2026-9-26

The 12th ISCA-SAC Doctoral Consortium

Handbook

Tianhua Qi, Fabian Ritter-Gutierrez,
Tina Raissi, Spyretta Leivaditi,
João Vitor Possamai de Menezes

ISCA-SAC's 12th Doctoral Consortium - Program		
Room	**Room 1 (Gallery 1)**	**Room 2 (Gallery 2)**
9:30-9:40	**Welcome Message at Gallery 1 (10 min)**	
9:40-10:15	Zhaofeng Lin: *Towards Agile and Generalisable Audio-Visual Speech Recognition*	Kemal Altwlkany: *Enhancing Voice Communication Services with Audio Fingerprinting Techniques*
10:15-10:50	Poppy Welch: *Privacy Considerations for Audio-Visual Speech Enhancement*	Baptiste Ramonda: *Characterization and measurement of listenability for vulnerable individuals*
10:50-11:10	**Coffe Break (20 min)**	
11:10-11:45	Himashi Rathnayake: *Community-Oriented Development of Speech Emotion Recognition Framework for te reo Māori*	Dushanthi Madhushika Manamalage: *Towards Trustworthy Speech-Based Multimodal Depression Detection*
11:45-12:00	**Q&A / Discussion (15 min)**	
12:00-13:00	**Lunch Break (60 min)**	
13:00-13:35	Rong Li: *Detection and Perception of Smiled Speech*	Shrankhla Pandey: *Computational Measures of Speech and Language for Schizophrenia Spectrum Disorders*
13:35-14:10	Yuxuan Jiang: *Toward a Holistic Modeling Paradigm for Accurate, Versatile, and Controllable Latent Audio Diffusion*	Yang Xiao: *From Representation to Retention: Locating and Protecting Knowledge for Continual Speech Adaptation*
14:10-14:30	**Coffe Break (20 min)**	
14:30-15:05	Shreeram Suresh Chandra: *Natural Language as the Interface to Speaking Style: Describing and Controlling How Speech Is Delivered*	Debasish Ray Mohapatra: *Physics-Based Simulation of Vowel Utterances using a Biomechanical-Acoustic Model*
15:05-15:35	**Panel Discussion (30 min) at Gallery 1**	
15:35-16:00	**Closing Message at Gallery 1**	

Zhaofeng Lin

Trinity College Dublin, Ireland

Title

Towards Agile and Generalisable Audio-Visual Speech Recognition

Self-Introduction

I'm a final year PhD student working on multimodal speech recognition

PhD Project Source / Funding

This PhD project is supported by Research Ireland on multimodal speech recognition

Towards Adaptive and Generalisable Audio-Visual Speech Recognition

Zhaofeng Lin

Trinity College Dublin, Ireland

linzh@tcd.ie

1. Background and Motivation

1.1. Audio-Visual Speech Recognition

Audio-Visual Speech Recognition (AVSR) systems exploit visual information from the speaker's face, such as lip movements, to enhance speech recognition. Traditional Automatic Speech Recognition (ASR) systems rely solely on auditory input, making them vulnerable to performance degradation in noisy environments. By incorporating visual information, AVSR systems offer enhanced robustness and reliability, especially under acoustically challenging conditions.

Recent advancements in deep learning, computer vision, and speech technology have facilitated significant progress in AVSR. These developments have led to sophisticated models capable of learning complex multi-modal representations, enabling improved recognition accuracy. However, the extent to which visual cues are effectively utilised by current AVSR systems remains unclear.

Furthermore, the evaluation of these recent, highly advanced architectures relies almost entirely on a single benchmark: LRS3-TED [1]. State-of-the-art methods are aggressively optimised for this specific dataset, and hence we are now witnessing severe performance saturation, with models achieving less than a 1% word error rate (WER) on clean speech and less than 5% under 0 decibels (dB) of noise. Crucially, the LRS3 test set is exceptionally small, comprising only 0.9 hours of video against a 433-hour training set. This imbalance raises a critical question: does this near-perfect performance indicate a genuine ability to recognise audio-visual speech, or does it merely reflect strong overfitting to the LRS3 distribution?

1.2. Research Questions

- **RQ1:** How can insights from human speech perception inform the development and improvement of AVSR systems?
- **RQ2:** How do AVSR systems generalise beyond the traditional benchmark?

2. Results to Date

2.1. Quantifying Visual Information

The effective Signal-to-Noise Ratio (SNR) gain [2, 3] is measured by the difference in SNR at which the AVSR WER equals the reference WER for audio-only recognition at 10 dB. This metric quantifies the benefit of the visual modality in reducing WER compared to audio-only speech recognition across various acoustic channel SNRs. However, with the development of modern neural architectures, systems generally perform well at 10 dB, making SNR gains at this level less effective for assessing performance in noisy environments. Therefore, this paper proposes quantifying SNR gains with reference to the audio-only WER at 0 dB.

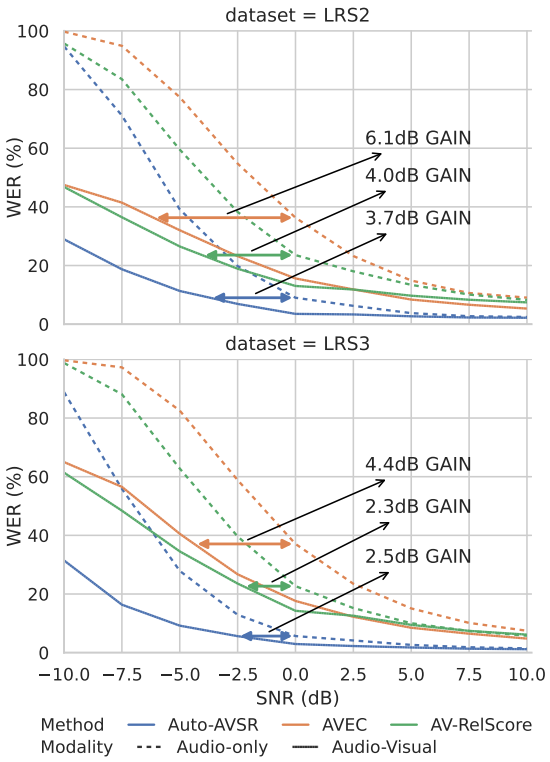


Figure 1: *Performance of six models (three methods, both audio-only and audio-visual) on LRS2 and LRS3 test sets under different levels of pink noise.*

Figure 1 shows the audio-only and audio-visual performance of three methods (Auto-AVSR [4], AVEC [5] and AV-RelScore [6]) across various SNR conditions on LRS2-BBC [7] and LRS3-TED datasets [1]. The effective SNR gains were measured compared to audio-only performance at 0 dB.

As reported in [2], both machines and humans achieved an effective SNR gain of 6 dB when compared to audio-only performance at 10 dB on the IBM ViaVoice audio-visual database. Although all three methods here were tested on LRS2 and LRS3, none reached this level of effective SNR gain when compared at 10 dB.

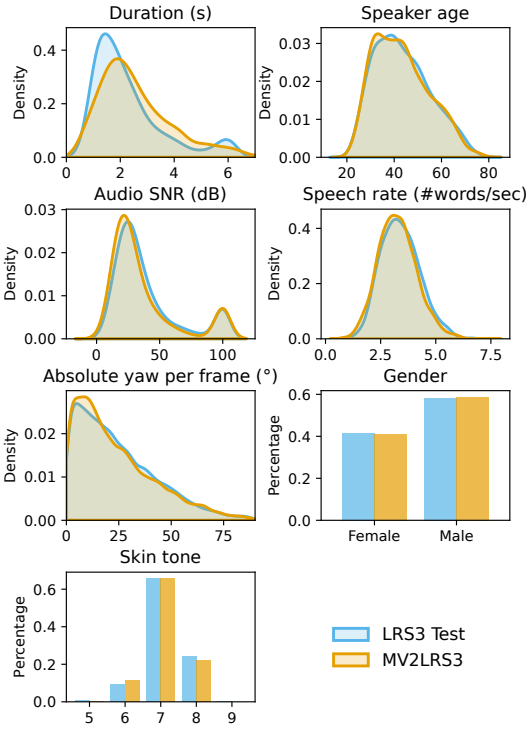


Figure 2: *Distribution of all 7 factors on the LRS3 Test set and MV2LRS3 set.*

2.2. AVSR Generalisability

We construct a new evaluation set strictly matching the demographic, acoustic and visual distributions of the LRS3 test set. We derive this subset from a recently created, massive lip-reading speech dataset, MultiVSR [8]. Specifically, we establish the LRS3 Test set as our reference distribution and utilise MultiVSR as our candidate pool. Our primary objective is to subsample this candidate pool to extract a distributionally aligned evaluation set, which we term the **MultiVSR2LRS3 (MV2LRS3)**. Figure 2 shows the distributions of 7 attributes (duration, age, gender, skin tone, head pose, SNR and speech rate) for both the LRS3 Test set and the MV2LRS3.

We select five state-of-the-art models representing the rapid architectural evolution of AVSR over recent years. These encompass fully-supervised end-to-end models, self-supervised models fine-tuned on LRS3, and the latest architectures integrating speech foundation models and Large Language Models. They are AV-HuBERT [9], Auto-AVSR [4], USR [10], Whisper-Flamingo [11] and Llama-AVSR [12].

Figure 3 shows the relationship between WER on the MV2LRS3 and the LRS3 Test set. We observe that the performance degradation can be approximated by a linear function:

$$WER_{MV2LRS3} = 10.4 \times WER_{LRS3} + 8.1 \qquad (1)$$

The steep slope of 10.4 demonstrates a high degree of sensitivity; small performance variations on the LRS3 benchmark are amplified around tenfold on the MV2LRS3 set. We can compare this to established generalisability literature for image classification and VSR [13, 14], where linear fit slopes typically range between 1.0 and 2.0. In contrast, our observed slope of 10.4 is substantially steeper than theirs.

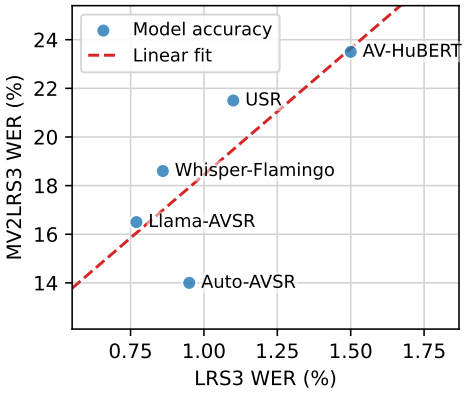


Figure 3: *Model performance on LRS3 Test vs. MV2LRS3 set.*

To understand how these architectures process distinct streams of information, we tested their performance across different modality settings. We evaluated the models on the MV2LRS3 set using audio-visual (AV), audio-only (AO), and video-only (VO) inputs. Table 1 reveals an unexpected trend: for several SoTA models, adding visual data actually harms performance on the MV2LRS3 set. Llama-AVSR, Whisper-Flamingo, and USR all achieve lower WERs in the audio-only setting compared to the audio-visual setting.

Table 1: *WER (%) of the audio-visual, audio-only, and video-only models on the LRS3 Test set and the MV2LRS3 set.*

Model	Unified model	LRS3			MV2LRS3		
		AV	AO	VO	AV	AO	VO
AV-HuBERT	✓	1.5	2.0	34.1	23.5	23.6	65.1
Auto-AVSR		0.95	0.99	19.1	14.0	16.4	45.5
USR	✓	1.1	1.2	22.3	21.5	21.0	57.9
Whisper-F.		0.86	0.85	–	18.6	16.6	–
Llama-AVSR		0.77	0.81	24.0	16.5	15.3	81.5

3. Plans for the Future

Based on previous research, current AVSR systems do not utilise visual information optimally. Specifically, they lack the human ability to adaptively shift focus between modalities based on the environment. In the next stage, I will focus on the following question: How can we train AVSR systems to know when to shift attention between different modalities? The initial step will be to build a baseline system with simple dynamic modality fusion, using an existing model. I will then investigate how much the models actually focus on each modality, both with and without this dynamic fusion.

4. Main Contributions

Our main contributions are as follows:

- Quantify the visual information in end-to-end AVSR systems.
- Conduct the first systematic assessment of generalisability in these AVSR systems.
- Introduce the MV2LRS3 dataset for the community to benchmark AVSR models.

5. References

[1] T. Afouras, J. S. Chung, and A. Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.

[2] G. Potamianos, C. Neti, G. Iyengar, and E. Helmuth, "Large-vocabulary audio-visual speech recognition by machines and humans," in *Seventh European Conference on Speech Communication and Technology*. Citeseer, 2001.

[3] Z. Lin and N. Harte, "Uncovering the visual contribution in audio-visual speech recognition," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] K. R. Prajwal, S. Hegde, and A. Zisserman, "Scaling multilingual visual speech recognition," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[9] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," in *International Conference on Learning Representations*, 2022.

[10] A. Haliassos, R. Mira, H. Chen, Z. Landgraf, S. Petridis, and M. Pantic, "Unified speech recognition: A single model for auditory, visual, and audiovisual inputs," *Advances in Neural Information Processing Systems*, vol. 37, pp. 139 673–139 699, 2024.

[11] A. Rouditchenko, Y. Gong, S. Thomas, L. Karlinsky, H. Kuehne, R. Feris, and J. Glass, "Whisper-Flamingo: Integrating Visual Features into Whisper for Audio-Visual Speech Recognition and Translation," in *Interspeech 2024*, 2024, pp. 2420–2424.

[12] U. Cappellazzo, M. Kim, H. Chen, P. Ma, S. Petridis, D. Falavigna, A. Brutti, and M. Pantic, "Large language models are strong audio-visual speech recognition learners," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[13] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do ImageNet classifiers generalize to ImageNet?" in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 97. PMLR, 09–15 Jun 2019, pp. 5389–5400. [Online]. Available: https://proceedings.mlr.press/v97/recht19a.html

[14] Y. A. D. Djilali, S. Narayan, E. LeBihan, H. Boussaid, E. Almazrouei, and M. Debbah, "Do vsr models generalize beyond lrs3?" in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 6635–6644.

Kemal Altwlkany

Infobip and University of Sarajevo, Bosnia and Herzegovina

Title

Enhancing Voice Communication Services with Audio Fingerprinting Techniques

Self-Introduction

I am Kemal Altwlkany, a Voice AI engineer at Infobip, and part-time teaching assistant at the Faculty of Science, University of Sarajevo. My research area is applicable speech technology in voice communications: fair and secure speech coding and transmission, audio fingerprinting, and speech enhancement. I'm currently about to start the 4th (final) year of my PhD.

PhD Project Source / Funding

PhDs in my country are self-funded and unpaid. About 90% of my tuition fees are government subsidized.

I am full-time employed as an engineer. My company (Infobip) supports my research by financing conference attendances, publication costs, and provides access to computing facilities.

Enhancing Voice Communication Services with Audio Fingerprinting Techniques

Kemal Altwlkany [1,2]

[1] Infobip
[2] Faculty of Science, University of Sarajevo, Bosnia and Herzegovina
[name.surname]@infobip.com

Abstract

Audio fingerprinting techniques have seen widespread adoption in the field of content based music retrieval. This paper provides an overview of an ongoing joint academic and industrial PhD thesis that explores how audio fingerprinting methods, specifically self-supervised neural fingerprinting, can be applied to solve various problems in voice communication services such as enhancing network security, optimizing network performance, and improving the efficiency of data annotation processes for industry-specific speech data.
Index Terms: voice communication, telephony, audio fingerprinting, robocalls, data annotation

1. Introduction

Audio fingerprinting techniques identify an audio file based on its content and are trained to be robust to degradations that can affect the content such as background noise, reverb, lossy compression, altered playback rate, and similar [1, 2, 3, 4, 5, 6]. They have seen widespread adoption in music retrieval and have been successfully applied to song identification, detecting copyrights, deleting duplicated contents, monitoring broadcasts, and tracking advertisements [3]. Even though they are called *audio* fingerprinting methods, applications are usually limited to music, while those that operate in the domain of *speech* have seen limited attention [7, 8, 9].

The doctoral thesis lies in the overlap between academia and industry; it considers the applications of audio fingerprinting (primarily speech fingerprinting) via self-supervised neural fingerprinting methods with the goal of enhancing voice communication services, specifically telephony-based services.

1.1. Research questions

Among others, the thesis addresses the following research questions:

1. Can audio fingerprinting methods trained on music data generalize well to speech retrieval tasks and be successfully applied to downstream speech-related tasks?
2. Can audio fingerprinting methods be used to enhance voice communication services in terms of network security and network performance (reliability)?
3. Can audio fingerprinting methods be used to improve the efficiency of speech data annotation processes?

2. Motivation and Applications

Voice communication is part of everyday modern life. With more than 96% of the global population covered by a mobile broadband connection and more than 9.2 billion mobile-cellular subscriptions in the world, the telephone network remains the largest network for human communication globally [10]. It is therefore important to ensure that the network is reliable, secure, and unbiased for all its users irrespective of their gender, ethnicity, or other demographic characteristics.

2.1. Enhancing network security

The US Federal Trade Commission (FTC) has reported that in 2025 more than 2.6 million complaints were filed, of which 61.2% were robocalls (automated prerecorded or AI generated calls) [11]. Robocalls are also the top priority of the US Federal Communications Commission (FCC) [12]. Incidents are candidly not limited to the US, with incidents of robocalls and telephony fraud reported across the entire globe [13].

Most robocalls are mass orchestrated campaigns, with the FTC estimating mass campaigns to contribute to about 64% of all robocalls [14]. Detecting robocalls in real time has been approached in three major paradigms: (1) transcribing the call and using text-based classification [15, 16], (2) classification using audio input or joint audio-text input [17, 18], and (3) deploying virtual AI assistants that intercept and filter calls before forwarding them to the callee [19, 20, 21].

This thesis presents a new approach; exploiting the high reuse rate of mass robocalls and thereby shifting detection from inference-heavy classification to low-cost audio fingerprinting. The proposed framework uses a GNN-encoder [4] to create audio embeddings of robocalls that are robust to noise and other audio degradations, while maintaining high sensitivity. The proposed framework carries several advantages compared to existing approaches; it is multilingual and independent of underlying language, it is light-weight and can perform in real-time without strong hardware demands, it is inherently robust to numerous audio degradations that can occur in practice, and it is non-intrusive i.e. it does not intercept and delay the call, it simply operates in the background. Early findings show an accuracy of 93.9%, which is close to that of existing state-of-the-art solutions [18, 17, 20], while still providing the aforementioned advantages.

