*

We test two versions of our proposed approach which each use different processes for providing **linguistic feedback**.

Results in Table 1 show that the proposed methods *ReAct1 & ReAct2* both improve retrieval performance (recall@1) and increase linguistic quality, accuracy, and adequacy of incorporating retrieved information (measured by *UniEval* scores).

Table 1: *Comparison of retrieval and* UniEval *scores.*

System	Retriev r@1	Ling. Quality nat	coh	eng	und	Faithfulness grnd	fact
Gold	NA	83.0	92.0	1.1	82.3	95.2	68.4
Baseline	50.6	83.8	95.7	1.2	83.1	87.1	67.7
GPT-3 based models							
ReAct1	51.2	88.7	99.3	2.7	89.6	98.0	**73.0**
ReAct2	59.4	89.4	**99.6**	3.0	90.6	**98.1**	72.9

Conclusion We address limitations in using LLMs as knowledge bases by designing two systems that exceed the Retrieval-Augmented Generation baseline system in terms of accuracy, linguistic quality, and faithfulness.

1. References

[1] S. Feng, S. S. Patel, H. Wan, and S. Joshi, "MultiDoc2Dial: Modeling dialogues grounded in multiple documents," in *Proc. of the 2021 Conf. on Empirical Methods in NLP*, Nov. 2021.

[2] J. Liu, "LlamaIndex," 11 2022. [Online]. Available: https://github.com/jerryjliu/llama_index

[3] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "ReAct: Synergizing Reasoning and Acting in Language Models," *arXiv preprint arXiv:2210.03629*, 2022.

[4] M. Zhong, Y. Liu, D. Yin, Y. Mao, Y. Jiao, P. Liu, C. Zhu, H. Ji, and J. Han, "Towards a Unified Multi-Dimensional Evaluator for Text Generation," *arXiv preprint arXiv:2210.07197*, 2022.

The UKIS 2024 Committee would like to thank ISCA and Interspeech 2023 organisers for kindly providing this template.

Acoustic Knowledge and Impact of Explainable Artificial Intelligence (XAI) on Audio Models Explainability

Marjan Golmaryami and Jennifer Williams

School of Electronics and Computer Science, University of Southampton, UK

m.golmaryami@soton.ac.uk, j.williams@soton.ac.uk

In an era dominated by deep learning, the interpretability of audio classification models remains a critical concern, especially given their increasing application in various tasks such as health care, security and surveillance, smart homes, and speech recognition. This poster explores the field of explainable artificial intelligence (XAI) with a focus on audio data, presenting a comprehensive survey of XAI methodologies tailored for enhancing the interpretability of audio models. Our analysis relies on a twofold approach: firstly, exploring different explanations for interpreting audio models and discussing the necessity of audio explanation, which are vital for validating and understanding model decisions in audio-related tasks such as speech synthesis. Secondly, investigating various XAI methods and their applications in interpreting audio models from different perspectives as well as studying unique challenges posed by audio data. We start with the necessity of interpretability of AI models, moving towards different concepts and categories of explainability in terms of stage, scope, and model dependency. Figure 1 shows an overview of this concepts. We discuss different types of explanations for audio models in detail and investigate why audio models require audio explanation [1].

Then we talk about different XAI methods including but not limited to partial dependence plot (PDP), local interpretable model-agnostic explanations (LIME), Deep Learning important feaTures (DeepLIFT), layer-wise relevance propagation (LRP), and SHapley Additive exPlanations (SHAP) followed by focusing on those methods that are used for audio models interpretation. By exploring XAI techniques designed specifically for understanding audio models, we will address challenges that are specific to audio data and models[1, 2]. Then we delve into the effects of different data representations on model performance, such as transformation of raw audio signals to more informative formats and frequency band impacts [3, 4]. By leveraging advanced XAI methods, we interpret these models from multiple aspects to uncover how they process and analyse complex audio inputs. Our work is significant in that it not only clarifies the internal mechanisms of audio classifiers but also demonstrates how XAI can be employed to ensure that these models are reliable, effective and understandable.

Through a systematic exploration of current literature and methodologies, including recent findings in tailored methods for interpreting models in specific scenarios, this poster aims to provide further discussion on the integration of XAI into audio processing frameworks. We aim to review XAI methods applied to different models and application fields with a focus on speech related problems. It highlights the ongoing need for models that have high performance as well as transparency and accountability, paving the way for safer and more reliable audio analysis technologies. This work is still in progress, however Table 1 shows some papers that were investigated in this study.