2.2. Improving network performance

The International Telecommunication Union Standardization Sector (ITU-T) defined the *answer seizure ratio* (ASR) as the ratio of successfully established telephone calls and one of the key metrics for monitoring network performance within its recommendations E.411 [22] and E.425 [23]. Recent work [24, 25] has shown that this metric can be improved using internal signaling, for example, if a connection was not established it can be retried using another gateway which might improve the ASR. In [24, 25], the authors argued that sometimes telecommunica-

Table 1: *Comparison of audio fingerprinting retrieval rates between music and speech data.*

		Background noise			
		5 dB	10 dB	15 dB	20dB
Speech	audfprint	33.6 ± 5.4	51.7 ± 5.7	64.2 ± 5.5	72.9 ± 5.1
	GraFPrint	59.9 ± 5.6	76.5 ± 4.8	84.4 ± 4.1	88.4 ± 3.8
	PTC	35.9 ± 5.4	50.7 ± 5.7	67.1 ± 5.4	82.9 ± 4.2
Music	audfprint	63.4 ± 5.4	76.2 ± 4.9	85.3 ± 4.3	89.1 ± 3.8
	GraFPrint [4]	98.3	99.3	99.6	99.9
	PTC [6]	93.9	97.1	97.8	/

tion service providers (TSPs) use audio instructions to encode information instead of proper SIP/SS7 signaling codes.

The thesis proposes using audio fingerprinting to quickly identify any received audio instructions instead of signaling codes. The difficulties that arise with this approach is that identification must be performed in real-time and it must be computationally efficient, as TSPs often operate at very large scales. The thesis shows why this approach is more efficient, robust, and scalable compared to transcribing the instructions in order to determine concrete actions.

2.3. Data diversification

A recent study [26] has found that neural codecs exhibit a small-to-medium disparity between the quality of transcoded audio depending on language, while traditional telephony codecs are significantly more biased towards male speech compared to female speech, with the estimated audio quality of transcoded female voices being more than a standard deviation apart from transcoded male voices. Aside from the obvious quality drop that this introduces against female users, it can have severe repercussions for other common telephony applications, as more than 77% of virtual assistants are female-voiced [27].

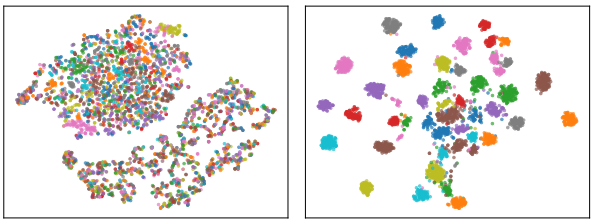
The thesis argues that the speech technology used in the telephone network can (and candidly already does) introduce biases. Since most industry applications are based on labeled/annotated data (more than 95% of industry machine learning is based on supervised data [28]), ensuring that data annotation processes are efficient and encapsulate sufficiently diverse data is of paramount interest to reduce such biases.

In [29], the authors show how audio fingerprinting techniques can help diversify data annotation processes. Due to limited space, this paper only presents a visualization of this result. In Figure 1a, audio embeddings were computed using the popular wav2vec2.0-base model [30] of the VCTK speech corpus [31]. Figure 1b shows embeddings of the same dataset, computed using a neural fingerprinting method [6]. The authors of [29] further show that embeddings of the same speaker tend to form clusters in the latent embedding space when using the neural fingerprinting as the encoder. Using this observation, the paper describes how a technique from image processing – farthest point sampling [32], can be used to diversify the data. This makes the data annotation process more efficient, ensuring that a small number of annotated samples provides a better and more diverse representation of the data.

3. Challenges

The thesis is concerned with applying audio fingerprinting to real industrial problems within the telecommunications domain. Identified challenges and difficulties:

- It is difficult to present results that can be reproduced or verified since data is sometimes subject to GDPR regulations or other legal constraints.
- Unlike automatic speech recognition or speech emotion detection, the problems are not well-known which limits academic interest and the availability of related work and literature.
- This paper presented a few applications of audio fingerprinting with the goal of enhancing telecommunication services. Although the applications have in common the use of audio fingerprinting techniques applied to speech content in a fast, real-time, and scalable way, the applications themselves are not too related. This makes framing the contributions and research questions of the thesis more difficult.
- Audio fingerprinting is not optimized for speech. All novel, state-of-the-art neural fingerprinting models [3, 5, 4, 6] were trained on music data [34]. A recent paper [35] investigates this disparity of the performance of audio fingerprinting methods on speech vs. music, shared in Table 1. The consequences of this disparity is that practical results on speech data are often of subpar quality, requiring additional fine-tuning and adaptation of fingerprinting methods.
- It is difficult to find appropriate publication venues. Related work is often dependent on audio/music processing venues, the work is being applied to telecommunication/cybersecurity related venues, while the data and technical contributions are mostly made in speech related venues.



(a) *wav2vec 2.0-base (95M)* (b) *Conformer small (1.5M)*

Figure 1: *t-SNE [33] visualization of embeddings generated by wav2vec 2.0 [30] and the small pretrained conformer [6].*

4. Summary

This paper presented an overview of an ongoing academic PhD motivated by industry applications of neural fingerprinting for enhancing voice communication services. Future work should address the quality of neural fingerprinting techniques applied to the speech domain and address the challenges presented in this paper, such as adapting the work for academic publishing and ensuring accessible data for easy reproduction of results.

5. References

[1] J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system." in *Proc. of the Int. Society for Music Information Retrieval (ISMIR)*, vol. 3, 2002, pp. 107–115.

[2] A. Wang *et al.*, "An industrial strength audio search algorithm." in *Proc. of the Int. Society for Music Information Retrieval (ISMIR)*, vol. 4, 2003, pp. 7–13.

[3] S. Chang, D. Lee, J. Park, H. Lim, K. Lee, K. Ko, and Y. Han, "Neural audio fingerprint for high-specific audio retrieval based on contrastive learning," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 3025–3029.

[4] A. Bhattacharjee, S. Singh, and E. Benetos, "Grafprint: A gnn-based approach for audio identification," in *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[5] G. Cortès-Sebastià, B. Martin, E. Molina, X. Serra, and R. Hennequin, "Peaknetfp: Peak-based neural audio fingerprinting robust to extreme time stretching," in *Proc. of the Int. Society for Music Information Retrieval (ISMIR)*, vol. 26, 2025, pp. 206–214.

[6] K. Altwlkany, E. Selmanovic, and S. Delalic, "Pretrained conformers for audio fingerprinting and retrieval," *arXiv preprint arXiv:2508.11609*, 2025.

[7] H. Özer, B. Sankur, N. Memon, and E. Anarım, "Perceptual audio hashing functions," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 12, p. 658950, 2005.

[8] D. Renza, J. Vargas, and D. M. Ballesteros, "Robust speech hashing for digital audio forensics," *Applied Sciences*, vol. 10, no. 1, p. 249, 2019.

[9] K. Altwlkany, S. Delalić, A. Alihodžić, E. Selmanović, and D. Hasić, "Application of audio fingerprinting techniques for real-time scalable speech retrieval and speech clusterization," in *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*. IEEE, 2024, pp. 91–96.

[10] International Telecommunication Union, "Measuring digital development: Facts and figures 2025," International Telecommunication Union (ITU), Geneva, Switzerland, Tech. Rep., 2025. [Online]. Available: https://www.itu.int/itu-d/reports/statistics/facts-figures-2025/

[11] US Federal Trade Commission, "National do not call registry," US Federal Trade Commission, Public Tableau, 2025. [Online]. Available: https://public.tableau.com/app/profile/federal.trade.commission/viz/DoNotCall/Infographic

[12] Federal Communications Commission, "Stop unwanted robocalls and texts," https://www.fcc.gov/consumers/guides/stop-unwanted-robocalls-and-texts, 2025, [Online] Accessed: 2026-05-02.

[13] M. Pietri, M. Mamei, and M. Colajanni, "Telecom spam and scams in the 5g and artificial intelligence era: analyzing economic implications, technical challenges and global regulatory efforts," *International Journal of Information Security*, vol. 24, no. 3, p. 139, 2025.

[14] US Federal Trade Commission, "Fraud facts," US Federal Trade Commission, Public Tableau, 2025. [Online]. Available: https://public.tableau.com/app/profile/federal.trade.commission/viz/FraudReports/FraudFacts

[15] A. Natarajan, A. Kannan, V. Belagali, V. N. Pai, R. Shettar, and P. Ghuli, "Spam detection over call transcript using deep learning," in *Proceedings of the Future Technologies Conference*. Springer, 2021, pp. 138–150.

[16] M. Lee and E. Park, "Real-time korean voice phishing detection based on machine learning approaches," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 7, pp. 8173–8184, 2023.

[17] B. Elizalde and D. Emmanouilidou, "Detection of robocall and spam calls using acoustic features of incoming voicemails," in *Proceedings of Meetings on Acoustics*, vol. 45, no. 1. AIP Publishing, 2021.

[18] Z. Yan, R. Tan, Q. Song, and C. X. Lu, "Telesonar: Robocall alarm system by detecting echo channel and breath timing," in *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, 2022, pp. 61–74.

[19] M. Sahin, M. Relieu, and A. Francillon, "Using chatbots against voice spam: Analyzing {Lenny's} effectiveness," in *Thirteenth symposium on usable privacy and security (SOUPS 2017)*, 2017, pp. 319–337.

[20] S. Pandit, J. Liu, R. Perdisci, and M. Ahamad, "Applying deep learning to combat mass robocalls," in *2021 IEEE Security and Privacy Workshops (SPW)*, 2021, pp. 63–70.

[21] S. Pandit, K. Sarker, R. Perdisci, M. Ahamad, and D. Yang, "Combating robocalls with phone virtual assistant mediated interaction," in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023, pp. 463–479.

[22] International Telecommunication Union, "International network management - operational guidance," International Telecommunication Union, Geneva, Switzerland, ITU-T Recommendation E.411, Mar. 1992. [Online]. Available: https://www.itu.int/rec/T-REC-E.411

[23] ——, "Internal automatic observations," International Telecommunication Union, Geneva, Switzerland, ITU-T Recommendation E.425, Mar. 2002. [Online]. Available: https://www.itu.int/rec/T-REC-E.425-200203-I

[24] K. Pribic, S. Delalic, H. Hadzic, K. Altwlkany, A. Kuric, L. Cizmek, and I. Dumlija, "Systems and methods for media analysis for call state detection," May 8 2025, US Patent App. 18/501,134.

[25] K. Altwlkany, H. Hadžić, A. Kurić, and E. Lacic, "Knowledge distillation for real-time classification of early media in voice communications," in *2024 32nd International Conference on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS)*, 2024, pp. 1–4.

[26] K. Altwlkany, A. Kuric, and E. Lacic, "On the Language and Gender Biases in PSTN, VoIP and Neural Audio Codecs," in *Interspeech 2025*, 2025, pp. 1348–1352.

[27] S. Tolmeijer, N. Zierau, A. Janson, J. S. Wahdatehagh, J. M. M. Leimeister, and A. Bernstein, "Female by default?–exploring the effect of voice assistant gender and pitch on trait and trust attribution," in *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, 2021, pp. 1–7.

[28] T. Fredriksson, D. I. Mattos, J. Bosch, and H. H. Olsson, "Data labeling: An empirical investigation into industrial challenges and mitigation strategies," in *International conference on product-focused software process improvement*. Springer, 2020, pp. 202–216.

[29] K. Altwlkany and E. Selmanovic, "Leveraging discriminative capabilities of self-supervised neural audio fingerprinting for efficient speech data annotation," in *Interspeech 2026*, 2026, (To appear).

[30] 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.

[31] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit," 2019, doi: doi.org/10.7488/ds/2645.

[32] T. Schlömer, D. Heck, and O. Deussen, "Farthest-point optimized point sets with maximized minimum distance," in *Proceedings of the ACM SIGGRAPH Symposium on High Performance Graphics*, 2011, pp. 135–142.

[33] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[34] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "Fma: A dataset for music analysis," in *Proc. of the Int. Society for Music Information Retrieval (ISMIR)*, vol. 18, 2017, pp. 316–323.

[35] K. Altwlkany, E. Selmanovic, S. Delalic, and E. Lacic, "Applicability and generalization of audio fingerprinting methods to speech," in *Speech and Computer: 28th International Conference, SPECOM 2026*, ser. Lecture Notes in Computer Science. Ohrid, North Macedonia: Springer, 2026, (To appear).

Poppy Welch

University of Southampton, United Kingdom

Title

Privacy Considerations for Audio-Visual Speech Enhancement

Self-Introduction

Poppy Welch is a 3rd year computer science PhD student at the University of Southampton, UK. Her research develops privacy-preserving methods for audio-visual speech enhancement in hearing aids, and investigates the risks introduced by the use of multimodal data throughout the machine-learning lifecycle. She is additionally interested in socio-technical trust in assistive technology, particularly how privacy protections translate into trust and understanding for users of these devices.

PhD Project Source / Funding

This PhD project is supported by a scholarship from the University of Southampton

Privacy Considerations for Audio-Visual Speech Enhancement

Poppy Welch

Electronics and Computer Science, University of Southampton, United Kingdom

p.welch@soton.ac.uk

Abstract

Recent advancements in speech enhancement leverage both audio and visual information to overcome prior performance limitations in audio-only methods. Multimodal solutions have the potential to transform the way in which the deaf and hard-of-hearing communities use hearing-assistive devices; however the use of the visual modality introduces additional critical privacy concerns that remain largely unaddressed. We address this by investigating a set of privacy-preserving techniques throughout the machine learning lifecycle, at data collection, in the trained model's representations, and at deployment, with an evaluation of the visual privacy each affords. In doing so, we hope to understand how audio-visual hearing aids can be developed safely for their users and society.

Index Terms: speech enhancement, audio-visual, multimodal, privacy, hearing aids

1. Introduction

Over 1.5 billion people across the globe are currently affected by hearing loss to some degree, with this number expected to rise to 2.5 billion by 2050 [1]. Despite this, the uptake of hearing assistive devices is estimated to be at approximately 11% [2], largely due to the poor performance of commercially available hearing aids in noisy environments [3]. Natural hearing perception is inherently multimodal, with listeners making use of congruent visual cues to suppress interfering noise and focus on the target speech in complex sound environments [4]. It is therefore a natural progression to incorporate a similar, cognitively-inspired approach into hearing aids for performance improvements. Audio-visual speech enhancement (AV-SE) models have been shown to outperform traditional audio-only (AO) methods, particularly in challenging sound environments [5, 6, 7]. Before these models can be deployed at scale, however, there are a number of critical privacy concerns introduced by use of the visual stream that remain largely unaddressed in existing research. This thesis investigates the additional visual privacy cost introduced, and strategies to mitigate it across the system.

1.1. Contributions

This thesis aims to investigate the additional visual privacy cost at three points in an AV-SE system, throughout the machine learning lifecycle, at deployment (completed), at data collection (ongoing) and in the model's representations (future work). The overarching research questions are:

RQ1: Can the visual data collected at deployment be reduced while maintaining enhancement performance?

RQ2: Can real visual biometric data in training be reduced or replaced with synthetic data without degrading enhancement performance?

RQ3: What identifiable visual information persists in an AV-SE model's learned representations, and how exposed is it to attack?

1.2. Related Work

In recent years there have been increasing concerns regarding the idea of pervasive surveillance, and attitudes towards wearable smart glasses are largely negative [8, 9], particularly among those who do not use such devices [10] and when the devices are used in social interactions [11]. Wearers of smart glasses have indicated they feel responsible for preserving bystander privacy, and may refrain from wearing the devices when concerned about bystander attitude [12]. Additionally, the European Data Protection Supervisor has recommended that for smart glasses to comply with UK regulations, data minimisation should be applied, and the design must embed the principles of privacy-by-design, which requires privacy to be proactively implemented throughout the processing lifecycle and applied as the default in a user-centric manner transparent to all stakeholders [13, 14]. By implementing privacy-preserving techniques as the default in AV hearing aids, it may be possible to abide by regulation and alleviate wearer concerns, allowing widespread uptake and usage of the devices.

Visual privacy-enhancing techniques from adjacent fields such as biometrics [15], assisted living [16], and wearable devices [17] are largely inadequate for AV-HAs. Obfuscation techniques such as image filtering or face de-identification are often applied post-data collection and may degrade performance [18]. Location-based solutions that prevent data capture in "sensitive" environments [19], may conflict with contexts in which the visual modality is most needed. Intervention methods such as gesture-based solutions [20] or wearable adversarial devices [21] disrupt the initial collection of data but inappropriately place the burden of privacy preservation solely on bystanders who may not even be aware of the presence of a camera. Within AV-SE, the predominant visual privacy-preserving measures focus on protecting speaker identity by modifying data representations. The use of lip-landmark flow features has been proposed due to the inability to reconstruct the face from this input [6], or using compressed visual images, in which speaker identity is largely removed from the visual modality [22]. Existing visual privacy techniques are therefore either applied insufficiently throughout the processing pipeline, degrade the performance gains the visual modality was introduced to provide or shift the burden onto bystanders, leaving the privacy cost of the visual modality in AV-SE largely unaddressed across the system as a whole. We treat visual privacy

as a challenge arising at multiple stages of the system and address each point, at deployment, data collection and the model's learned representations.

1.3. Challenges

There are a number of challenges when considering privacy for audio-visual speech enhancement. Firstly, privacy and performance are often in direct tension with no single correct operating point. The visual information is both the factor that increases performance, and introduces privacy concerns, and considering a balance between the two is use-case dependent. Additionally, privacy perceptions are wide-ranging and often context-dependent [23]. AV-HAs will be deployed across unbounded environments, and bystanders will have their own privacy perceptions formed based on their experiences. Furthermore, these visual privacy gains can only be assessed by testing against specific attacks, which cannot rule out unforeseen or stronger adversaries.

2. Thesis Outline

The work in Section 2.1 has been completed, Section 2.2 describes ongoing work and Section 2.3 outlines research that will be completed in the final year of the degree.

2.1. Modality-Centric Privacy using Back-off

Problem Statement: Privacy-preservation solutions should be easily understandable by users and those that interact with the technology in order to increase trust [24]. In natural human interlocutor interaction, listeners make use of visual cues more heavily in challenging environments of multiple interfering speakers or lower signal-to-noise ratios (SNRs) [25]. Traditional cryptographic data minimisation techniques are challenging for end users due to a lack of real-world analogies that can be used to form correct mental models, and users often do not trust the claim that privacy-enhancing technologies protect their privacy [26]. By replicating interlocutor interaction, a concept for which users have a natural mental model, we may be able to design a trustworthy data minimisation method for AV-SE.

Method: We design a hard-switching model as a feasibility study to investigate whether an AV-SE model can replicate natural interlocutor interaction while increasing visual privacy by reducing the quantity of visual data captured by the sensor. We introduce a back-off strategy for a streaming scenario that controls which input modality is used by the AV-SE framework and when. When the short-time objective intelligibility (STOI) [27] is within 95% of the maximum intelligibility obtained by an AV-model, the algorithm may back off to use the audio-only model, thereby increasing visual privacy through visual data minimisation. We additionally define and quantify visual privacy by considering how frequently the model uses audio-visual input. The back-off model is evaluated in 4 constant SNR scenes, and 3 fluctuating scenes to simulate real-world environments.

Results & Conclusion: In our analysis, we observe that when using a back-off strategy, we can maintain STOI within 95% of the maximum intelligibility with a 58.7% increase in visual privacy (relative to the audio-visual model) when averaged across all SNR conditions. In challenging, lower SNRs the model can rely more heavily on audio-visual input to selectively enhance the target voice, and at higher SNRs, the model can adaptively switch to audio-only input without significantly reducing STOI while increasing visual privacy.

2.2. Synthetic Data Augmentation

Problem statement: Training AV-SE models requires large-scale AV corpora of real speakers, raising privacy concerns regarding the collection and storage of sensitive, biometric information. Augmenting training data with synthetic samples has been shown effective in automatic speech recognition [28] and speaker recognition [29], in addition to audio-only speech enhancement [30]. Using synthetic visual training data could reduce potential privacy risk in data collection and storage, particularly when a large number of samples are required for a single speaker in personalised speech enhancement.

Method: We are using the DeepSpeak v1.1 dataset [31], a corpus with 3 categories of synthetic content: face swap, lip-sync with real audio, and lip-sync with synthetic audio. Using a shared fusion backbone and audio encoder [7], we train 3 models with different visual encoders and inputs: pre-trained AVHuBERT embeddings [32], facial landmarks and raw lip pixel images. Each model is trained across 5 different training splits representing different synthetic/real data ratios (with the total data quantity held constant), to understand whether synthetic data can augment or replace real data. We evaluate only on real audio-visual data. We additionally intend to investigate whether this differs across different visual front-ends and how the real and synthetic visual embeddings differ across encoders, how much of the original identity is retained and the impact of the synthetic generation technique used.

Results: Initial results indicate that synthetic visual training data can augment real training data with minimal performance decrease. At an SNR of -5dB, a model trained on data containing only synthetic visual samples achieves a STOI score of 0.82, and perceptual evaluation of speech quality (PESQ) [33] score of 2.107 vs a STOI score of 0.824 and PESQ score of 2.18 for a model trained on real visual data. The same holds at high SNRs, e.g. at SNR of 10dB, the synthetic data model achieves a STOI=0.944 and PESQ=3.29 vs STOI=0.948 and PESQ=3.353 for the model trained on real data.

2.3. Analysis of Information Leakage

Problem Statement: Audio-visual hearing aids capture a significant amount of rich biometric data from interlocutors and bystanders, some of whom may not be aware of the presence of such a device due to its discreet appearance. Understanding the information that may be obtained by an attacker is vital to understand the potential privacy risk and violations that may occur to these bystanders. Attribute inference attacks consist of attackers attempting to obtain sensitive or undisclosed user attributes from internal model feature embeddings [34]. We intend to investigate what visual identity information is retained in the latent representations of an AV-SE model, and understand how it differs from that in an audio-only SE model.

Proposed Method: We intend to assess information leakage in a speech enhancement model through an attribute inference attack to understand what sensitive information is retained in the latent representations of an audio-only speech enhancement model vs an audio-visual one. In particular, we intend to investigate whether attributes that are not present in the speech enhancement output may be represented in the latent encodings, and how accessible they are. We plan to report leakage at matched performance across audio-only and audio-visual models to reduce confounding factors between model capability and model modalities.

3. References

[1] World Health Organization, "World report on hearing," *World Health Organization*, 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] N. A. Lesica, "Why do hearing aids fail to restore normal auditory perception?" *Trends in neurosciences*, vol. 41, no. 4, pp. 174–185, 2018.

[4] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[5] T. Afouras, J. S. Chung, and A. Zisserman, "My Lips Are Concealed: Audio-Visual Speech Enhancement Through Obstructions," in *Proceedings of Interspeech*, 2019, pp. 4295–4299.

[6] M. Gogate, K. Dashtipour, and A. Hussain, "Towards real-time privacy-preserving audio-visual speech enhancement," in *2nd Symposium on Security and Privacy in Speech Communication*, 2022, pp. 7–10.

[7] T. Ma, S. Yin, L.-C. Yang, and S. Zhang, "Real-Time Audio-Visual Speech Enhancement Using Pre-trained Visual Representations," in *Proceedings of Interspeech*, 2025, pp. 61–65.

[8] S. Zuboff, "The age of surveillance capitalism," in *Social theory re-wired*. Routledge, 2023, pp. 203–213.

[9] M. Koelle, M. Kranz, and A. Möller, "Don't look at me that way! understanding user attitudes towards data glasses usage," in *Proceedings of the 17th international conference on human-computer interaction with mobile devices and services*, 2015, pp. 362–372.

[10] F. Kaviani, B. Lyall, and S. Koppel, "Exploring social perceptions of everyday smartglass use in australia," *PloS one*, vol. 19, no. 11, p. e0313001, 2024.

[11] V. Schwind, J. Reinhardt, R. Rzayev, N. Henze, and K. Wolf, "Virtual reality on the go? A study on social acceptance of VR glasses," in *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*, 2018, p. 111–118.

[12] D. Bhardwaj, A. Ponticello, S. Tomar, A. Dabrowski, and K. Krombholz, "In focus, out of privacy: The wearer's perspective on the privacy dilemma of camera glasses," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024.

[13] European Data Protection Supervisor, "Smart glasses and data protection," European Data Protection Supervisor, Technology Report 1, 2019. [Online]. Available: https://www.edps.europa.eu/data-protection/our-work/publications/reports/smart-glasses-and-data-protection_en

[14] A. Cavoukian *et al.*, "Privacy by design: The 7 foundational principles," *Information and privacy commissioner of Ontario, Canada*, vol. 5, p. 12, 2009.

[15] B. Meden, P. Rot, P. Terhörst, N. Damer, A. Kuijper, W. J. Scheirer, A. Ross, P. Peer, and V. Štruc, "Privacy–enhancing face biometrics: A comprehensive survey," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 4147–4183, 2021.

[16] S. Ravi, P. Climent-Pérez, and F. Florez-Revuelta, "A review on visual privacy preservation techniques for active and assisted living," *Multimedia Tools and Applications*, vol. 83, no. 5, pp. 14 715–14 755, 2024.

[17] R. Alharbi, M. Tolba, L. C. Petito, J. Hester, and N. Alshurafa, "To Mask or Not to Mask? Balancing Privacy with Visual Confirmation Utility in Activity-Oriented Wearable Cameras," in *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, Sep. 2019.

[18] T. Bäckström, S. Ravi, and F. Florez-Revuelta, "Privacy preservation in audio and video," in *Privacy-Aware Monitoring for Assisted Living: Ethical, Legal, and Technological Aspects of Audio- and Video-Based AAL Solutions*, 2025, pp. 77–98.

[19] R. Templeman, M. Korayem, D. J. Crandall, and A. Kapadia, "PlaceAvoider: Steering First-Person Cameras away from Sensitive Spaces," in *Network and Distributed System Security (NDSS) Symposium*, vol. 14, 2014, pp. 23–26.

[20] M. Koelle, S. Ananthanarayan, S. Czupalla, W. Heuten, and S. Boll, "Your smart glasses' camera bothers me! exploring opt-in and opt-out gestures for privacy mediation," in *Proceedings of the 10th Nordic Conference on Human-Computer Interaction*, 2018, pp. 473–481.

[21] S. Komkov and A. Petiushko, "Advhat: Real-world adversarial attack on arcface face id system," in *25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 819–826.

[22] S.-Y. Chuang, Y. Tsao, C.-C. Lo, and H.-M. Wang, "Lite Audio-Visual Speech Enhancement," in *Proceedings of Interspeech*, 2020, pp. 1131–1135.

[23] H. Nissenbaum, "Privacy as contextual integrity," *Washington law review*, vol. 79, no. 1, p. 119, 2004.

[24] D. Harborth, S. Pape, and K. Rannenberg, "Explaining the technology use behavior of privacy-enhancing technologies: The case of tor and jondonym," *Proceedings on Privacy Enhancing Technologies*, 2020.

[25] L. D. Rosenblum, *Primacy of Multimodal Speech Perception*. John Wiley & Sons, Ltd, 2005, ch. 3, pp. 51–78.

[26] S. Pearson, "Privacy, security and trust in cloud computing," in *Privacy and security for cloud computing*. Springer, 2012, pp. 3–42.

[27] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *IEEE ICASSP*, 2010, pp. 4214–4217.

[28] A. Fazel, W. Yang, Y. Liu, R. Barra-Chicote, Y. Meng, R. Maas, and J. Droppo, "SynthASR: Unlocking Synthetic Data for Speech Recognition," in *Interspeech 2021*, 2021, pp. 896–900.

[29] Y. Huang, Y. Chen, J. Pelecanos, and Q. Wang, "Synth2aug: Cross-domain speaker recognition with tts synthesized speech," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 316–322.

[30] L. Zhang, W. Zhang, C. Li, and Y. Qian, "Scale this, not that: Investigating key dataset attributes for efficient speech enhancement scaling," *arXiv preprint arXiv:2412.14890*, 2024.

[31] S. Barrington, M. Bohacek, and H. Farid, "The deepspeak dataset," *arXiv preprint arXiv:2408.05366*, 2024.

[32] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," *arXiv preprint arXiv:2201.02184*, 2022.

[33] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE ICASSP*, vol. 2, 2001, pp. 749–752.

[34] J. Jia and N. Z. Gong, "Attriguard: A practical defense against attribute inference attacks via adversarial machine learning," in *27th USENIX Security Symposium (USENIX Security 18)*, 2018, pp. 513–529.

Baptiste Ramonda

Queensland University of Technology, Australia

Title

Characterization and measurement of listenability for vulnerable individuals

Self-Introduction

I'm a second-year joint PhD candidate co-supervised between the Queensland University of Technology (QUT, Australia) and the University of Toulouse (UT, France). My doctoral research focuses on establishing an interdisciplinary computational framework for speech listenability by bridging natural language processing, acoustic signal processing, and psycholinguistic modeling of listening effort.

PhD Project Source / Funding

This PhD is conducted under an international joint supervision agreement between the Queensland University of Technology (QUT) and the University of Toulouse (UT). The research in France was supported by SAMOVA project residuals, while the onshore research phase at QUT is supported by the QUT Postgraduate Research Award (QUTPRA), a QUT HDR Tuition Fee Sponsorship, and an institutional research funding allowance provided by the QUT Faculty of Science.

Characterization and measurement of listenability for vulnerable individuals

Baptiste Ramonda [1,2]

[1] IRIT, CNRS, UT, Toulouse, France
[2] QUT, Brisbane, Australia
b.ramonda@hdr.qut.edu.au

1. Introduction

The concept of listenability originally emerged in the 1950s within the fields of radio and telecommunications [1, 2, 3], focusing primarily on ensuring that radio speech signals were perceived as clearly as possible by listeners. Today, listenability can be formalized as the global optimization of spoken discourse to achieve a dual objective: maximizing the conveyed information while minimizing the listener's cognitive load. In other words, it represents the capacity to adapt speech to render it both simple and effective. By nature, listenability lies at the intersection of several distinct fields, including natural language processing (NLP), audio signal processing, cognitive science, or human-computer interaction (HCI). The objective of this doctoral research is to establish a unified theoretical model capable of capturing and mapping the interactions across these dimensions. A direct application of this framework is to adapt the generic speech parameters of our future model to the needs of vulnerable audiences, such as individuals with intellectual disabilities or speech impairments. Ultimately, this model aims to foster digital inclusion and enhance their accessibility within our societies. This paper presents the preliminary results obtained during the first year of this PhD project and outlines the theoretical framework and the research roadmap drawn up for the remaining two years.

2. Research Questions

In order to develop a viable theoretical model of listenability, this doctoral research decomposes the problem into three distinct yet interrelated fundamental questions:

- **RQ1 (Definition):** How can the concept of 'listenability' be formally defined in terms of its core objectives and multidisciplinary boundaries?
- **RQ2 (Optimization):** How can we optimize each underlying dimension that influences listenability while accounting for their complex interdisciplinary relationships?
- **RQ3 (Application and Adaptation):** Once a general theoretical model is established, how can its parameters be dynamically adapted to the specific needs of vulnerable audiences to ensure digital accessibility and practical deployment?

3. Progress So Far

The first year of this thesis was devoted to conducting an initial analysis of listenability and related concepts, as well as to studying the interaction between text complexity (readability) and the performance of ASR systems (intelligibility).

3.1. Experiments on Text Readability

In order to properly assess the readability of a text, we first addressed the limitation of traditional readability metrics (e.g., Flesch [4], SMOG [5]), which often exhibit individual methodological biases. To address this issue, we formulated a Global Readability Index (GRI) that standardizes and fuses six classic formulas into a singular metric scored from 0 to 10 (The higher the score, the more complex the text)

$$GRI = \frac{1}{N_i} \sum_{i=1}^{N_i} \frac{I_i - I_{i,\min}}{I_{i,\max} - I_{i,\min}} \times 10 \quad (1)$$

where I_i represents the score of index i, $I_{i,\min}$ and $I_{i,\max}$ define its theoretical boundaries, and N_i is the number of integrated indices.

Using this index and an automated pipeline powered by the Gemini 1.5 Flash API[1], we evaluated text simplification strategies to optimize content readability [6]. We compared a progressive iterative method against a child-directed reformulation model. The latter proved highly efficient, reducing formal text difficulty by over 50% in a single execution (dropping from a complex baseline of $\mu = 9.97$ to $\mu = 4.30$). To verify information fidelity, we then analyzed the semantic proximity between summaries of the original and simplified texts using Sentence-BERT dense embeddings [7]. The valid pairs achieved a high cosine similarity ($\mu = 0.82$, $\sigma = 0.07$) compared to random controls ($\mu = 0.17$, $\sigma = 0.16$), validating our ability to structurally simplify discourse while strictly preserving its core semantic concepts.

3.2. Dissociating Readability from Intelligibility

The second phase of our work challenged the historical paradigm stating that automatic speech recognition (ASR) performance is correlated to the linguistic complexity of the source text. In the context of this study, 'intelligibility' is defined as the objective performance of automated ASR models, utilizing the Word Error Rate (WER) as an established proxy metric for transcription accuracy [8]. To isolate formal text difficulty from human elocution variations, we developed an acoustic control pipeline [9] using the CLEAR corpus and Amazon Polly speech synthesis. To evaluate architectural robustness under realistic environmental constraints, the generated audio files were subjected to acoustic degradations, including additive babble noise (0 dB to 20 dB) and simulated room impulse reverberations.

Our evaluations across multiple ASR architectures revealed that modern end-to-end (E2E) models have reached a state of

[1] https://ai.google.dev/gemini-api/docs/models/gemini

structural decoupling. While older hybrid Kaldi [10] or CTC-based models [11] remain sensitive to text complexity, OpenAI's Whisper (Tiny and Medium) [12] exhibited near-perfect orthogonality with the linguistic structure ($r = 0.03$, $p = 0.059$), a trend validated on natural speech via the LibriSpeech corpus [13] ($r = -0.02$). An assessment of the computational load of the models (predictive token entropy $r = -0.01$ [14] and zlib compression ratios $r = -0.08$) confirmed that there was absolutely no variation in the end-to-end model's transcription performance between very complex texts and very simple ones.

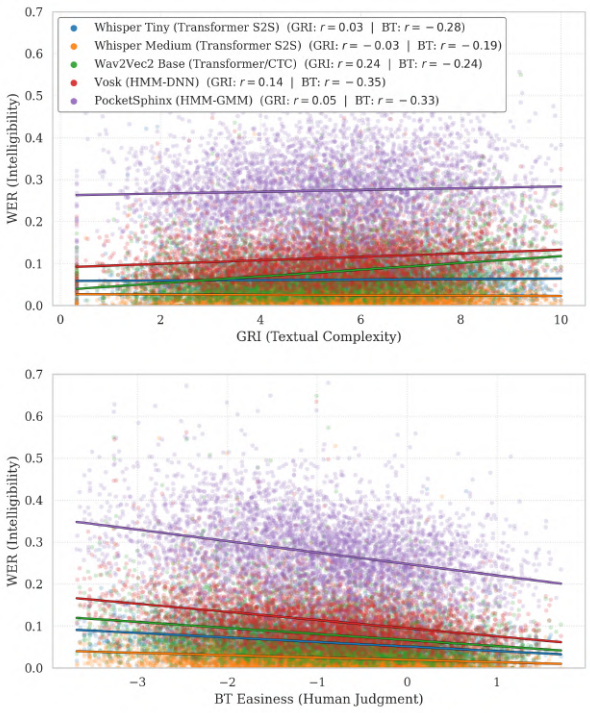


Figure 1: *Impact of Textual Complexity and Human Judgment on ASR Architectures.*

However, a persistent correlation was observed between ASR errors and human-annotated difficulty ($r \approx -0.28$), demonstrating a clear human-machine divergence: contemporary models are immune to formal syntax but remain sensitive to latent perceptual factors that humans find challenging.

4. Future Work & Roadmap

Over the remaining two years of this doctoral project, the focus will shift from isolating independent text-speech variables to modelling the systemic, multi-modal and human-centred dimensions of listenability. The research roadmap is structured around three core areas:

4.1. Systematic Review

Within the speech and computational linguistics communities, there is a need for a robust and unified definition of listenability. While specific dimensions of the concept have been individually addressed since the 1950s [1, 3], it currently suffers from disciplinary fragmentation [15, 16] and lacks a modern, cross-disciplinary formalization. To bridge this gap, we are currently conducting a systematic literature review aimed at synthesizing the state-of-the-art. The objective of this review is to establish a formal, consensus-based definition of listenability. This theoretical groundwork is designed to serve as a standardized baseline for future related research. This study is actively underway, with the future goal of submitting the findings to a peer-reviewed journal in the field.