This work was supported by the Southampton University MINDS CDT.

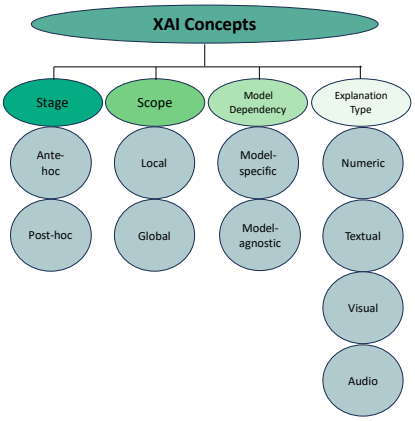


Figure 1: *Figure 1. An overview of different concepts used in classifying XAI*

Table 1: *Table 1: Publications that have implemented XAI methods to enhance audio model explainability*

Ref	XAI method(s)	Applications
[3]	DFT-LRP	Audio event detection
[4]	LIME, SHAP, GradCAM	Fake speech detection
[5]	LRP	Audio classification

1. References

[1] A. Akman and B. W. Schuller, "Audio explainable artificial intelligence: A review," *Intelligent Computing*, vol. 2, p. 0074, 2024.

[2] C. Molnar, *Interpretable machine learning*. Lulu. com, 2020.

[3] A. Frommholz, F. Seipel, S. Lapuschkin, W. Samek, and J. Vielhaben, "Xai-based comparison of audio event classifiers with different input representations," in *Proceedings of the 20th International Conference on Content-based Multimedia Indexing*, 2023, pp. 126–132.

[4] A. Govindu, P. Kale, A. Hullur, A. Gurav, and P. Godse, "Deepfake audio detection and justification with explainable artificial intelligence (XAI)," 2023.

[5] S. Becker, J. Vielhaben, M. Ackermann, K.-R. Müller, S. Lapuschkin, and W. Samek, "Audiomnist: Exploring explainable artificial intelligence for audio analysis on a simple benchmark," *Journal of the Franklin Institute*, vol. 361, no. 1, pp. 418–428, 2024.

Accounting for vocal mismatch within an automatic speaker recognition system

Tallulah Buckley[1], *Vincent Hughes*[1], *Paul Foulkes*[1], *Philip Harrison*[1], *Jessica Wormald*[1], *Finnian Kelly*[2], *David van der Vloed*[3]

[1]Department of Language and Linguistic Science, University of York, UK.
[2]Oxford Wave Research, Oxford, UK.
[3]Netherlands Forensic Institute, The Hague, Netherlands.

{tb1267|vincent.hughes|paul.foulkes|philip.harrison|Jessica.wormald}@york.ac.uk,
finnian@oxfordwaveresearch.com,
d.van.der.vloed@nfi.nl

The performance of Automatic Speaker Recognition (ASR) systems is improving with forensically realistic data [1]. However, mismatches between speech samples of the kind often found in forensic casework remain a considerable challenge for ASR systems. While previous research has tended to focus on sources of technical mismatch, only a small amount of work has considered sources of vocal mismatch. [2] examined the effects of speaker variation on the performance of an x-vector ASR system. They demonstrated that mismatches involving whispered speech had the most detrimental effect on performance, both in terms of discrimination and calibration error. [3] explored the use of whispered speech samples at the calibration stage in order to minimise calibration error. In the present study, we replicate and extend previous work to explore techniques for improving performance with samples of whispered speech through adaptation at the LDA (embedding dimension reduction) and PLDA (scoring) stages of an x-vector ASR system, as well as testing with tailored calibration data.

All of our testing was conducted using VOCALISE 2021 version 3.0.0.1746; a commercial x-vector PLDA ASR system produced by Oxford Wave Research, which has been used around the world in forensic casework. We used a small but highly controlled test set collected as part of [2]. The set consists of six male phoneticians recorded in an anechoic chamber on three separate occasions. Participants read an excerpt from The Rainbow Passage multiple times in a variety of vocal conditions, although for the purposes of the present study we focus on the modal and whispered samples recorded via a head-band microphone. LDA/PLDA adaptation data were taken from the CHAINS corpus [4], which contains cross-session samples of modal and whispered speech from 20 male speakers and is well-matched to the test set for style, technical conditions, and speaker sex. Same- and different-speaker test comparisons were conducted (a) without adaptation, (b) with modal-modal adaptation data (*cross-session whisper-whisper adaptation data was not available in CHAINS), and (c) with modal-whisper adaptation data. This produced sets of scores which were then calibrated initially using cross-session samples of read speech from 20 male speakers from the DyViS dataset [5], which is also well-matched to the test set. Performance was evaluated in terms of the equal error rate (EER) and log LR cost function (C_{llr}) including calibration loss (C_{llr}^{cal}) and discrimination loss (C_{llr}^{min}).

With the default adaptation settings, there was a marginal improvement in EER when using modal-modal adaptation data compared with no adaptation. However, using modal-whisper adaptation data, reflective of the mismatch in the test data, did not improve discrimination. No further improvements in performance were found when increasing the adaptation weight. This may be due to the fact that the adaptation data is relatively small. In this paper, we will expand on this initial testing to explore the effects of using a larger set of whisper data for both adaptation and calibration data (namely the Whisper TIMIT data set). We will also explore techniques for simulating data to increase sample size where available datasets are small, as is likely to be the case in forensics.

1. References

[1] Morrison, G. S. and Enzinger, E. (2019) Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (*forensic_eval_01*) – Conclusion. Speech Communication 112: 37-39

[2] Hughes, V., Wormald, J. H., Foulkes, P., Harrison, P., Kelly, F., van der Vloed, D., Welch, P., & Xu, C. (2023). Automatic speaker recognition with variation across vocal conditions: a controlled experiment with implications for forensics. In Proceedings of Interspeech 2023, 591-595 https://doi.org/10.21437/Interspeech.2023-443

[3] Kelly, F., & Hansen, J. H. L. (2021). Analysis and Calibration of Lombard Effect and Whisper for Speaker Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 927–942.

[4] F. Cummins, M. Grimaldi, T. Leonard, and J. Simko. (2006). The CHAINS corpus: CHAracterizing INdividual Speakers (pp. 431–435). SPECOM.

[5] Nolan, F., McDougall, K., Jong, G. D., & Hudson, T. (2009). The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. International Journal of Speech Language and The Law, 16, 31–57.

Evaluating the Effectiveness of the Conformer ASR Model in Handling Spanish Morphology

Mourbi Basak

*University of Edinburgh, United Kingdom
m.basak@sms.ed.ac.uk, mourbibasak@gmail.com

Abstract

Despite advancements in state-of-the-art speech recognition models, there is a lack of research concerning their generalisation capabilities to instances within languages characterised by morphological productivity. Rare or out-of-vocabulary (OOV) words have received limited attention in recent end-to-end models, as they do not significantly increase the overall word error rate (WER). However, this study aims to evaluate the performance of a Conformer model fine-tuned on over 400 hours of Spanish data, specifically targeting the accurate transcription of challenging phrases. To achieve this, we designed two distinct algorithms for sentence selection from our test sets to assess any potential degradation in model performance. Through a comparative analysis involving 2,000 sentences in each test set, we observed a WER degradation of up to 10% absolute value compared to our pre-split test set, highlighting the model's susceptibility to morphologically challenging language constructs.

1. Methodology

We create datasets for model testing using two different algorithms. The first algorithm (Algorithm A), utilises an open-source toolkit to select morphologically complex sentences from our test set. For the second algorithm (Algorithm B), linguistic rules of Spanish morphology were employed to devise an algorithm for selecting sentences to create our test sets. Our goal is to choose sentences with maximum inflections.

2. Datasets and Model

We run the two algorithms, A and B, on two different base datasets. The first dataset comprises the test sets of the datasets used to train our model: Common Voice Spanish, Fisher Spanish, and Media Speech. The second dataset is an out-of-domain dataset from the TEDx Spanish corpus, which was not used to train the model. The rationale behind using the second dataset was to evaluate the model's performance on unseen spontaneous speech data.

We implement continuous fine-tuning of a large Conformer CTC model with 121M parameters from NVIDIA NeMo pre-trained models. The model is fine-tuned on 416 hours of data.