4.2. GRI Validation

In parallel with this theoretical work, we will conduct a dedicated study to firmly establish the structural foundations of the Global Readability Index (GRI) introduced in previous evaluation frameworks [6, 9]. This axis aims to refine the metric mathematically, notably by optimizing and recalibrating the weight coefficients assigned to each constituent classic formula. The objective is to publish these validation results alongside a public, open-source code repository, providing the speech and language communities with a tool for text readability assessment that minimizes individual index biases. To ensure linguistic and statistical rigour, this study will be pursued in collaboration with specialists in computational text analysis.

4.3. Perceptual Cognitive Dimension

In the longer term, this doctoral research will transition from automated frameworks to ecological, human-centred validation to validate our results in real-world conditions. Incorporating the cognitive dimension represents a critical layer that will be directly superimposed onto our existing findings. To achieve this, we will design behavioural experimental paradigms with human participants to evaluate information retention, attention spans, and objective listening effort in natural settings. This phase will investigate both inter-individual variations — how distinct user profiles, specifically vulnerable audiences with intellectual or speech impairments, modulate discourse processing — and intra-individual variations, tracking dynamic attentional fluctuations and cognitive fatigue over time. This human-centred phase constitutes an important milestone of this doctoral research, which is conducted under an international joint supervision arrangement. This planned research mobility will provide the necessary infrastructure and field-testing opportunities required to translate our theoretical model into actionable digital accessibility solutions for vulnerable populations.

5. Acknowledgements

This doctoral research is conducted within the framework of a joint PhD supervision arrangement between the Université de Toulouse, France, and the Queensland University of Technology (QUT), Brisbane, Australia. The author wishes to express his deepest gratitude to his supervisors, Julien Pinquier, Laurianne Sitbon, and Greig de Zubicaray, for their invaluable guidance, continuous support, and insightful feedback throughout this project. The author also thanks both institutions and their respective research teams for providing the necessary infrastructure and a welcoming collaborative environment.

6. Generative AI Use Disclosure

The authors declare that no generative AI or AI-assisted technologies were used in the writing of this manuscript.

7. References

[1] K. Harwood and F. Cartier, "On definition of listenability," *Southern Journal of Communication*, vol. 18, no. 1, pp. 20–23, 1952.

[2] K. A. Harwood, "I. listenability and readability," *Communications Monographs*, vol. 22, no. 1, pp. 49–53, 1955.

[3] F. A. Cartier, "Ii. listenability and "human interest"," *Communications Monographs*, vol. 22, no. 1, pp. 53–57, 1955.

[4] R. Flesch, "A new readability yardstick." *Journal of applied psychology*, vol. 32, no. 3, p. 221, 1948.

[5] G. H. Mc Laughlin, "Smog grading-a new readability formula," *Journal of reading*, vol. 12, no. 8, pp. 639–646, 1969.

[6] B. Ramonda, I. Ferrané, and J. Pinquier, "Amélioration de la lisibilité de textes via l'utilisation de llm," in *Actes des 18e Rencontres Jeunes Chercheurs en RI (RJCRI) et 27ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL)*, 2025, pp. 1–13.

[7] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[8] M. Karbasi and D. Kolossa, "Asr-based speech intelligibility prediction: A review," *Hearing Research*, vol. 426, p. 108606, 2022.

[9] B. Ramonda, L. Sitbon, and J. Pinquier, "Dissocier la lisibilité et l'intelligibilité: un nouveau paradigme pour l'écoutabilité," in *Proc. JEP 2026*, 2026, pp. 50–59.

[10] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

[11] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

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

[13] 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)*. IEEE, 2015, pp. 5206–5210.

[14] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," *Advances in neural information processing systems*, vol. 31, 2018.

[15] I. E. Fang, "The "easy listening formula"," *Journal of Broadcasting & Electronic Media*, vol. 11, no. 1, pp. 63–68, 1966.

[16] D. L. Rubin, "Listenability as a tool for advancing health literacy," *Journal of Health Communication*, vol. 17, no. sup3, pp. 176–190, 2012.

Himashi Rathnayake

University of Auckland, New Zealand

Title

Community-Oriented Development of Speech Emotion Recognition Framework for te reo Māori

Self-Introduction

I am a final-year PhD student in Computer Systems Engineering at the University of Auckland. My PhD focuses on developing a culturally appropriate speech emotion recognition system for te reo Māori, the Indigenous language of New Zealand. My research interests include natural language processing and speech processing, with a particular focus on developing technologies for low-resource languages.

PhD Project Source / Funding

This PhD project is supported by the Science for Technological Innovation National Science Challenge, funded by the Ministry of Business, Innovation and Employment, New Zealand, in collaboration with Te Hiku Media, and by a University of Auckland Doctoral Scholarship.

Community-Oriented Development of a Speech Emotion Recognition Framework for te reo Māori

Himashi Rathnayake

Department of Electrical, Computer and Software Engineering, University of Auckland, New Zealand

himashi.rathnayake@auckland.ac.nz

Abstract

Speech emotion recognition (SER) has become an increasingly important area in human-computer interaction; however, such technologies remain largely unavailable for low-resource and Indigenous languages such as te reo Māori, the Indigenous language of New Zealand. Existing approaches often rely on Western-derived emotion categories and methodologies, with limited community involvement. This paper presents an overview of a doctoral thesis that addresses these gaps by developing a community-oriented SER system for te reo Māori. The research involves identifying culturally grounded emotion categories through community engagement, developing the first Māori emotional speech corpus, and conducting acoustic feature analysis to investigate patterns of emotional expression. This process will inform the development of the first Māori SER system. This work contributes a culturally grounded framework for emotion technology development that can be extended to other Indigenous and low-resource languages.

Index Terms: speech emotion recognition, community-oriented, Māori, Indigenous language

1. Introduction and Motivation

Identifying emotions from speech, known as speech emotion recognition (SER), is widely used in human-computer interaction, particularly due to the growing demand for contactless interaction. However, the development of such speech technologies has been largely limited to a small subset of languages, accounting for only about 100 of the more than 7,000 languages spoken worldwide [1]. Most SER research has focused on high-resource languages, such as English, German, Mandarin and French [2, 3, 4], while low-resource languages remain significantly underexplored. Notably, there is currently no SER system available for te reo Māori, the Indigenous language of New Zealand (Māori referring to the people and te reo Māori to the language). This study, therefore, aims to address this gap by developing a SER system for te reo Māori.

A review conducted in the initial stages of this thesis on SER for low-resource and Indigenous languages found that existing systems often classify speech into predefined emotion categories derived from English or their direct translations, assuming emotions are universal [5]. There is an ongoing debate about how emotions are categorised differently across languages due to cultural variations. For instance, researchers argue that experiments designed to verify the universality of emotions have flaws due to the lack of native speaker involvement and the failure to consider Indigenous knowledge [6]. There is also discussion and evidence that categorisation and perception of basic emotions can vary across cultures [7, 8, 9, 10]. Specifically, for Māori, studies have found that Māori and New Zealand Europeans exhibit different patterns of emotional reactions to various social situations [11]. A Māori psychiatrist, Elder [12], suggests that the nuances of Māori emotions are lost when they are referenced in English. Further, Ritchie et al. [13] noted that Māori contains a wide vocabulary of emotions that lack direct English translations. Hence, using emotions defined in English or their closest translations might be inadequate for te reo Māori, and there is a need to experimentally verify how emotions are categorised in te reo Māori.

Another limitation of current SER systems is that they often do not employ community-oriented approaches or incorporate cultural knowledge when developing technologies for Indigenous languages [5]. Further, many existing SER systems rely on end-to-end deep learning approaches, with limited integration of established speech processing knowledge.

These gaps lead to the research question: ***How can a culturally grounded speech emotion recognition system for te reo Māori be co-developed through community engagement?***. To address this question, this thesis adopts a community-driven, multi-stage approach. It begins with the identification of Māori emotion categories through community engagement, followed by the development of the first annotated emotional speech corpus for te reo Māori, and an acoustic analysis of how these emotions are expressed in speech. These insights then inform the development of a SER system for Māori.

2. Methodology and Preliminary Results

The proposed community-oriented methodology is illustrated in Figure 1 and outlined below.

1. Community media collection: Since there was no annotated emotional speech corpus for te reo Māori, emotional speech data were collected from community media sources, including Te Hiku Media radio recordings [14] and Māori television programmes. A preliminary annotation was carried out

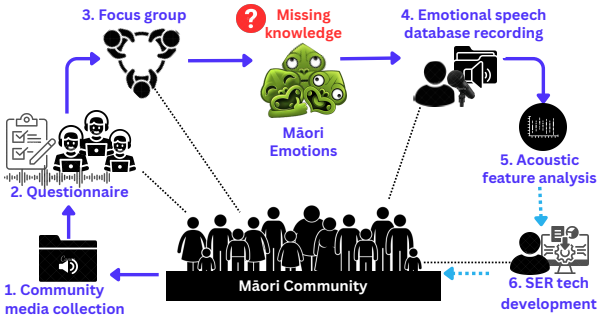


Figure 1: *Community-oriented framework for Māori SER.*

to identify audio clips containing emotional content, with non-emotional segments excluded from further analysis.

2. Online questionnaire: A questionnaire was designed to obtain Māori community feedback on emotions conveyed in the audio clips identified in the previous stage [15]. Participants were asked to listen to selected clips and provide free-form descriptions of the emotions they perceived. The questionnaire was designed to be bilingual, as all Māori speakers are bilingual in English and te reo Māori, and to reflect the impact of te reo Māori as a revitalised language. In this context, speakers may use English to express emotions while still conveying an emotional framework that is culturally Māori. From these responses, over 200 emotion terms were collected, with several identified as synonymous.

3. Focus group: The terms were further analysed and categorised through focus group discussions with Māori participants, ensuring that cultural nuances in the emotions were preserved during the grouping process. This process yielded 16 emotion categories, which are essential for meaningful emotion analysis and the development of emotion technology. These categories include: ngenge (tired), haumaru (safe), aroha (love), huakore (hopeless), pōuri (sad), harikoa (happy), manawanui (serious), pai (good, well), kaikā (impatient), whai whakaaro (curious), whakarihariha (disgusted), whakahīhī (proud), hōhā (annoyed), hīkaka (excited), ohorere (surprised), āwangawanga (anxious). While the English translations provided here are the close equivalents, they do not fully capture the cultural nuances of the original Māori terms. Some of these emotions correspond broadly to presumptive universal emotions (e.g., harikoa and pōuri). Others reflect culturally specific concepts, such as *pai*, for which available English equivalents do not fully capture its cultural meaning, and *aroha*, a culturally rich concept that is not typically represented in mainstream emotion research. These distinctions highlight the limitations of universal emotion models and the importance of culturally grounded approaches.

4. Emotional speech corpus development: Pā-Kakare, the first Māori emotional speech corpus, was developed as the next step, facilitating further understanding of Māori emotion expression [16]. Unlike existing emotional speech corpora that rely on presumed universal emotions, this corpus is grounded in the community-derived 16 emotion categories, ensuring cultural relevance. The corpus was recorded with four professional Māori actors, with 15 sentences per emotion, each sentence recorded four times over two days, resulting in 3,840 high-quality recordings (16 emotions $\times$ 15 sentences $\times$ 4 actors $\times$ 4 takes). Database validation was subsequently conducted with 28 participants, who evaluated a randomly selected 15% of the recordings for perceived emotion. To uphold principles of data sovereignty and kaitiakitanga (guardianship), the corpus is treated as a community resource rather than as researcher-owned data. Its use is governed by a Kaitiakitanga license [17], ensuring culturally appropriate access, use, and protection.

5. Acoustic feature analysis: As the next step, acoustic feature analysis was conducted using prosodic, voice-quality, and spectral features to investigate how these features differentiate Māori emotions. In existing literature, such differences are commonly interpreted in terms of arousal (the level of activation) and valence (positive or negative affect) [18, 19]. For example, fundamental frequency (f0) and intensity tend to be higher for high-arousal emotions and lower for low-arousal emotions. However, these interpretations are largely based on Western-derived emotion frameworks. Therefore, this study examines whether emotions expressed in Māori speech exhibit similar or differing patterns. Preliminary statistical analysis revealed significant differences across several prosodic and spectral features and identified emotion pairs with overlapping acoustic characteristics. For example, kaikā (impatient) and hōhā (annoyed) exhibit similar acoustic profiles in Māori, despite being associated with different arousal levels in English-based emotion models. These findings highlight both cross-cultural similarities and the culturally specific nature of emotional expression, reinforcing the importance of developing culturally grounded approaches to SER.

The next stage of this work, which focuses on SER technology development, is described in the following section.

3. Future Work

The final objective of this research is to develop a SER system for te reo Māori. This involves investigating which approaches are most effective for modelling culturally grounded emotions. Using the developed Pā-Kakare corpus, along with the extracted acoustic features, a range of machine-learning and deep-learning approaches will be evaluated. In particular, this study evaluates the effectiveness of acoustic features using traditional machine learning approaches and of frame-level representations using deep learning models as baselines. Recent self-supervised models such as Emotion2Vec will also be evaluated [20]. These models are trained on large-scale datasets based on Western-derived emotion categories. Their ability to represent culturally grounded emotion distinctions in te reo Māori will therefore be evaluated rather than assumed. Model performance will be assessed using metrics such as accuracy, precision, recall, and F1-score under both speaker-dependent and speaker-independent evaluation settings.

4. Challenges

The development of the first Māori SER system presents several challenges. First, as a revitalised Indigenous language, there is a significant lack of available resources, including annotated data and linguistic tools. In addition, identifying culturally grounded emotions was time-consuming, requiring extensive community engagement and iterative analysis. Furthermore, the current state of the art in emotion recognition for low-resource languages is largely based on pre-trained models using self-supervised learning. However, these models may not transfer well to Māori, as they are trained on Western-derived emotion categories and lack exposure to Māori data, limiting their ability to model the 16 culturally grounded emotion categories identified in this study. Conducting research with Indigenous communities requires careful ethical consideration, particularly regarding data governance, representation, and cultural sensitivity. These were addressed through close collaboration with a transdisciplinary team, including Māori researchers, ensuring culturally appropriate and community-oriented research.

5. Contribution

The main contributions of this study are as follows: (1) The identification of culturally grounded common social emotion categories in te reo Māori speech, (2) The development of Pā-Kakare, the first emotional speech corpus for te reo Māori, (3) An acoustic characterisation of emotional expression in Māori speech, (4) The planned development of the first SER system for te reo Māori, and (5) A community-oriented framework for developing speech emotion technologies that can be adapted for other Indigenous and low-resource languages.

6. Acknowledgment

This research has been a collaborative effort, where technology developers, language experts, and the Māori community have worked together to co-develop emotion technology for Māori. Research decisions were discussed within the team, drawing on the diverse expertise and perspectives of its members. I sincerely thank my supervisors, Dr Jesin James, Dr Ake Nicholas, Dr Gianna Leoni, Prof Catherine Watson, and A/Prof Peter Keegan, for their guidance and support throughout this research. I also sincerely thank all the Māori participants who contributed their time, knowledge, and perspectives at different stages of this thesis. I gratefully acknowledge the financial support provided by the New Zealand Science for Technological Innovation National Science Challenge, Te Hiku Media, and the University of Auckland.

7. References

[1] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi *et al.*, "Scaling speech technology to 1,000+ languages," *Journal of Machine Learning Research*, vol. 25, no. 97, pp. 1–52, 2024.

[2] S. Taj, G. Mujtaba, S. M. Daudpota, and M. H. Mughal, "Urdu speech emotion recognition: A systematic literature review," *ACM Transactions on Asian and Low-Resource Language Information Processing*, 2023.

[3] S. Chopra, P. Mathur, R. Sawhney, and R. R. Shah, "Meta-learning for low-resource speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Toronto, ON, Canada: IEEE, 2021, pp. 6259–6263.

[4] Y. Nam and C. Lee, "Chung-ang auditory database of korean emotional speech: A validated set of vocal expressions with different intensities," *IEEE Access*, vol. 10, pp. 122 745–122 761, 2022.

[5] H. Rathnayake, J. James, G. Leoni, A. Nicholas, C. Watson, and P. Keegan, "A review on speech emotion recognition for low-resource and indigenous languages," *Speech Communication*, p. 103342, 2025.

[6] K. Crawford, "Artificial intelligence is misreading human emotion," *The Atlantic*, vol. 27, 2021.

[7] R. E. Jack, O. G. Garrod, H. Yu, R. Caldara, and P. G. Schyns, "Facial expressions of emotion are not culturally universal," *Proceedings of the National Academy of Sciences*, vol. 109, no. 19, pp. 7241–7244, 2012.

[8] M. Gendron, D. Roberson, J. M. van der Vyver, and L. F. Barrett, "Perceptions of emotion from facial expressions are not culturally universal: evidence from a remote culture." *Emotion*, vol. 14, no. 2, p. 251, 2014.

[9] R. A. Shweder, *Thinking through cultures: Expeditions in cultural psychology*. Harvard University Press, 1991.

[10] J. A. Russell, "Culture and the categorization of emotions." *Psychological bulletin*, vol. 110, no. 3, pp. 426–450, 1991.

[11] C. Banks, "A comparison study of maori and pakeha emotional reactions to social situations that involve whakamaa," Ph.D. dissertation, Massey University, 1996.

[12] H. Elder, *Aroha: Maori wisdom for a contented life lived in harmony with our planet*. Random House, 2020.

[13] J. Ritchie, "Staying with the Troubles of Colonised Emotional Well-Being of Young Children in Aotearoa (New Zealand)," 2021.

[14] Te Hiku Media, "Te hiku media," https://tehiku.nz/, 2026.

[15] H. Rathnayake, J. James, A. Nicholas, G. Leoni, C. I. Watson, and P. Keegan, "Towards identifying the common social emotions in spoken te reo maori: A community-oriented approach," in *Proceedings of the Nineteenth Australasian International Conference on Speech Science and Technology*, 2024.

[16] H. Rathnayake, J. James, A. Nicholas, G. Leoni, C. I. Watson, and P. J. Keegan, "Pā-kakare: The first emotional speech database for te reo māori," in *Proceedings of Interspeech*, 2026.

[17] Te Hiku Media. (2026) Kaitiakitanga license. whare kōrero kaitiakitanga license. https://xn--wharekrero-v3b.nz/kaitiakitanga.