3. Test sets

After implementing both algorithms on two of our test sets, we get 4 test sets with 2000 entries each. The test sets are then used to evaluate the performance of our model. The 4 test set names with descriptions are stated in table 1

Test set names	Desciption
base	the pre-split test set
A.1	in-domain test set from Algorithm A
A.2	out-of-domain test set from Algorithm A
B.1	in-domain test set from Algorithm B
B.2	out-of-domain test set from Algorithm B

Table 1: *List of all the test sets used for evaluation of the model.*

4. Results

Both greedy search and beam search were employed for decoding purposes. We use a 4-gram KenLM with beam search parameters set to a size of 128, an alpha value of 0.8, and a beta value of 1.5. As evident from Figure 1, when the model's performance was assessed on the test set using the proposed algorithm B, the accuracy was found to be inferior compared to algorithm A, which utilises an off-the-shelf tool.

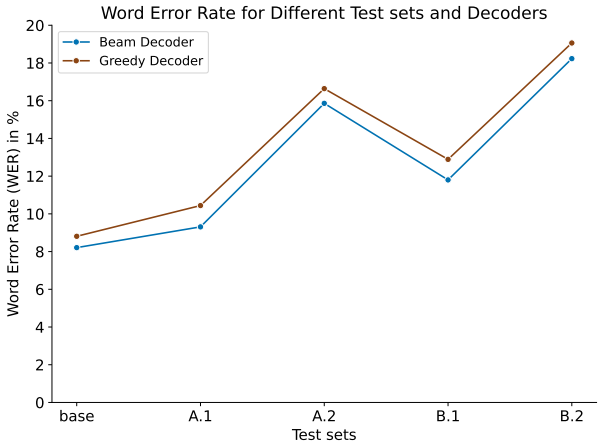


Figure 1: *Accuracy of the model evaluated on five test sets.*

The UKIS 2024 Committee would like to thank ISCA and Interspeech 2023 organisers for kindly providing this template.

*I would like to thank Dialpad for providing me with the compute for the entirety of the project.

Muting Whisper: A Universal Acoustic Adversarial Attack on Speech Foundation Models

Vyas Raina, Rao Ma*, Charles McGhee, Kate Knill, Mark Gales*

University of Cambridge

{vr313, rm2114, cgm43, kmk1001, mjfg100}@cam.ac.uk

Motivation: Recent advances in foundation models like Whisper have significantly enhanced automatic speech recognition (ASR) capabilities. These models often use special language tokens, such as `<start>` and `<end>`, to guide the generation of output sequences. Although these tokens are necessary, this work demonstrates that their use can make ASR systems vulnerable to adversarial attacks. Specifically, we explore how these tokens can be exploited to mute, or silence, an ASR model, thereby preventing it from transcribing audio inputs. This vulnerability poses potential threats in contexts where ASR systems are used to detect harmful content, but conversely it can also be viewed as an approach to protect speech privacy.

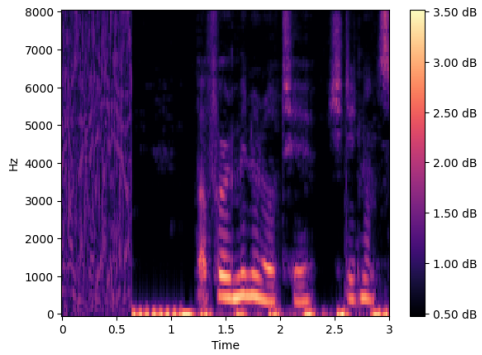


Figure 1: *Mel spectrogram of universal acoustic segment (0.64s) prepended to a random speech sample (truncated to a total length of 3s).*

Method: We developed a universal adversarial audio attack capable of muting Whisper, a popular foundation model for ASR. Our method involves creating a 0.64-second adversarial audio segment that acts as an acoustic realization of the `<endoftext>` token used by Whisper. This segment, when prepended to a speech signal, effectively mutes the model by preventing transcription of the speech. The adversarial segment is trained to be 'universal', meaning it can be applied to any speech signal irrespective of its content. The effectiveness and imperceptibility of the adversarial segment were enhanced by limiting its duration (to 0.64-seconds) and amplitude (to a maximum of 0.02), ensuring that it remains undetected to human listeners and automated systems alike. Figure 1 shows the Mel spectogram for a speech signal where the universal acoustic adversarial segment has been prepended to it, demonstrating the relatively lower power and thus inaudibility of the adversarial segment.

Main Results: Table 1 shows the impact of the universal adversarial segment on the LibriSpeech dataset. Two metrics are used: the percentage of samples that are successfully attacked, ∅, where an attack is defined as the predicted transcription having no generated tokens; and the average sequence length, `asl`, for the predicted transcriptions. Our experiments conducted across eight different Whisper ASR models show that the universal adversarial segment successfully mutes more than 97% of all tested speech samples. In our experiments we also find that the attack segment can transfer across different speech domains and even to different speech processing tasks (specifically speech translation).

Model	Metric	No Attack	Attack
tiny.en	∅ (%) ↑	0.0	99.7
	`asl` ↓	17.9	0.06
tiny	∅ (%) ↑	0.0	99.6
	`asl` ↓	17.9	0.04
base.en	∅ (%) ↑	0.0	99.0
	`asl` ↓	17.8	0.20
base	∅ (%) ↑	0.0	99.5
	`asl` ↓	17.8	0.05
small.en	∅ (%) ↑	0.0	98.6
	`asl` ↓	17.7	0.14
small	∅ (%) ↑	0.0	98.7
	`asl` ↓	17.3	0.15
medium.en	∅ (%) ↑	0.0	99.5
	`asl` ↓	17.7	0.10
medium	∅(%)↑	0.0	97.8
	`asl` ↓	17.8	0.56

Table 1: *The percentage of successfully 'muted' speech samples, ∅, and the Average Sequence Length (`asl`) of transcriptions. Results are presented for no attack, and for a trained (per model) universal acoustic adversarial attack, where the same universal adversarial segment is prepended to each speech sample.*

Implications: The ability to mute Whisper has significant implications. On the one hand, it represents a security risk, potentially allowing malicious actors to circumvent moderation systems designed to filter harmful content. On the other hand, it offers a new method for protecting speech privacy against unauthorized transcription. This dual nature of the technology highlights the need for ongoing research into both the potential and the risks of adversarial attacks on ASR systems. As these systems become more ingrained in our digital infrastructure, ensuring their resilience against such attacks becomes crucial to maintaining trust and functionality in deployed ASR systems.

Zipfian Properties of English and Finnish Documents

Martin Tunnicliffe, Gordon Hunter

Kingston University
M.J.Tunnicliffe@kingston.ac.uk, G.Hunter@kingston.ac.uk

English and Finnish belong to two very different language types: while English is largely noninflected with limited agglutination, Finnish is highly agglutinative and inflected with a high ratio of wordforms to lemmas. Here we show some more subtle differences between the two languages relating to Zipf's statistical type-token laws.

Zipf's first law $f(r) \propto 1/r^{\alpha}$ relates the frequency f of a word type to its rank r, while the "second law" $n(f) \propto 1/f^{\beta}$ describes the number of word types exhibiting each frequency. The indices α and β are approximately constant. These trends were investigated using items from the Standardized Project Gutenberg Corpus (SPGC), trimmed to 100,000 words each to eliminate size variation as noise. Figure 1 shows n vs. f graphs for typical English and Finnish texts; while the second law holds approximately for lower frequencies, the graph is log-log convex for English and log-log concave for Finnish. This holds for approximately 95% of English and Finnish texts.

Although the second law is said to be a "straightforward consequence" of the first [1] (with $\beta = 1 + 1/\alpha$), this relies on a continuum approximation with Bernoulli word selection and unlimited word types. Furthermore, discrete analysis shows the two laws are not even mutually consistent [2, 3]. If the first law is universally correct then the "true" second law should be

$$n(f) = \frac{1}{\alpha f!}\left(\frac{t}{\zeta(\alpha)}\right)^{\frac{1}{\alpha}} \Gamma\left(f - \frac{1}{\alpha}\right) \qquad (1)$$

where t is the document size in word tokens and $\zeta(\alpha) = \sum_{i=1\ldots\infty} 1/i^{\alpha}$. (Though rarely seen in the literature, this expression dates back at least to 1961, when it appeared in a paper by Mandelbrot [4].) While for large f it aligns asymptotically with the "classic" second law, the effective β rises with decreasing f, creating a slight log-log concavity. With a finite maximum vocabulary V the equation becomes [3]

$$n(f) = \frac{1}{\alpha f!}\left(\frac{t}{\zeta_V(\alpha)}\right)^{\frac{1}{\alpha}} \Gamma\left(f - \frac{1}{\alpha}, \frac{t}{\zeta_V(\alpha) V^{\alpha}}\right) \qquad (2)$$

where $\zeta_N(\alpha) = \sum_{i=1\ldots N} 1/i^{\alpha}$. With a suitably chosen V, $n(f)$ acquires a log-log convexity, suggesting that (1) best describes the Finnish document and (2) the English.