[18] M. D. Pell, S. Paulmann, C. Dara, A. Alasseri, and S. A. Kotz, "Factors in the recognition of vocally expressed emotions: A comparison of four languages," *Journal of Phonetics*, vol. 37, no. 4, pp. 417–435, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0095447009000448

[19] J. James, "Modeling prosodic features for empathetic speech of a healthcare robot," Ph.D. dissertation, ResearchSpace@ Auckland, 2020.

[20] Z. Ma, Z. Zheng, J. Ye, J. Li, Z. Gao, S. Zhang, and X. Chen, "emotion2vec: Self-supervised pre-training for speech emotion representation," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, pp. 15 747–15 760.

Dushanthi Madhushika Manamalage

University of Auckland, New Zealand

Title

Towards Trustworthy Speech-Based Multimodal Depression Detection

Self-Introduction

I am a third-year PhD candidate at the University of Auckland, New Zealand. My research focuses on developing trustworthy, privacy-preserving AI for speech-based depression assessment, with particular interests in multimodal learning, federated learning, and interpretable AI. Through my PhD, I aim to address challenges related to limited and sensitive mental health data while improving the reliability and practical applicability of AI-based assessment systems.

PhD Project Source / Funding

This PhD project is not supported by external project funding. I am supported by a full doctoral scholarship from the University of Auckland, New Zealand.

Towards Trustworthy Speech-Based Multimodal Depression Detection

Dushanthi Madhushika Manamalage [1,**]

[1] DeepNet Discovery Network, The University of Auckland, New Zealand

dmad651@aucklanduni.ac.nz

Abstract

Speech-based multimodal artificial intelligence has demonstrated considerable potential for automatic depression detection, offering scalable and non-invasive approaches for mental health screening. However, real-world deployment remains challenging due to limited clinical data, privacy concerns surrounding sensitive speech information, and the limited interpretability of deep learning models. This study investigates how multimodal speech-based depression detection systems can be made trustworthy by jointly addressing predictive performance, privacy preservation, and interpretability. The research progresses through three interconnected stages: developing effective multimodal depression detection under data-scarce conditions, enabling privacy-preserving learning through federated learning, and investigating representation-level interpretability in privacy-aware multimodal systems. Together, these stages establish a unified pathway toward trustworthy multimodal depression detection for clinically relevant mental health assessment.

Index Terms: Speech-based Depression Detection, Federated Learning, Privacy-Performance, Interpretability

1. Motivation

Depression is a common mental health disorder characterized by persistent low mood, loss of interest, and reduced daily functioning. It affects approximately 280 million people worldwide and remains one of the leading causes of disability [1]. Despite the availability of effective treatments, many individuals do not receive timely support due to limited access to mental health services, shortages of trained professionals, and increasing demands on healthcare systems [2, 3]. These challenges highlight the need for scalable, accessible, and non-invasive approaches to early depression screening. Speech, which can be conveniently collected through smartphones and telehealth platforms, offers a promising opportunity to address this need.

Speech has emerged as a promising digital biomarker for depression because depressive symptoms influence both *how* people speak (acoustic) and *what* they say (linguistic/context) [4]. Combined with recent advances in artificial intelligence (AI), multimodal speech-based depression detection systems that jointly model acoustic and linguistic information have demonstrated considerable potential for accessible and scalable mental health assessment.

However, several challenges continue to limit their real-world adoption. First, depression datasets are often small, imbalanced, and noisy, making effective multimodal learning difficult [5, 6, 7]. Second, speech recordings contain highly sensitive personal information, creating significant privacy concerns for centralized learning approaches [8]. Third, although privacy-preserving learning can mitigate these risks, it often degrades predictive performance. Moreover, existing approaches provide limited insight into how clinically relevant representations are learned and preserved, while lacking a unified framework for evaluating trustworthiness in clinical deployment [9]. These challenges motivate three connected research stages: improving multimodal learning under limited, imbalance and noisy data, enabling privacy-preserving decentralized learning, and developing interpretable privacy-aware models with systematic trustworthiness evaluation.

2. Research Questions

These challenges motivate the overarching research question of this doctoral research: **How can multimodal speech-based depression detection systems be made trustworthy by jointly addressing performance, privacy preservation, and interpretability?**

To address this question, the dissertation is organized around three interconnected research questions:

- **RQ1:** How can complementary speech (acoustic and linguistic) information be effectively integrated to improve depression detection under limited, imbalanced, and noisy data conditions?
- **RQ2:** How can privacy-preserving techniques preserve predictive utility and speaker privacy when jointly learning acoustic and linguistic depression representations across decentralized data?
- **RQ3:** How can representation-level insights be leveraged to develop interpretable privacy-aware multimodal depression detection systems with systematic trustworthiness evaluation?

Together, these research questions establish a progression from effective multimodal learning (RQ1) to privacy-preserving deployment (RQ2), and finally to trustworthy and interpretable federated depression detection systems (RQ3).

2.1. Improving Performance in Data-Scarce Environments

The first stage of this research investigated effective multimodal depression detection under limited, imbalanced, and noisy data conditions. Existing depression datasets often contain transcription errors, recording variability, speaker heterogeneity, and class imbalance, all of which negatively affect model learning [5, 6, 7].

To address these challenges, I proposed ALiDeR [10], a multimodal depression detection framework that jointly integrates transformer-based acoustic (Whisper) [11] and linguis-

** indicates the corresponding author.

tic (DepRoBERTa) [12] representations with a quality-aware data processing pipeline, including transcript refinement, participant speech extraction, and the proposed Signal-to-Noise Ratio-based Adaptive Data Selection (SNR-ADS) method [13].

Experimental evaluation on the E-DAIC (English) [5, 6] and MODMA (Chinese) [7] datasets demonstrated that ALiDeR consistently outperformed unimodal and existing multimodal baselines. The proposed framework achieved accuracies of 0.92 and 0.89, and F1-scores of 0.91 and 0.88 on the E-DAIC and MODMA datasets, respectively. Comprehensive ablation studies further confirmed that effective multimodal fusion, quality-aware data selection, and transformer-based representation learning contributed to the observed improvements.

These findings validate the proposed multimodal learning strategy under limited, imbalanced, and noisy data conditions and establish a strong centralized baseline for the privacy-preserving investigations in RQ2.

2.2. Privacy-Preserving Multimodal Depression Detection

While centralized multimodal learning achieved strong predictive performance, sharing sensitive speech and mental health data introduced substantial privacy risks. The second stage therefore investigated privacy-preserving learning through federated learning (FL).

To address this challenge, I developed FedMPA, a novel federated multimodal learning framework that combines decentralized optimization, prototype-guided knowledge sharing, and privacy-aware prototype aggregation mechanisms [14, 15]. In addition, I proposed a multi-channel privacy evaluation pipeline together with novel Privacy–Performance Score (PPS), providing a unified quantitative framework for evaluating predictive utility and privacy preservation. Experimental evaluation on the E-DAIC [5, 6] and MODMA [7] datasets demonstrated that FedMPA consistently preserved predictive performance while improving privacy protection compared with existing federated baselines, achieving accuracies of 0.88 and 0.83, F1-scores of 0.89 and 0.84, and PPS values of 0.92 and 0.87, respectively.

These findings demonstrate that privacy-preserving multimodal learning can maintain competitive predictive performance while enhancing privacy protection, validating the proposed FedMPA framework for decentralized clinical AI. More importantly, they motivate the next stage of this research, which investigates how clinically relevant representations are learned and preserved under privacy-aware learning and how these insights can support interpretable and trustworthy clinical AI.

2.3. Interpretability in Privacy-Aware Learning

The third stage of this research investigates how representation-level insights can be leveraged to develop interpretable privacy-aware multimodal depression detection systems and systematically evaluate their trustworthiness. While existing explainability methods primarily rely on post-hoc feature attribution or attention visualization, they provide limited insight into how clinically relevant depression representations are learned and preserved under privacy-aware learning [9]. Moreover, current FL studies predominantly evaluate predictive performance and privacy protection, with little understanding of the representational effects of decentralized optimization [16, 17, 18].

To address this gap, this research investigates the layer-wise organization of depression-related acoustic and linguistic representations and exploits these insights to guide representation-aware multimodal learning in federated settings. The proposed framework is being evaluated through three complementary criteria: (i) identification of clinically relevant layer-wise representations, (ii) improved predictive performance through representation-guided learning under privacy-aware FL, and (iii) a unified trustworthiness evaluation incorporating predictive performance, privacy preservation, and multi-level interpretability.

Preliminary analyses have identified modality-specific features and task representations across transformer layers, providing a foundation for developing representation-guided multimodal learning and systematic trustworthiness evaluation. Ongoing work builds on these findings to establish a unified framework for evaluating predictive performance, privacy preservation, and interpretability.

3. Current and Future Research Directions

This research has progressively evolved from improving multimodal depression detection under limited, imbalance and noisy data conditions (RQ1) to enabling privacy-preserving learning through FL (RQ2), and currently focuses on advancing representation-level interpretability to achieve trustworthy privacy-aware multimodal depression detection (RQ3). Together, these stages establish a pathway towards trustworthy multimodal depression detection by jointly addressing predictive performance, privacy preservation, and interpretability.

Future work will investigate how clinically relevant depression-related information emerges across transformer layers, how acoustic and linguistic representations align during multimodal learning, and how privacy-preserving mechanisms influence these representations in FL. Representation-level analyses are being developed to characterize layer-wise information emergence and multimodal alignment in federated settings. Additional studies will include cross-domain validation and the development of quantitative metrics for evaluating representation-level interpretability and overall trustworthiness. Ultimately, this research aims to establish a unified approach for trustworthy multimodal depression detection that jointly optimizes predictive performance, privacy preservation, and interpretability.

4. Conclusion

This doctoral research investigates how multimodal speech-based depression detection systems can be made trustworthy by jointly addressing predictive performance, privacy preservation, and interpretability. To this end, the research has progressively advanced from developing ALiDeR, a multimodal framework for effective depression detection under data-scarce conditions [13], to proposing FedMPA, together with a multi-channel privacy evaluation pipeline and the Privacy–Performance Score (PPS), for privacy-aware federated multimodal learning [14, 15]. Ongoing research extends these contributions by investigating representation-level interpretability to understand how clinically relevant depression-related information is learned and preserved under privacy-aware learning. Collectively, these contributions establish a unified research pathway toward trustworthy multimodal depression detection by jointly advancing predictive performance, privacy preservation, and interpretability.

5. Generative AI Use Disclosure

During the preparation of this work, the author used GPT-5.2 to review grammar and improve readability. After using

this tool/service, the author reviewed and edited the content as needed and took full responsibility for the content of the published article.

6. References

[1] W. H. Organization, "Depression," https://www.who.int/health-topics/depression#tab=tab_1, 2026, accessed: 2026-02-12.

[2] ——, "Depressive disorder (depression)," https://www.who.int/news-room/fact-sheets/detail/depression, 2026, accessed: 2026-02-12.

[3] K. Bryant, N. Greer-Williams, N. Willis, and M. Hartwig, "Barriers to diagnosis and treatment of depression: Voices from a rural african-american faith community," *Journal of National Black Nurses' Association : JNBNA*, vol. 24, no. 1, p. 31, 2013. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC3983966/pdf/nihms565271.pdf

[4] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639315000369

[5] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, S. Rizzo, and L.-P. Morency, "The distress analysis interview corpus of human and computer interviews," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 3123–3128. [Online]. Available: https://aclanthology.org/L14-1421/

[6] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, G. M. Lucas, S. Marsella, F. Morbini, A. Nazarian, S. Scherer, G. Stratou, A. Suri, D. R. Traum, R. Wood, Y. Xu, A. A. Rizzo, and L. philippe Morency, "Simsensei kiosk: a virtual human interviewer for healthcare decision support," in *Adaptive Agents and Multi-Agent Systems*, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:11087454

[7] H. Cai, Z. Yuan, Y. Gao, S. Sun, N. Li, F. Tian, H. Xiao, J. Li, Z. Yang, X. Li, Q. Zhao, Z. Liu, Z. Yao, M. Yang, H. Peng, J. Zhu, X. Zhang, G. Gao, F. Zheng, R. Li, Z. Guo, R. Ma, J. Yang, L. Zhang, X. Hu, Y. Li, and B. Hu, "A multi-modal open dataset for mental-disorder analysis," *Scientific Data*, vol. 9, no. 1, Apr. 2022. [Online]. Available: http://dx.doi.org/10.1038/s41597-022-01211-x

[8] S. G. Brederoo, F. G. Nadema, F. G. Goedhart, A. E. Voppel, J. N. D. Boer, J. Wouts, S. Koops, and I. E. Sommer, "Implementation of automatic speech analysis for early detection of psychiatric symptoms: What do patients want?" *Journal of Psychiatric Research*, vol. 142, pp. 299–301, 10 2021. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/34416548/

[9] K. Sokol and P. Flach, "Interpretable representations in explainable ai: from theory to practice," *Data Mining and Knowledge Discovery*, vol. 38, no. 5, p. 3102–3140, Apr. 2024. [Online]. Available: http://dx.doi.org/10.1007/s10618-024-01010-5

[10] D. M. Manamalage, R. N. Rao, M. Bilal, G. Cheung, H. Thabrew, P. S. Roop, S. R. Shahamiri, and F. Sundram, "With data in mind: Ai applications in depression, autism and dementia care," *New Zealand Medical Student Journal*, vol. 0, no. 39, Sep. 2025. [Online]. Available: http://dx.doi.org/10.57129/001c.144915

[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. ICML'23. JMLR.org, 2023.

[12] R. Poświata and M. Perełkiewicz, "OPI@LT-EDI-ACL2022: Detecting signs of depression from social media text using RoBERTa pre-trained language models," in *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, B. R. Chakravarthi, B. Bharathi, J. P. McCrae, M. Zarrouk, K. Bali, and P. Buitelaar, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 276–282. [Online]. Available: https://aclanthology.org/2022.ltedi-1.40/

[13] D. M. Manamalage, S. R. Shahamiri, P. S. Roop, and F. Sundram, "Adaptive signal-to-noise ratio data selection for robust multimodal depression assessment using transformer representations," 2026, manuscript under review at Cognitive Computation.

[14] ——, "Fedmpa: A novel privacy-performance optimization approach for multimodal speech-based depression detection," in *Proceedings of Interspeech 2026*, 2026, accepted for publication.

[15] ——, "Privacy-enhanced prototype-based federated learning with multi-channel privacy evaluation for speech-based depression detection," 2026, manuscript under review at ACM Transactions on Computing for Healthcare.

[16] P. Kairouz and H. B. McMahan, "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1-2, p. 1–210, Jun. 2021. [Online]. Available: http://dx.doi.org/10.1561/2200000083

[17] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[18] A. Pasad, B. Shi, and K. Livescu, "Comparative layer-wise analysis of self-supervised speech models," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

Rong Li

University of Twente, The Netherlands

Title

Detection and Perception of Smiled Speech

Self-Introduction

I am Rong Li, a PhD candidate in Human–Media Interaction at the University of Twente in the Netherlands, and I have just started my second year. My research focuses on smiled speech, particularly how smiling is expressed in the voice, how it can be automatically detected, and how it is perceived by listeners. I am interested in speech perception, paralinguistic communication, and affective computing.

PhD Project Source / Funding

This PhD project is supported by the EU Horizon Europe Marie Skłodowska-Curie Actions (MSCA) Doctoral Network Voice Communication Sciences (VoCS) (Grant Agreement No. 101168998).

Detection and Perception of Smiled Speech

Rong Li

Human Media Interaction, University of Twente, The Netherlands

rong.li@utwente.nl

Abstract

This PhD investigates smiled speech as an audible social signal. While smiles are usually studied visually, smile-related cues also shape the speech signal and can be perceived by listeners. The project examines how these cues are expressed, perceived, and modeled across smile types and communicative contexts. Understanding smiled speech is relevant for everyday communication, human–machine interaction, and speech technologies in noisy or audio-only settings. This paper summarizes the completed first study, planned human–machine comparison and database work, expected contributions, and key methodological challenges.

Index Terms: smiled speech, vocal smile, smile perception, paralinguistic speech processing, affective computing

1. Introduction and Research Questions

Smiles are socially rich signals. They can signal amusement, happiness, embarrassment, discomfort, or even sadness [1, 2], and they support social functions ranging from reward and affiliation to hierarchy negotiation [3]. Although smiles are usually treated as facial expressions, they also leave traces in the speech signal. Prior work has shown that smiling can shift formant structure, alter intensity and voice quality, and create an audible impression of "smiling" even without visual access [4, 5, 6]. Listeners can also distinguish different kinds of vocal smile-related speech, including mechanically spread-lip speech and affective amused speech [7, 8, 9], and auditory smiles can trigger facial imitation in listeners [10].

Despite these findings, smiled speech remains a fragmented research area compared with facial smiling, laughter, or broad speech emotion recognition (SER). Existing resources are often audio only with limited speakers, focused on acted or scripted material, and collected under clean laboratory conditions. As a result, it remains unclear how smile-related cues are produced, perceived, and modeled in the kinds of settings where vocal smile cues may matter most, such as telephone and contact-center speech and human–machine conversations where visual access may be limited or the signal is noisy.

Taken together, these gaps motivate the central question of this PhD project: how are smile-related cues expressed, perceived, and modeled in speech across different smile types and communicative contexts? This question is addressed through four linked research questions:

1. **Smiled speech production:** Which acoustic and visual cues distinguish smiled speech from neutral speech across smile types and contexts?
2. **Human perception:** How do listeners perceive smiled speech, and different smile types, under realistic listening conditions?
3. **Automatic modeling:** How can smiled speech be automatically detected and modeled using acoustic representations?
4. **Communicative interpretation:** How do smile-related cues shape social impressions, such as perceived warmth, trustworthiness, or friendliness?

2. Work Completed So Far

The first completed study examined whether listeners can still hear smiles in multi-talker babble noise [11]. The experiment used the AMuS corpus [9], which distinguishes neutral, spread-lip, and amused speech. Seventy-five listeners completed two binary categorization tasks: smile-like detection, contrasting neutral with smile-like speech, and smile-type categorization, contrasting amused with spread-lip speech. The stimuli were presented in quiet, -3 dB SNR, and -6 dB SNR. Trial-level generalized linear mixed-effects models (GLMMs) were combined with signal detection theory (SDT) analyses to separate classification accuracy from response bias.