Figure 2 shows low frequency β ($1 \le f \le 10$) plotted against mid-range α ($100 \le r \le 1000$), revealing distinct English and Finnish clusters. Adherence to Zipf's first law within this mid-range was assessed using the Kolmogorov-Smirnov significance, which was much closer to 1 for the Finnish than the English items.

A few additional observations can be made. Firstly, despite the aforementioned $\beta = 1 + 1/\alpha$ rule, a positive correlation exists within each language between the low-frequency β and mid-range α (though for higher frequency β the law has been found to hold [3]). Secondly, the bulk of the English items form an approximately elliptical cluster which may be amenable to

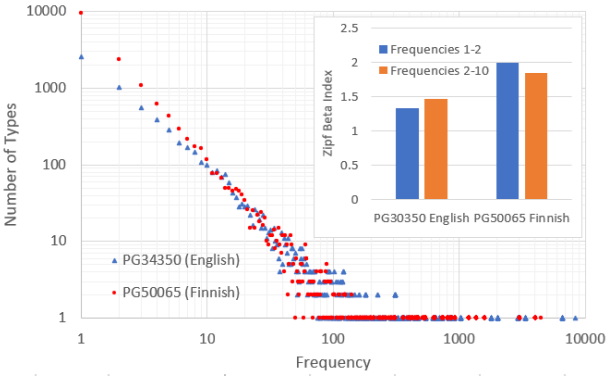


Figure 1: *Types vs. frequency plot for typical English and Finnish documents*

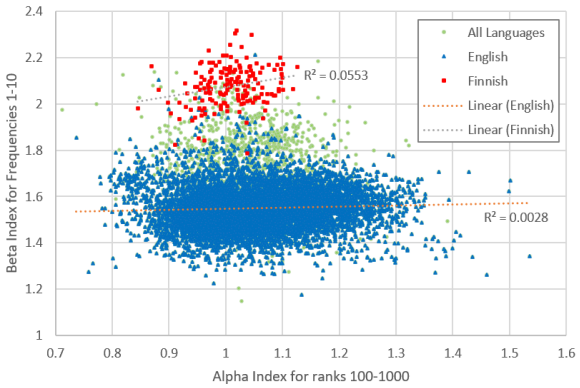


Figure 2: *Alpha vs. beta plot for the Standardized Project Gutenberg Corpus*

principal component analysis. The Finnish cluster has both a steeper slope and a higher coefficient of correlation than the English cluster, though at present the sample size is insufficient to reveal much about its precise shape. Other important topics still to be examined are the effects of document length (currently set to 100,000 word instances), and the scaling of the statistics of English and Finnish texts.

References

[1] A. Kornai, "Zipf's Law Outside the Middle Range," in 6th Meeting on Mathematics of Language, University of Central Florida, 1999.
[2] Á. Corral, I. Serra, Ferrer-i-Cancho, R, "Disstinct Flavours of Zipf's Law and its Maximum Likelihood Fitting: Rank-Size and Size-Ddistribution Representations," *Phys. Rev. E.*, vol. 102, 052113, 2020.
[3] M. Tunnicliffe, G. Hunter, "Random Sampling of the Zipf-Mandelbrot Distribution as a Representation of Vocabulary Growth", Physica A, vol. 608, 128259, 2022.
[4] B. Mandelbrot, "On the Theory of Word Frequencies and on Related Markovian Models of Discourse," in *Structure of Language and its Mathematical Aspects, R. Jakobson (Editor)*, Providence, Rhode Island, American Mathematical Society, 1961, pp. 190-219.

Low-resource Speech Recognition and Dialect Identification of Irish in a Multi-task Framework

[1]*Mengjie Qian,* [2]*Liam Lonergan,* [2]*Neasa Ní Chiaráin,* [2]*Christer Gobl,* [2]*Ailbhe Ní Chasaide*

[1]Engineering Department, Cambridge University, UK
[2]Phonetics and Speech Laboratory, School of Linguistics, Speech and Communication Sciences, Trinity College Dublin

[1]mq227@cam.ac.uk, [2]{llonerga, nichiarn, cegobl, anichsid}@tcd.ie

Irish (Gaelic) is a highly inflected, minority language indigenous to the island of Ireland, which exhibits significant dialectal variation, including Ulster (Ul), Connaught (Co), and Munster (Mu). Developing effective speech technology for Irish is challenging due to the limited availability of data and the lack of a standardised spoken variety. As part of the ABAIR initiative, which is developing speech technologies for the Irish language, a TDNN-HMM ASR system *Éist* [1] has been developed and is available for public use[1]. This ASR system addresses some of these challenges, but dialect bias remains an issue. This study explores a novel approach using Hybrid CTC/Attention encoder-decoder models [2], enhanced with Intermediate CTC (InterCTC), to improve automatic speech recognition (ASR) and dialect identification (DID) for Irish.

The dataset for this study includes recordings from the MíleGlór and Synthesis corpora, as well as audiobooks and spontaneous speech from the Corpas na Cainte Beo. The training set consists of 290 hours of audio data, with validation and test sets constructed to ensure no overlap in speakers or texts. The Hybrid CTC/Attention model leverages both CTC and attention mechanisms, optimising a combination of losses to generate accurate predictions. InterCTC applies CTC to intermediate encoder layers to support multi-task learning, explicitly capturing dialects by prepending dialect tags to the text output. Figure 1 illustrates the architecture.

For experiments, a previously trained hybrid TDNN-HMM model served as the ASR baseline, and an ECAPA-TDNN model was used as the DID baseline. We tested two innovative encoder architectures for the end-to-end ASR model: Conformer and E-branchformer, using XLS-R 300M as the front-end module for acoustic feature extraction paired with a transformer decoder. Various InterCTC configurations were evaluated for joint ASR and DID to optimize performance. The optimal InterCTC settings improved DID accuracy by 10.8% over the ECAPA-TDNN baseline and achieved ASR performance approaching that of the TDNN-HMM model. Experiments comparing the Conformer and E-branchformer encoders show that E-branchformer Small yields the highest DID accuracy and E-branchformer Large shows the best ASR performance. Applying LM shallow fusion to the multitask trained model can improve both the DID accuracy and ASR performance. Results for the baseline models and multitask models are presented in Table 1.

In conclusion, this multi-task approach using Hybrid CTC/Attention-based encoder-decoder models with InterCTC demonstrates significant improvements in DID accuracy and ASR performance. The inclusion of LM shallow fusion further enhances these results, making this a promising strategy for addressing the challenges of low-resource ASR and DID for Irish.

[1]https://abair.ie

Table 1: *DID & ASR performance for Conformer and E-branchformer encoders and multi-task LM shallow fusion.*

Task	Model	DID	WER			
		Acc	All	Ul	Co	Mu
ASR	TDNN-HMM	-	13.2	18.0	15.3	6.2
DID	ECAPA-TDNN	72.5	-	-	-	-
Multitask	Conformer	80.3	16.7	16.4	18.2	15.2
	E-Branchformer-S	**81.5**	17.7	17.6	19.1	16.3
	E-Branchformer-L	80.0	**15.7**	**15.5**	**17.0**	**14.5**
	+ LM	80.8	**13.5**	**13.0**	**14.9**	**12.3**

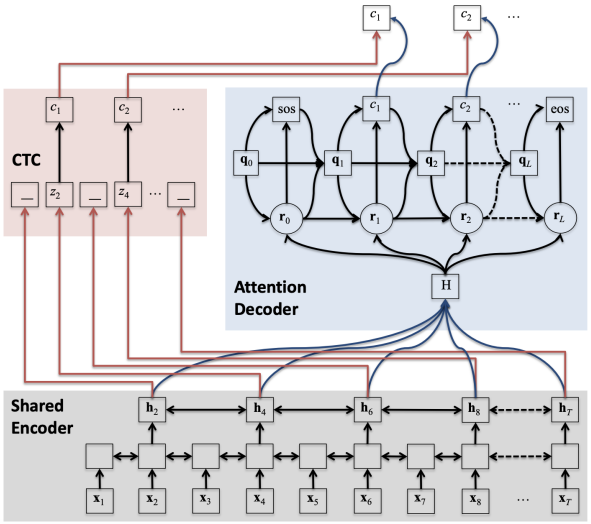


Figure 1: *Hybrid CTC/attention-based end-to-end architecture. The shared encoder is trained by both CTC and attention model objectives simultaneously.*

1. References

[1] L. Lonergan, M. Qian, H. Berthelsen, A. Murphy, C. Wendler, N. N. Chiaráin, C. Gobl, and A. N. Chasaide, "Automatic speech recognition for irish: the abair-éist system," in *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, 2022, pp. 47–51.