As shown in Figure 1, noise did not simply make smile perception less accurate. In both tasks, increasing noise reduced sensitivity, indicating that listeners had less access to smile-related auditory cues. At the same time, the decision criterion shifted upward, showing that listeners became less likely to choose the smile-like or more affective response under uncertainty. In smile-like detection, sensitivity decreased from $d' = 1.87$ in quiet to $d' = 0.58$ at -6 dB SNR, while the criterion shifted from $c = 0.09$ to $c = 0.69$. For smile-type categorization, an initial bias toward "amused" responses in quiet reversed under strong noise ($c = -0.51$ to $c = 0.64$).

The main contribution of this study is therefore to show that auditory smile perception in noise is not only a problem of degraded sensory evidence. Rather, background noise also changes how listeners use the remaining evidence: as the signal becomes less reliable, they become more conservative in interpreting speech as smile-like or affectively amused. The study also exposed limitations that shape the next stages of the PhD. The AMuS corpus is valuable because it separates spread-lip and amused speech, but it only contains a small number of speakers and was not designed to capture audiovisual or more naturalistic smile behavior. The next steps therefore focus on human–machine comparison using the existing material and on collecting new audiovisual data for a broader investigation of smiled speech.

3. Planned Work for the Future

At the time of submission, I am still in the first year of my PhD. The planned work follows the four research questions in sequence. The completed perception study addresses RQ2, and

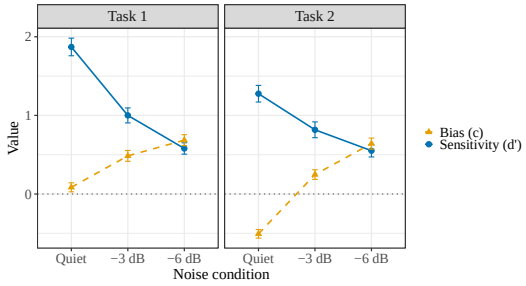


Figure 1: *SDT metrics for two tasks, from [11]. Mean sensitivity (d') and decision criterion (c) are shown as a function of noise condition, with error bars indicating standard errors.*

the next human-machine comparison on AMuS links RQ2 and RQ3 using the same tasks and conditions for human listeners and computational models. The new audiovisual database will address RQ1 and support further work on RQ2 and RQ3. RQ4 is planned for a later stage of the PhD.

Human–machine comparison on AMuS. The completed perception study will be extended by evaluating computational models on the same AMuS stimuli, tasks, and listening conditions. Task 1 will address smile-like detection, contrasting neutral with smile-like speech, whereas Task 2 will address smile-type classification, contrasting amused with spread-lip speech. Baselines will include acoustic-feature pipelines and self-supervised speech representations such as WavLM and wav2vec 2.0 [12, 13]. The aim is not only to test whether models can reach human-level classification accuracy, but also to examine what kind of smile-related information is preserved in different representations and how robust this information is under acoustic masking. Model behavior will therefore be compared with human responses at multiple levels, including performance degradation across noise conditions, category-level confusion patterns, and signal-detection-inspired measures. This will allow the project to ask whether computational systems fail in ways that are merely accuracy-based, or whether their response patterns show comparable changes in uncertainty under noise.

Audiovisual smiled-speech database. The next major output is an audiovisual database involving at least fifty English-speaking adults aged 18-65. The protocol includes posed, elicited, and more spontaneous smiled speech, providing increasing naturalness while retaining experimental control. In the posed condition, participants produce semantically neutral sentences following explicit instructions to speak with or without a smile. In the elicited conditions, neutral and humorous short videos are followed by the same sentence-production task, while additional humorous videos are followed by free-speech retelling. More spontaneous speech is elicited through autobiographical recall of amusing experiences. Synchronized audio and video of the face and upper body are recorded throughout.

Elicitation condition will be treated as contextual information rather than a definitive smile label. Annotation will include elicitation context, participants' funniness ratings of the video stimuli, observable facial smiling, and perceptual judgments of vocal smiling. Visual behavior will provide the primary behavioral reference for observable smile production, while listener judgments will separately capture whether smiling is audible. This allows convergence and divergence between visible and perceived smiling to be examined rather than collapsed into a single ground-truth category.

The in-person recordings will form the core resource, with Prolific potentially supporting larger-scale perceptual judgments. Depending on progress and feasibility, exploratory extensions may include EMG-based measures of smile-related muscle activity and recordings from speakers with Parkinson's disease. These are not core objectives of the PhD.

A later-stage study, to be developed in Year 3, will investigate how smile-related cues affect social impressions such as warmth, friendliness, and trustworthiness. Together, the project is expected to make three main contributions. First, it shows how noise affects smiled-speech perception via both access to smile-related cues and listeners' response strategies under uncertainty. Second, it will provide a multi-purpose audiovisual smiled-speech database. Third, it will develop a computational framework for detecting and analyzing smiled speech across acoustic representations, with particular attention across speakers and communicative conditions.

4. Key Challenges

The first challenge of this thesis is conceptual and concerns variability. Prior work has distinguished between different smile types and functions from the perspectives of emotion, such as felt, false, and miserable smiles, and social signaling, including reward, affiliation, and dominance smiles [14, 1]. This diversity means that smile categories are not discrete acoustic objects. A smile may arise from positive affect, deliberate posing, politeness, embarrassment, or other social goals, and these functions may overlap in speech. In addition, smile-related cues are likely to be expressed relative to a speaker's own neutral baseline rather than as fixed acoustic or visual patterns. Speakers may differ in habitual expressiveness, facial anatomy, and cultural display norms, while speaker characteristics such as gender and language background may also modulate vocal-affect perception [15, 16, 17]. Therefore, it is neither realistic nor theoretically desirable to treat smiled speech as a set of exhaustive and mutually exclusive categories with definitively correct labels. Instead, this project treats smile type as a context-dependent construct, combining elicitation context, observable facial and vocal behavior, speaker self-report, and listener interpretation.

The second challenge is methodological and concerns the measurement of smile intensity. Smiled speech is not simply the addition of a smile signal to a speech signal. Facial posture and articulatory movement can constrain each other. For example, lip spreading during smiling may conflict with the lip closure required for bilabial consonants. Studies show that such conflicts can lead to articulatory compromises, such as labiodentalization, and that speech targets may suppress or temporally reshape smile-related muscle activity [18, 19]. This may make it difficult to infer smile intensity directly from acoustic output or visible mouth movement, especially for automated systems that rely on changes in lip shape as evidence of smiling. Elicitation condition alone also does not guarantee that a speaker would produce the intended smile type or intensity. At the same time, no single measurement provides a perfect ground truth: computer vision captures visible facial movement but may miss or misattribute covert muscle activity, while EMG provides more direct information about muscle activation but could be highly sensitive to the interaction between speech and facial expression. How smile intensity should best be labeled and modeled therefore remains an open methodological question, and the project will consider behavioral, contextual, and perceptual measures as complementary sources of information.

5. References

[1] M. Rychlowska, R. E. Jack, O. G. B. Garrod, P. G. Schyns, J. D. Martin, and P. M. Niedenthal, "Functional smiles: Tools for love, sympathy, and war," *Psychological Science*, vol. 28, no. 9, pp. 1259–1270, 2017.

[2] D. Keltner, "Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame," *Journal of Personality and Social Psychology*, vol. 68, no. 3, pp. 441–454, 1995.

[3] P. M. Niedenthal, M. Mermillod, M. Maringer, and U. Hess, "The simulation of smiles (SIMS) model: Embodied simulation and the meaning of facial expression," *Behavioral and Brain Sciences*, vol. 33, no. 6, pp. 417–433, 2010.

[4] V. C. Tartter, "Happy talk: Perceptual and acoustic effects of smiling on speech," *Perception & Psychophysics*, vol. 27, pp. 24–27, 1980.

[5] H. Barthel and H. Quené, "Acoustic-phonetic properties of smiling revised: Measurements on a natural video corpus," in *Proceedings of the 18th International Congress of Phonetic Sciences*, 2015.

[6] E. Ponsot, P. Arias, and J.-J. Aucouturier, "Uncovering mental representations of smiled speech using reverse correlation," *The Journal of the Acoustical Society of America*, vol. 143, no. 1, pp. EL19–EL24, 2018.

[7] V. Aubergé and M. Cathiard, "Can we hear the prosody of smile?" *Speech Communication*, vol. 40, no. 1, pp. 87–97, 2003.

[8] A. Drahota, A. Costall, and V. Reddy, "The vocal communication of different kinds of smile," *Speech Communication*, vol. 50, pp. 278–287, 2008.

[9] K. El Haddad, I. Torre, E. Gilmartin, H. Cakmak, S. Dupont, T. Dutoit, and N. Campbell, "Introducing AMuS: The amused speech database," in *Statistical Language and Speech Processing*. Cham: Springer International Publishing, 2017, pp. 229–240.

[10] P. Arias, P. Belin, and J.-J. Aucouturier, "Auditory smiles trigger unconscious facial imitation," *Current Biology*, vol. 28, no. 14, pp. R782–R783, 2018.

[11] R. Li, E. Janse, D. Heylen, and K. P. Truong, "Hearing smiles in the crowd: How babble noise shapes smiled speech perception," in *Proceedings of Interspeech 2026*, 2026.

[12] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, and M. Zeng, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[13] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, 2020.

[14] P. Ekman and W. V. Friesen, "Felt, false, and miserable smiles," *Journal of nonverbal behavior*, vol. 6, no. 4, pp. 238–252, 1982.

[15] L. M. Tschinse, A. Asadi, A. Gutnyk, and O. Niebuhr, "Keep on smiling...? an exploration of the gender-specific connections between smiling duration and perceived speaker attributes in business pitches," in *Proceedings of SpeechProsody 2022*, 2022, pp. 624–628.

[16] M. Cuciniello, T. Amorese, A. Alterio, D. Pepe, A. Esposito, O. Scharenborg, and G. Cordasco, "The role of speaker gender in vocal emotion recognition," in *2025 IEEE 16th International Conference on Cognitive Infocommunications (CogInfoCom)*, 2025, pp. 000 121–000 126.

[17] M. D. Pell, S. Paulmann, C. Dara, A. Alasseri, and S. A. Kotz, "Factors in the recognition of vocally expressed emotions: A comparison of four languages," *Journal of Phonetics*, vol. 37, no. 4, pp. 417–435, 2009.

[18] Y. Liu, T. Chan, G. Purnomo, and B. Gick, "Talking while smiling: Suppression in an embodied model of coarticulation," in *Proceedings of the International Seminar on Speech Production*, 2021, pp. 130–133.

[19] Y. Liu, K. Hung, M. Villasenor, S. Colcleugh, E. Chung, and B. Gick, "Timing of perioral muscle suppression in smiled speech," *Canadian Acoustics*, vol. 51, no. 3, pp. 206–207, 2023.

Shrankhla Pandey

University of Cambridge, United Kingdom

Title

Computational Measures of Speech and Language for Schizophrenia Spectrum Disorders

Self-Introduction

Shrankhla Pandey is a final-year PhD candidate at the Department of Psychiatry under the School of Clinical Medicine in the University of Cambridge, and a member of Darwin College. She is supervised by Prof Graham Murray and Dr Sarah Morgan. Her research is in assessing computational measures of speech and language for Schizophrenia Spectrum Disorders. She is keen to connect with mentors and peers working at the intersection of speech technology and healthcare.

PhD Project Source / Funding

This PhD project is funded by WD Armstrong Trust Fund and the Accelerate Program for Scientific Discovery.

Computational Measures of Speech and Language for Schizophrenia Spectrum Disorders

Shrankhla Pandey [1]

[1] Department of Psychiatry, University of Cambridge, UK
[2] Department of Computer Science and Technology, University of Cambridge, UK
sp2147@cam.ac.uk

Abstract

Schizophrenia-spectrum disorders (SSDs) are characterised by disrupted emotional processing, reflected in speech and language. In this work, I investigate automated computational measures of SSDs using the Narrative Emotion Task, a semi-structured interview eliciting emotional narrative speech. I report findings on *emotional alignment* (EA) scores and acoustic features for distinguishing healthy controls from individuals with SSD. I then outline a plan for repeated-measures prediction of symptoms using acoustic, linguistic, and affective speech features across five timepoints, moving towards a speech-based framework for longitudinal symptom monitoring in SSDs.
Index Terms: schizophrenia, schizoaffective disorder, PANSS, speech analysis, acoustic features, natural language processing

1. Introduction

1.1. Background

Schizophrenia-spectrum disorders (SSDs) carry a substantial personal and societal burden [1]. Measuring speech and language aberrations offers a non-invasive, scalable window into the emotional and cognitive disruptions central to these disorders [2, 3, 4]. Speech and language aberrations in SSDs may be more prominent when speaking in an emotional context [5].

Advances in automated speech analysis and natural language processing have provided initial evidence that acoustic, prosodic, and linguistic markers can objectively index psychotic symptoms and differentiate individuals with SSD from healthy controls [6, 7, 8, 9, 10]. However, few studies use speech elicited through personal emotional narrative, despite evidence that speech disturbances in SSDs are particularly reactive to affective conditions [5]. This combination of narrative structure and emotional content remains underexplored as a source of computational markers.

1.2. Data

To address this gap, I use a clinical dataset where speech was elicited in an emotional context [5]. Data were obtained from the MPP-S study (Modified Psychodynamic Psychotherapy for Patients with Schizophrenia - a Randomized Controlled Efficacy Study; ClinicalTrials.gov ID: NCT02576613), conducted by the Charité Berlin and the International Psychoanalytic University.

Speech samples were collected in German using the Narrative Emotion Task (NET) [11], a semi-structured interview assessing social cognition. The NET captures how individuals make sense of emotional events. Trained clinicians administered it without time limits, prompting participants to respond to four emotion prompts in random order: happy, sad, anger, and fear. For each emotion, three questions were posed: a definition (e.g., "What does ⟨emotion⟩ mean?"), a personal narrative (e.g., "Tell me about a time when you felt ⟨emotion⟩."), and a causal attribution (e.g., "Why did that make you feel ⟨emotion⟩?"). On average, the interview took 7.5 minutes [12].

1.3. Hypotheses

The majority of previous automated speech analysis studies in SSD have examined speech elicited on neutral topics, which may be less sensitive to group differences than speech in emotional contexts [5]. Beyond quantification of group differences, speech analysis may provide insights into within patient associations with symptom profiles and/or prognostic utility.

In my PhD research, I analyse group differences in speech, and associations with symptoms cross-sectionally and longitudinally, in the context of a task requiring emotional reflection [13, 10]. It is organised around three research questions:

RQ1: Does *emotional alignment* (EA) [14], an automated measure of social cognition derived from narrative speech in the emotional context of the NET, discriminate healthy controls (HC) from SSD?

RQ2: Can acoustic features of speech elicited in the NET (a) HC from SSD population and (b) schizophrenia (SCZ) from schizoaffective-disorders (SZA)?

RQ3: Can acoustic, linguistic, and emotion features predict individual Positive and Negative Syndrome Scale (PANSS [15]) item scores cross-sectionally and longitudinally, and which features most strongly track clinically observed symptoms?

This submission forms one out of two studies within my PhD thesis.

1.4. Progress Report

I have completed RQ1 and RQ2. RQ3 is in progress.

2. RQ1: Emotional Alignment for Group Classification

Original study: Bradley et al. [14] proposed a rapid method of quantifying abnormal emotion processing: how similar is a participant's emotional response to the typical response to the same stimulus? A deep learning model generated multi-emotion vector per participant; the *emotional alignment* (EA) score reflects the similarity between a participant's emotion vector and the average emotion vector of controls on a continuous scale. Bradley et al.[14] reported impaired EA in patients ($\mu = 0.45, \sigma = 0.15$) relative to controls ($\mu = 0.55, \sigma = 0.13$) in a task eliciting emotional speech.

Given that the NET similarly elicits emotional speech and the MPP-S patient cohort is larger ($n = 95$) than the original study ($n = 37$), I anticipated finding significant group differences in EA score from NET data in the MPP-S study.

Method: NET interviews were automatically transcribed and manually reviewed. I manually segmented transcripts into 4 segments (one per emotion).

Results: At the whole-interview level, EA did not distinguish HC from SSD in the MPP-S cohort ($t = -1.20$, $p = .24$). At the emotion level, only Anger was found significant ($t = 2.81$, $p = .01$), with HC showing higher EA score than SSD, consistent with Bradley et al.'s [14] direction of effect. For Happy, Fear, and Sad, the direction reversed: SSD scored numerically higher than HC, though none of these differences were significant.

Conclusion: Four factors may explain this divergence from Bradley et al. [14]. First, the original study recruited only male patients, while the MPP-S cohort includes both sexes. Second, the small HC sample ($n = 17$) here may have poorly estimated the distribution of typical emotional responses. Thirdly, the paradigm used by Bradley et al, involving viewing of emotional videos, may be more sensitive at eliciting speech differences than reflecting on past emotional experiences as in the NET. Fourthly, Bradley et al. did not stratify their analysis by emotion, so it is unknown whether their discrimination between groups differed by specific emotions.

EA's ability to distinguish SSD from HC in the MPP-S cohort was present only in the Anger condition, which echoes the finding of Docherty et al, 1998 [5].

3. RQ2: Acoustic Classification of Healthy and Diagnostic Groups

For RQ2, I used acoustic features to classify (a) HC vs. SSD and (b) SCZ vs. SZA. The sample comprised 112 participants (17 HC, 70 SCZ, 25 SZA), but one healthy control was dropped due to missing demographic data, leaving 111 for analysis.

Acoustic Features: Features were extracted via Parselmouth/Praat [16] and openSMILE [17]. Filtering a priori based on clinical markers for SSDs and applying Pearson correlation threshold $|r| > 0.8$ resulted in twelve acoustic features for modelling: temporal-prosodic features (speaking rate, pause rate, phonation ratio), affective flattening proxies (fundamental frequency F_0 as a proxy for pitch monotonicity, and intensity standard deviation), laryngeal-motor-control indices (jitter, shimmer, harmonics-to-noise ratio), and articulatory precision measures (first and second formant bandwidths); Spearman correlations reported alongside to guard against outlier sensitivity.

Training: XGBoost [18] classifiers were trained under a nested cross-validation setup to obtain unbiased generalisation estimates. The outer loop used 5-fold `StratifiedKFold` [19]; the inner loop used 3-fold `GridSearchCV` [19] performing a search over the XGBoost hyperparameter grid, selecting models by macro-F_1. For binary and three-class classification task we compared two feature sets: demographics only (age, sex) and demographics combined with the twelve selected acoustic features. Final models were refit on the full sample using the best hyperparameter configuration.

Results: Using age and gender alone, outer CV macro F_1 was 0.45 ± 0.12 (2-class: HC/SSD) and 0.30 ± 0.05 (3-class: HC/SCZ/SZA); adding acoustic features improved these to 0.87 ± 0.21 and 0.57 ± 0.11, respectively. For the three-class task, acoustic features identified HC ($F_1 = 0.86$) and SCZ ($F_1 = 0.77$), but the model struggled to recognise SZA ($F_1 = 0.21$).

Conclusion: Acoustic features in the speech captured under emotional context during the NET task carry signal well beyond age and gender, robustly separating SSD from HC. However, they fail to resolve SCZ from SZA. This difficulty reflects clinical findings: Maj et al. reported inter-rater agreement of only 0.40 for the SCZ/SZA distinction across five centres [20], underscoring that this boundary is inherently unstable.

4. RQ3: PANSS Item Score Prediction

Moving from classification to symptom-level prediction can capture the nuances of speech and language aberrations in emotional context, offering more clinically relevant insight into symptomatology in SDDs [10, 21, 22].

Data: Samples are available across five timepoints: $n_1 = 95$, $n_2 = 95$, $n_3 = 70$, $n_4 = 40$, $n_5 = 70$, enabling repeated-measures analysis. Given the limited sample size, PANSS items will be hand-selected based on established speech-behaviour correlates [23, 24], alignment with the NET's emotional elicitation design, and clinician input [22].

Analyses will be conducted at three granularities: (i) full NET interview, (ii) four emotion segments, and (iii) twelve segments (three questions per emotion).

Method: Analysis proceeds in two steps. First, Pearson and Spearman correlations will be computed between acoustic features, emotion probability scores, and selected PANSS items. Second, a linear mixed-effects model with a random intercept per patient will be fitted separately for each of four feature sets, with regularisation.

Four feature sets will be used: (a) confounders only (age, sex, country of birth); (b) (a) plus six emotion probability scores using a public model [25] ; (c) (a) plus the 11 acoustic features from RQ2; and (d) (a) plus text embeddings, PCA-reduced to 80% explained variance.

Performance will be reported as RMSE under patient-grouped, nested 5-fold cross-validation.

A technical challenge is automatic audio segmentation, with quality control, which is aligned with RQ1 transcript segmentation.

5. Concluding remarks

Results from RQ1 and RQ2 demonstrates that emotionally-elicited speech carries clinically meaningful signal for automated SSD assessment: acoustic features robustly separate SSD from HC, while EA score shows prompt-dependent sensitivity, highlighting the importance of elicitation design in computational psychiatry. RQ3 will extend this to symptom-level prediction, moving towards a longitudinal analysis, generating evidence for a speech-based monitoring framework.

I hope the Doctoral Consortium will provide feedback on both the methodological choices and the translational potential of this work.

6. Acknowledgments

I sincerely thank Prof Graham Murray, Dr Sarah Morgan, Dr Sandra Just, and Dr Nicholas Cummins for making this work possible. I gratefully acknowledge PhD funding from the W.D. Armstrong Trust Fund and the Accelerate Programme for Scientific Discovery.

7. References

[1] W. Rössler, H. Salize, J. van Os, and A. Riecher-Rössler, "Size of burden of schizophrenia and psychotic disorders," *European Neuropsychopharmacology*, vol. 15, no. 4, pp. 399–409, 2005.

[2] H. H. Stassen, M. Albers, J. Püschel, C. Scharfetter, M. Tewesmeier, and B. Woggon, "Speaking behavior and voice sound characteristics associated with negative schizophrenia," *Journal of Psychiatric Research*, vol. 29, no. 4, pp. 277–296, 1995.

[3] Q. Zhao, W.-Q. Wang, H.-Z. Fan, D. Li, Y.-J. Li, Y.-L. Zhao, Z.-X. Tian, Z.-R. Wang, Y.-L. Tan, and S.-P. Tan, "Vocal acoustic features may be objective biomarkers of negative symptoms in schizophrenia: A cross-sectional study," *Schizophrenia Research*, vol. 250, pp. 180–185, 2022.

[4] S. X. Tang, K. Hänsel, Y. Cong, A. H. Nikzad, A. Mehta, S. Cho, S. Berretta, L. Behbehani, S. Pradhan, M. John, and M. Y. Liberman, "Latent factors of language disturbance and relationships to quantitative speech features," *Schizophrenia Bulletin*, vol. 49, no. Supplement 2, pp. S93–S103, 2023.

[5] N. Docherty, M. Hall, and S. Gordinier, "Affective reactivity of speech in schizophrenia patients and their nonschizophrenic relatives," *Journal of Abnormal Psychology*, vol. 107, no. 3, pp. 461–467, 1998.

[6] F. Martínez-Sánchez, J. A. Muela-Martínez, P. Cortés-Soto, J. J. García Meilán, J. A. Vera Ferrándiz, A. Egea Caparrós, and I. M. Pujante Valverde, "Can the acoustic analysis of expressive prosody discriminate schizophrenia?" *The Spanish Journal of Psychology*, vol. 18, p. E86, 2015.

[7] J. Olah, K. Diederen, T. Gibbs-Dean, M. J. Kempton, R. Dobson, T. Spencer, and N. Cummins, "Online speech assessment of the psychotic spectrum: Exploring the relationship between overlapping acoustic markers of schizotypy, depression and anxiety," *Schizophrenia Research*, vol. 259, pp. 11–19, 2023.

[8] N. H. Julia, R. Lewis, C. Ferguson, S. Goldberg, W. Lau, C. M. Swords, G. Valdivia, C. D. Wilson-Mendenhall, R. Tartar, R. W. Picard, and R. Davidson, "Identifying vocal and facial biomarkers of depression in large-scale remote recordings: A multimodal study using mixed-effects modeling," *Interspeech 2025*, 2025.

[9] G. Premananth, Y. M. Siriwardena, P. Resnik, S. Bansal, D. L.Kelly, and C. Espy-Wilson, "A multimodal framework for the assessment of the schizophrenia spectrum," *Interspeech 2024*, p. 1470–1474, Sept 2024.

[10] G. Premananth, P. Resnik, S. Bansal, D. L. Kelly, and C. Espy-Wilson, "Multimodal biomarkers for schizophrenia: Towards individual symptom severity estimation," *Interspeech 2025*, p. 3065–3069, Aug 2025.

[11] B. Buck, K. Ludwig, P. S. Meyer, and D. L. Penn, "The use of narrative sampling in the assessment of social cognition: The narrative of emotions task (NET)," *Psychiatry Research*, vol. 216, no. 3, pp. 404–409, 2014.

[12] S. A. Just, B. Elvevåg, S. Pandey, I. Nenchev, A.-L. Bröcker, C. Montag, and S. E. Morgan, "Moving beyond word error rate to evaluate automatic speech recognition in clinical samples: Lessons from research into schizophrenia-spectrum disorders," *Psychiatry Research*, vol. 352, p. 116690, 2025.

[13] C. M. Corcoran and G. A. Cecchi, "Using language processing and speech analysis for the identification of psychosis and other disorders," *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, vol. 5, no. 8, pp. 770–779, 2020.

[14] E. R. Bradley, J. Portanova, J. D. Woolley, B. Buck, I. S. Painter, M. Hankin, W. Xu, and T. Cohen, "Quantifying abnormal emotion processing: A novel computational assessment method and application in schizophrenia," *Psychiatry Research*, vol. 336, p. 115893, 2024.

[15] S. R. Kay, A. Fiszbein, and L. A. Opler, "The positive and negative syndrome scale (PANSS) for schizophrenia," *Schizophrenia Bulletin*, vol. 13, no. 2, pp. 261–276, 1987.

[16] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing Parselmouth: A Python interface to Praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[17] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[18] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: ACM, 2016, pp. 785–794.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[20] M. Maj, R. Pirozzi, A. M. Formicola, L. Bartoli, and P. Bucci, "Reliability and validity of the DSM-IV diagnostic category of schizoaffective disorder: Preliminary data," *Journal of Affective Disorders*, vol. 57, no. 1, pp. 95–98, 2000.

[21] S. Ciampelli, J. N. de Boer, S. Koops, and I. E. C. Sommer, "Automated speech-based modeling of item-level symptom severity in schizophrenia," *JAMA Network Open*, vol. 9, no. 6, p. e2620239, 2026.

[22] R. He, M. Kirdun, C. Palominos, L. Navarrete Orejudo, S. Barthelemy, S. Bhola, S. Ciampelli, A. Decker, C. Demirlek, R. Fusaroli, J. T. García-Molina, G. Gimenez, R. Hüppi, K. Koelkebeck, A. Lecomte, R. Qiu, A. Simonsen, V. Tourneur, B. Verim, H. Wang, B. Yalincetin, S. Yin, Y. Zhou, M. Amblard, R. Ayesa Arriola, E. Bora, J. de Boer, A. I. Figueroa-Barra, S. Koops, M. Musiol, L. Palaniyappan, A. Parola, F. Spaniel, S. X. Tang, I. E. Sommer, P. Homan, and W. Hinzen, "Predicting PANSS symptoms in schizophrenia spectrum disorders using speech only: an international, multi-centre, retrospective, computational study across multiple languages," *medRxiv*, 2026.

[23] J. N. de Boer, A. E. Voppel, S. G. Brederoo, H. G. Schnack, K. P. Truong, F. N. K. Wijnen, and I. E. C. Sommer, "Acoustic speech markers for schizophrenia-spectrum disorders: a diagnostic and symptom-recognition tool," *Psychological Medicine*, vol. 53, no. 4, pp. 1302–1312, 2021.

[24] Q. Zhao, W.-Q. Wang, H.-Z. Fan, D. Li, Y.-J. Li, Y.-L. Zhao, Z.-X. Tian, Z.-R. Wang, Y.-L. Tan, and S.-P. Tan, "Vocal acoustic features may be objective biomarkers of negative symptoms in schizophrenia: A cross-sectional study," *Schizophrenia Research*, vol. 250, pp. 180–185, 2022.

[25] O. Ring, "Emotion_RoBERTa_german6_v7: A fine-tuned RoBERTa model for emotion classification in German," Hugging Face, 2024, accessed: 2026. [Online]. Available: https://huggingface.co/ringorsolya/Emotion_RoBERTa_german6_v7

Yuxuan Jiang

Tsinghua University, China

Title

Toward a Holistic Modeling Paradigm for Accurate, Versatile, and Controllable Latent Audio Diffusion

Self-Introduction

Second year PhD Student.

PhD Project Source / Funding

My mentor's fundings.

Toward a Holistic Modeling Paradigm for Accurate, Versatile, and Controllable Latent Audio Diffusion

Yuxuan Jiang[1]

[1] Tsinghua University, China

jiangyux25@mails.tsinghua.edu.cn

Abstract

Text-to-audio generation has advanced rapidly, producing remarkably diverse and high-fidelity audio from free-form natural language descriptions. However, these models still offer only coarse control and cannot customize the generated audio as users intend. We therefore aim to build a holistic modeling paradigm for controllable audio generation, extending the control capability across timing, long-form, acoustic, and speech content. We further identify four key research questions spanning data construction, model architecture, inference efficiency, and multi-task co-optimization. To this end, we develop four works, FreeAudio, ControlAudio, AnyAudio, and FreeSonic, that progressively tackle these questions and achieve strong experimental results. In the future, we will develop a more holistic modeling paradigm for accurate and versatile controllable audio generation that resolves these research questions.

Index Terms: audio generation, latent diffusion, controllability

1. Introduction

Latent diffusion models have driven rapid progress in text-to-audio (TTA) generation [1, 2, 3, 4]. By learning to denoise in a compressed latent space, these models synthesize diverse, high-fidelity audio from natural language descriptions (e.g., *"A bird is chirping"*), and are approaching the quality frontier set by dedicated sound design tools. The recent adoption of scalable diffusion transformers has further advanced both the fidelity and efficiency of audio generation [5, 6, 7, 8].

Despite this progress, real-world applications such as film and game sound design demand not merely open-ended synthesis but fine-grained, multi-dimensional controllability. However, prevailing TTA models still struggle to provide such control with the precision and breadth that real-world applications require. We identify four key dimensions that most constrain current TTA performance, namely timing, long-form, acoustic, and speech content. Timing refers to the precise temporal placement and duration of each sound event (e.g., *"a dog barking at 2–4 seconds"*). Long-form denotes the coherent and globally structured generation of extended audio (e.g., *"a two-minute narrative"*). Acoustic concerns the independent manipulation of perceptual attributes such as loudness and timbre (e.g., *"light rain and loud thunder"*, or matching the timbre of a reference recording). Speech content captures the intelligibility and semantic fidelity of spoken audio (e.g., *"a man says 'Today is a sunny day', then a girl responds 'Hello daddy' "*). These controls are not orthogonal. In speech-related audio, the timing of speech content and other events must be aligned [13]. In addition, every control becomes harder as the audio gets longer.

However, existing methods typically support only part of these controls, falling short of a holistic modeling paradigm, as summarized in Table 1. Training-based methods learn control through dedicated modules and supervised objectives, which improve performance but demand large-scale annotated data, dedicated architectures, and training pipelines [14, 15]. Moreover, each new control signal typically calls for retraining or an additional module, limiting their extensibility to new attributes. Training-free methods instead exploit the scalable capabilities of pre-trained audio diffusion models, steering the sampling process at inference time through attention manipulation, latent-space planning, or gradient guidance [16]. This paradigm unlocks more flexible and composable control at near-zero training cost, although it often sacrifices precision and stability compared with its training-based counterpart [17, 18].

We hypothesize that a progressively trained backbone can accurately model heterogeneous controls, while composable attention-, latent-, and waveform-level inference guidance can extend it to new attributes without retraining. Toward this goal, we present four complementary works that investigate these two routes across timing, long-form, acoustic, speech-content, and editing control, providing initial evidence toward a unified, accurate, and extensible framework:

- FreeAudio [19]: a training-free temporal planning approach for timing-controllable long-form audio generation.
- ControlAudio [20]: a training-based approach that progressively integrates text, timing, and phoneme conditions for accurate timing and intelligible audio generation.
- AnyAudio [21]: a training-free method that steers generation via differentiable guidance on attention maps, latent spectra, and waveforms for timing, long-form, and acoustic control.
- FreeSonic [22]: a training-free temporal-aware decoupled attention mechanism for precise and faithful audio editing.

2. Research Questions

The key to holistic controllable audio generation is to support multiple controls simultaneously while minimizing the conflicts among them. However, no existing method can support all these controls simultaneously. This gives rise to four tightly coupled research questions, spanning data construction, architecture design, inference efficiency, and multi-task optimization. We elaborate on each question and its core challenge below:

- RQ1: Training-based control requires costly annotation of multiple signals, especially for long-form audio.
- RQ2: A scalable architecture must encode heterogeneous signals and jointly model all control dimensions.
- RQ3: Training-free control often requires iterative inference-time optimization, increasing sampling cost.
- RQ4: Joint optimization creates task interference, risking the performance of individual controls.

Table 1: *Comparison of existing and our proposed methods across various control aspects and capabilities.* ✓ *and* ✗ *denote whether a capability is supported, while* – *marks general semantic pre-trained models trained on 10-second datasets.*

	Existing Methods						Ours			
	AudioLDM [1]	Stable Audio Open [8]	SaFa [9]	VoiceLDM [10]	Fugatto [11]	T2A-Adapter [12]	FreeAudio	ControlAudio	AnyAudio	FreeSonic
semantic	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
timing	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓
long-form	✗	✓	✓	✗	✓	✗	✓	✗	✓	✓
acoustic	✗	✗	✗	✗	✗	✓	✗	✗	✓	✗
speech	✗	✗	✗	✓	✓	✗	✗	✓	✗	✗
training-free	–	–	✓	✗	✗	✗	✓	✗	✓	✓
efficiency	✓	✓	✗	✓	✗	✓	✗	✓	✗	✓

3. Method

We develop four works to realize this paradigm, all built on a shared latent audio diffusion backbone pre-trained on large-scale 10-second audio-text pairs [23, 24, 25]. FreeAudio leverages LLM planning to decompose a caption into non-overlapping time windows, then decouples and aggregates cross-attention outputs to align each window for precise timing, while fusing overlapping latent segments for coherent long-form generation. AnyAudio steers the sampling trajectory via differentiable losses that operate on cross-attention maps for timing, on latent spectra for long-form coherence, and on decoded waveforms for acoustic attribute matching, all composable within a single inference pass. In contrast, ControlAudio takes the training-based route and recasts controllable generation as progressive diffusion modeling over text, timing, and phoneme conditions. It is supported by a dedicated data construction pipeline, multi-stage progressive training, and coarse-to-fine guided sampling, attaining accurate timing together with intelligible speech. Finally, FreeSonic performs training-free audio editing on a rectified-flow [26, 27] model, confining edits to the target region while preserving the unedited content.

For timing control, we evaluate on the AudioCondition test set [14], which contains 1110 clips with frame-level temporal annotations. We compare against TTA pre-trained models such as AudioLDM [1], AudioLDM 2 [2], Stable Audio Open (SAO) [8], and TangoFlux [7], as well as the training-free method TG-Diff [28] and training-based methods AudioComposer [29] and T2A-Adapter [12]. As shown in Table 2, ControlAudio achieves the best timing control among all methods, while FreeAudio and AnyAudio, despite being fully training-free, achieve competitive timing control and generation quality. For long-form generation, we evaluate on the AudioCaps test set [24] and MusicCaps [30] at target durations of 26 and 90 seconds, comparing against AudioLDM, AudioGen [31], AudioLDM 2, Stable Audio Open, MusicGen [32], and SaFa [9]. As shown in Table 3, both FreeAudio and AnyAudio consistently maintain stable generation quality across extended durations.

These four works each attempt to address different research questions within the latent audio diffusion framework. ControlAudio addresses RQ1 through its data construction pipeline, RQ2 through unified semantic modeling of heterogeneous conditions, and RQ4 through progressive multi-stage diffusion training. FreeAudio and AnyAudio sidestep RQ1 and RQ4 by requiring no training or annotation, though their additional inference cost remains the limitation posed by RQ3. FreeSonic also operates training-free and addresses RQ3 through efficient rectified-flow inversion, though its control capability remains limited. A unified solution that jointly resolves all four research questions within a single framework remains an open challenge and constitutes the primary goal of this thesis.