[2] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

Technical challenges and opportunities in building an Irish-language augmentative and alternative communication (AAC) system

Rían Errity[1], Emily Barnes[2], Julie Cummins[1], Oisín Morrin[1], Harald Berthelsen[1], Christoph Wendler[1], Andy Murphy[1], Christer Gobl[1], Neasa Ní Chiaráin[1], Ailbhe Ní Chasaide[1]

[1]Phonetics & Speech Laboratory, School of Linguistic, Speech and Communication Sciences, Trinity College Dublin, Ireland
[2]School of Education, Trinity College Dublin, Ireland
Rian.Errity@tcd.ie, ebarnes@tcd.ie, anichsid@tcd.ie

1. Introduction

While building an Irish-language Alternative and Augmentative Communication (AAC) system (Geabaire) from scratch brings many challenges, it presents the opportunity to satisfy both the urgent need for a usable system in the short-term and the need for a platform that can be developed and further enhanced in the long-term. The key design features underpinning the current system and enabling future development are briefly outlined. They cover (i) platform and software components, (ii) board layout that reflects the linguistic structure of Irish, (iii) speech output, adapted to current and future user requirements, and (iv) collection of user data to guide future research and development.

2. Key Design Features

(i) Platform and Software: React Native was chosen to ensure Geabaire's availability across various platforms (specifically iOS, Android, and web) with a single codebase developed in JavaScript. Figure 1 shows the structure of the various software components. The platform is modular in design, allowing for ease of iterative development. To improve compatibility with accessibility hardware, the platform will be further developed for desktop.

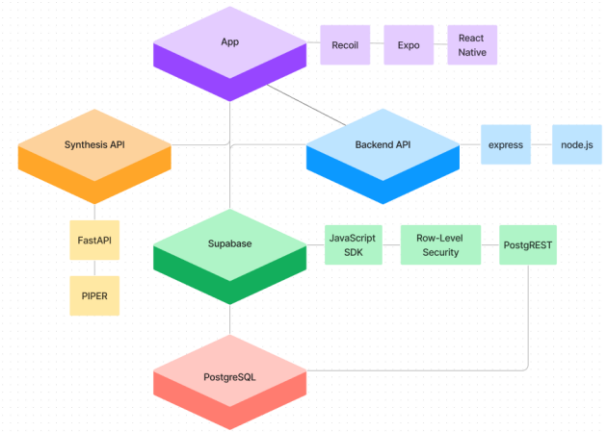


Figure 1: *Software architecture diagram of Geabaire.*

(ii) Board Layout: The board layout has been developed to reflect the morphophonological and syntactic systems of Irish, requiring a radically different design from existing English AAC implementations.



Figure 2: *Geabaire's user interface.*

(iii) Speech Output: Geabaire uses the synthetic voices for the main Irish dialects, developed in the ABAIR initiative and available at https://abair.ie. The current system uses ABAIR's API for synthesis, which requires an active internet connection. Implementing offline and on-device synthesis is ongoing and requires complex interaction with native APIs. The three main dialects of Irish are currently available, with pitch and speed controls. Future plans focus on the development of children's voices as well as bilingual (Irish-English) synthesis and eventually accommodating codeswitching. Parallel research is also being directed at the incorporation of affective prosody (the RóbóGlór project).

(iv) Data Collection: User data is collected to allow for research on AAC acquisition and to guide future development and personalisation of the platform.

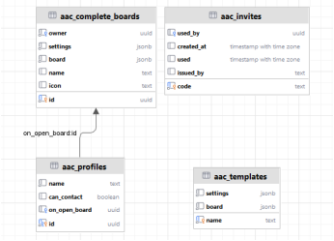


Figure 3: *Geabaire's SQL Schema.*

At the heart of development is the input from the community of users, their carers, therapists, teachers. The system is currently undergoing user evaluation.

3. Acknowledgements

This research is supported by An Chomhairle um Oideachas Gaeltachta agus Gaelscolaíochta (AAC don Ghaeilge project) and An Roinn Turasóireachta, Cultúir, Ealaíon, Gaeltachta, Spóirt agus Meán (ABAIR & RóbóGlór projects), as part of *Stráitéis 20 bliain don Ghaeilge, 2010-2030.*


UKIS 2024
Cambridge