Table 2: *Evaluation on the AudioCondition test set. For each metric, the best result is bold and the second-best is underlined.*

Method	Eb↑	At↑	FAD↓	KL↓	CLAP↑
Ground Truth	43.37	67.53	-	-	0.377
AudioLDM	9.21	49.35	4.14	2.18	0.378
AudioLDM 2	6.33	47.67	3.00	1.77	0.423
Stable Audio Open	3.37	45.21	3.12	2.87	0.280
TangoFlux	5.93	53.97	2.58	1.86	**0.494**
TG-Diff	26.70	60.06	2.66	-	0.244
AudioComposer-S	43.51	60.83	4.57	1.99	0.284
T2A-Adapter	54.36	68.26	1.87	**1.53**	0.307
FreeAudio	44.34	68.50	1.92	1.73	0.321
AnyAudio	48.62	75.12	**1.78**	1.80	0.427
ControlAudio	**55.58**	**79.42**	2.61	1.85	0.325

Table 3: *Evaluation of long-form generation on the AudioCaps test set and MusicCaps at different output durations.*

Method	Sec.	AudioCaps			MusicCaps		
		FAD↓	FD↓	CLAP↑	FAD↓	IS↑	CLAP↑
AudioLDM	10	8.37	37.27	0.200	6.34	2.49	0.220
AudioGen	10	1.72	**13.29**	0.270	4.55	2.24	0.180
AudioLDM 2	10	1.97	26.34	0.220	2.80	2.84	0.230
SAO	10	3.15	29.01	0.210	3.23	2.94	0.230
AnyAudio	10	**1.48**	19.83	**0.536**	**1.67**	**3.96**	**0.392**
AudioLDM	26	5.99	35.31	0.409	3.64	2.79	**0.427**
AudioGen	26	3.38	19.07	0.338	5.69	2.43	0.318
AudioLDM 2	26	4.26	28.71	0.406	3.16	3.29	0.361
SAO	26	4.17	28.82	0.281	2.94	2.99	0.410
SaFa	26	3.98	23.30	0.494	3.08	3.00	0.407
FreeAudio	26	1.85	**16.21**	0.527	2.46	2.17	0.359
AnyAudio	26	**1.59**	20.98	**0.531**	**1.81**	**3.77**	0.409
AudioLDM	90	4.81	28.49	0.430	3.74	2.74	**0.428**
AudioGen	90	3.38	20.37	0.314	6.12	2.27	0.306
AudioLDM 2	90	7.75	35.94	0.356	5.09	3.34	0.327
SaFa	90	4.08	24.69	0.485	2.34	3.09	0.410
FreeAudio	90	2.19	**19.95**	0.495	2.57	2.29	0.342
AnyAudio	90	**1.74**	22.41	**0.515**	**1.93**	**3.90**	0.391

4. Future Work

While our works have made progress on research questions, several key directions remain open. First, we plan to unify the training-based and training-free routes into one framework that retains the fidelity of the former and the extensibility of the latter, for example by learning a strong controllable backbone and augmenting it with composable training-free guidance for new attributes. Second, we aim to reduce the inference cost of training-free control, exploring techniques such as distillation or sampling acceleration to replace iterative per-step guidance. Third, we will extend the paradigm to further control aspects such as spatial audio, emotion, and speaker identity. Finally, we plan to establish a unified evaluation protocol and benchmark for multi-dimensional controllability, enabling systematic evaluation of compositional control in real creative workflows.

5. Generative AI Use Disclosure

Generative AI tools were used for the sole purpose of editing and polishing the manuscript's language to improve clarity and readability. These tools were not used to generate any scientific content, data, or conclusions. The author reviewed and edited the final manuscript and takes full responsibility for the accuracy and integrity of the work.

6. References

[1] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "Audioldm: Text-to-audio generation with latent diffusion models," *arXiv preprint arXiv:2301.12503*, 2023.

[2] H. Liu, Y. Yuan, X. Liu, X. Mei, Q. Kong, Q. Tian, Y. Wang, W. Wang, Y. Wang, and M. D. Plumbley, "Audioldm 2: Learning holistic audio generation with self-supervised pretraining," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2871–2883, 2024.

[3] R. Huang, J. Huang, D. Yang, Y. Ren, L. Liu, M. Li, Z. Ye, J. Liu, X. Yin, and Z. Zhao, "Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models," in *International Conference on Machine Learning*. PMLR, 2023, pp. 13 916–13 932.

[4] D. Ghosal, N. Majumder, A. Mehrish, and S. Poria, "Text-to-audio generation using instruction guided latent diffusion model," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 3590–3598.

[5] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4195–4205.

[6] Z. Evans, J. D. Parker, C. Carr, Z. Zukowski, J. Taylor, and J. Pons, "Long-form music generation with latent diffusion," *arXiv preprint arXiv:2404.10301*, 2024.

[7] C.-Y. Hung, N. Majumder, Z. Kong, A. Mehrish, A. A. Bagherzadeh, C. Li, R. Valle, B. Catanzaro, and S. Poria, "Tangoflux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization," *arXiv preprint arXiv:2412.21037*, 2024.

[8] Z. Evans, J. D. Parker, C. Carr, Z. Zukowski, J. Taylor, and J. Pons, "Stable audio open," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[9] Y. Dai, C. Wang, C. Li, C. Wang, K. Li, J. Du, L. Sun, J. Gao, R. Wang, and J. Ma, "Latent swap joint diffusion for 2d long-form latent generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 11 006–11 015.

[10] Y. Lee, I. Yeon, J. Nam, and J. S. Chung, "Voiceldm: Text-to-speech with environmental context," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12 566–12 571.

[11] R. Valle, R. Badlani, Z. Kong, S.-g. Lee, A. Goel, S. Kim, J. Santos, S. Dai, S. Gururani, A. Aljafari *et al.*, "Fugatto 1: Foundational generative audio transformer opus 1," in *International Conference on Learning Representations*, vol. 2025, 2025, pp. 4325–4353.

[12] H. Zhu, Y. Xiao, X. Li, Z. Ma, J. Yu, B. Zhang, M. Yang, and X. Chen, "Audio controlnet for fine-grained audio generation and editing," *arXiv preprint arXiv:2602.04680*, 2026.

[13] Y. Dai, K. Wang, B. Gao, Y. Jiang, W. Wang, Q. Ke, and J. Cai, "Cinedub: Scaling end-to-end video dubbing to multi-speaker dialogues with coherent sound effects," *arXiv preprint*, 2026.

[14] Z. Guo, J. Mao, R. Tao, L. Yan, K. Ouchi, H. Liu, and X. Wang, "Audio generation with multiple conditional diffusion model," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 16, 2024, pp. 18 153–18 161.

[15] Y. Dai, Z. Chen, Y. Jiang, B. Gao, Q. Ke, J. Zhu, and J. Cai, "Omni2sound: Towards unified video-text-to-audio generation," *arXiv preprint arXiv:2601.02731*, 2026.

[16] J. Wang, Z. Chen, B. Yuan, K. Zheng, C. Li, Y. Jiang, and J. Zhu, "Audiomog: Guiding audio generation with mixture-of-guidance," *arXiv preprint arXiv:2509.23727*, 2025.

[17] Z. Novack, J. McAuley, T. Berg-Kirkpatrick, and N. J. Bryan, "Ditto: Diffusion inference-time t-optimization for music generation," *arXiv preprint arXiv:2401.12179*, 2024.

[18] Z. Novack, Z. Zukowski, C. Carr, J. Parker, Z. Evans, J. Taylor, T. Berg-Kirkpatrick, J. McAuley, and J. Pons, "Low-resource guidance for controllable latent audio diffusion," in *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2026, pp. 15 072–15 076.

[19] Y. Jiang, Z. Chen, Z. Ju, C. Li, W. Dou, and J. Zhu, "Freeaudio: Training-free timing planning for controllable long-form text-to-audio generation," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 9871–9880.

[20] Y. Jiang, Z. Chen, Z. Ju, Y. Dai, W. Dou, and J. Zhu, "Controlaudio: Tackling text-guided, timing-indicated and intelligible audio generation via progressive diffusion modeling," *arXiv preprint arXiv:2510.08878*, 2025.

[21] Y. Jiang, Z. Chen, A. Wang, Y. Dai, Z. Chen, W. Dou, and J. Zhu, "Anyaudio: Accurate, versatile, and training-free audio control via steering generation," *arXiv preprint*, 2026.

[22] Y. Jiang, M. Han, Y. Dai, A. Wang, T. Zhou, J. Ye, D. Wang, H. Shi, B. Li, J. Song *et al.*, "Freesonic: Training-free temporal-aware decoupled attention for precise audio editing," *arXiv preprint arXiv:2606.15186*, 2026.

[23] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[24] C. D. Kim, B. Kim, H. Lee, and G. Kim, "Audiocaps: Generating captions for audios in the wild," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 119–132.

[25] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 721–725.

[26] Q. Liu, "Rectified flow: A marginal preserving approach to optimal transport," *arXiv preprint arXiv:2209.14577*, 2022.

[27] J. Wang, J. Pu, Z. Qi, J. Guo, Y. Ma, N. Huang, Y. Chen, X. Li, and Y. Shan, "Taming rectified flow for inversion and editing," *arXiv preprint arXiv:2411.04746*, 2024.

[28] T. Du, J. Chen, J. Lu, Q. Xu, H. Liao, Y. Chen, and Z. Wu, "Controllable text-to-audio generation with training-free temporal guidance diffusion," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2024, pp. 1–6.

[29] Y. Wang, H. Chen, D. Yang, Z. Wu, and X. Wu, "Audiocomposer: Towards fine-grained audio generation with natural language descriptions," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[30] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi *et al.*, "Musiclm: Generating music from text," *arXiv preprint arXiv:2301.11325*, 2023.

[31] F. Kreuk, G. Synnaeve, A. Polyak, U. Singer, A. Défossez, J. Copet, D. Parikh, Y. Taigman, and Y. Adi, "Audiogen: Textually guided audio generation," *arXiv preprint arXiv:2209.15352*, 2022.

[32] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, "Simple and controllable music generation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 47 704–47 720, 2023.

Yang Xiao

The University of Melbourne, Australia

Title

From Representation to Retention: Locating and Protecting Knowledge for Continual Speech Adaptation

Self-Introduction

I am currently a PhD candidate at the University of Melbourne in Australia during my second year. I am contributing to building “adaptive”, “efficient”, and “robust” next-generation speech AI systems. At the moment, I mainly work on the multimodal memory of large audio language models and voice agents to support long-term speech interactions.

PhD Project Source / Funding

My research journey is fully supported by the Melbourne Research Scholarship

From Representation to Retention: Locating and Protecting Knowledge for Continual Speech Adaptation

Yang Xiao, The University of Melbourne, Australia

Abstract

Speech foundation models (SFMs) encode broad multilingual and acoustic-linguistic knowledge directly in their parameters. Deploying them in the real world for extending coverage to low-resource languages, new domains, and individual users is a continual learning problem: the model must acquire new capabilities without forgetting existing ones. Yet naive fine-tuning catastrophically overwrites what a model already knows, and external retrieval (RAG) cannot supply a *capability* that must live in the weights. This doctoral research asks how linguistic and semantic knowledge is organized inside speech models, and how that organization can be exploited to acquire new knowledge while retaining existing knowledge. Initial work uncovers a "Semantic Valley" in the Whisper decoder, which is a band of middle layers holding language-agnostic semantic knowledge. Depth-Aware Model Adaptation (DAMA) exploits it to absorb a new language while keeping source-language WER stable, using roughly 80% fewer trainable parameters than LoRA. Future work extends this to continual, multi-step adaptation, speech-native data selection, and beyond-Transformer architectures for unbounded speech streams.

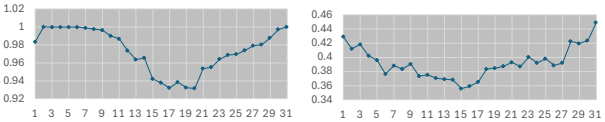
Index Terms: speech recognition, parametric knowledge, continual learning, catastrophic forgetting

1. Introduction

Consider a speech foundation model (SFM) [1, 2] such as Whisper deployed as the multilingual backbone of a transcription or voice-assistant service. Over its lifetime it is repeatedly asked to do something it was not originally trained for: support a newly requested low-resource language, handle a new acoustic domain, or adapt to an individual speaker. Retraining the whole model from scratch on each request is infeasible, yet every incremental fine-tune risks silently erasing languages the model already serves. Deployment is therefore a *continual learning* problem: the model must stay plastic enough to acquire new capabilities while remaining stable enough not to forget old ones.

Two intuitive solutions both fail for this setting. First, externalizing knowledge like the Retrieval-Augmented Generation (RAG) [3, 4, 5] paradigm works when the missing item is a *fact one can look up*, but recognizing a previously unseen language is a *capability* that cannot be retrieved; it has to be encoded in the parameters. Externalizing raw audio also incurs unacceptable latency and privacy risk for real-time use, and transcribing it to text discards the paralinguistic detail that carries part of the signal. Second, naive full-parameter fine-tuning writes new knowledge into the weights, but it indiscriminately overwrites existing knowledge, causing catastrophic forgetting [6].

My dissertation approaches this through representation analysis followed by anti-forgetting adaptation: first locating where linguistic knowledge resides inside an SFM, then designing parameter-efficient adaptation that adds new capabilities without disturbing existing ones, and finally sustaining this over a continual stream of adaptations. The first step is methodologically reminiscent of the locate stage in knowledge editing for large language models; but the objective here is to preserve and extend distributed capabilities, not to rewrite discrete facts, placing the work squarely in continual learning.



(a) *Five Seen Languages.* (b) *Ten Unseen Languages.*

Figure 1: *Layer-wise probing accuracy for different languages.*

2. Research Questions

- **RQ1 (Localization).** How is linguistic and semantic knowledge distributed across the internal layers of a speech foundation model and how can that structure be exploited to adapt the model to a new language or domain?
- **RQ2 (Retention).** How can we extend adaptation to a *continual* setting where many sequential adaptations over shifting distributions and where forgetting compounds across steps?
- **RQ3 (Evolution).** When adaptation never stops over unbounded speech streams, which incoming audio should be retained or discarded under a bounded budget (*data* level), and what architectures can sustain life-long model updating (*architecture* level)?

3. Current Progress & Results (RQ1)

Initial work probes how a base speech model stores knowledge across its internal layers [7, 8]. A layer-wise linear-probing analysis for Language Identification (LID) on the Whisper decoder, as Figure 1 reveals a U-shaped pattern of plasticity: early and late layers carry highly language-specific features, whereas the intermediate layers are largely insensitive to language identity, forming a "Semantic Valley" that holds language-agnostic, shared semantic knowledge [9]. This organization is exactly what naive adaptation destroys. Standard full-parameter fine-tuning (FFT) erases the Semantic Valley and triggers severe catastrophic forgetting: after adapting the model to Kinyarwanda, the Word Error Rate (WER) on the source language English collapses from 11.09% to 105.50%.

To prevent this, we developed **Depth-Aware Model Adaptation (DAMA)** as Figure 2, which uses the localization above to protect shared knowledge while adapting. DAMA intro-

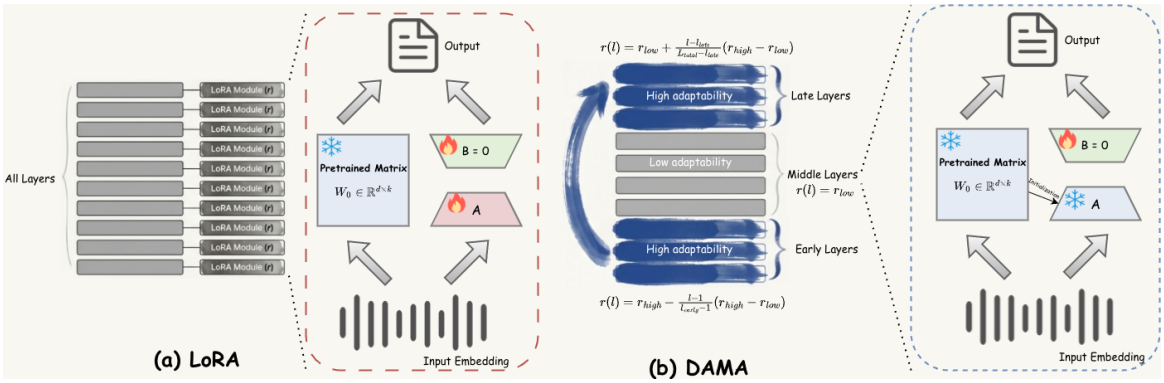


Figure 2: *Overview of the DAMA Framework compared with LoRA. (a) The standard LoRA with uniform rank. (b) The DAMA Framework. Specifically, the Depth-Aware Rank schedule allocates high plasticity to the early and late layers, while the Basis-Protected Projection physically locks the middle layers to protect "Semantic Valley". All layers mean all the layers from the decoder. The decoder processes acoustic embeddings from the encoder and transcribes them into output tokens.*

Table 1: *Performance comparison using Average WER and parameter efficiency on Common Voice and FLEURS. The rightmost column separately reports additional MACs .*

Method	Params (M)	Unseen Languages		Seen-Weak	Average	Extra MACs (G)
		Common Voice	FLEURS	FLEURS		
Fine-tuning	906.5	43.87	50.05	27.55	41.34	-
LoRA [12]	68.2	43.25	47.13	**25.19**	39.71	102.2
DoRA [13]	68.7	43.61	**46.46**	25.24	39.73	1415.4
LoRA-FA [14]	34.1	46.00	48.56	25.65	41.55	51.1
LoRA-XS [15]	**1.3**	57.36	55.23	30.29	50.06	104.2
VB-LoRA [16]	4.9	59.81	50.75	28.00	49.59	758.3
AdaLoRA [17]	51.1	51.66	52.36	28.41	46.02	76.7
DAMA (Ours)	14.9	**43.20**	47.25	25.26	39.73	**22.3**

duces: a *Depth-Aware Rank Schedule* that allocates higher adaptation capacity to the plastic early/late layers and restricts rank in the middle layers; *SVD-based Initialization* that updates only along trailing singular vectors to avoid semantic drift; and *Basis-Protected Projection (BPP)* that physically locks the projection bases within the middle layers. Evaluated across 18 low-resource languages on Common Voice [10] and FLEURS [11], DAMA matches the accuracy of standard LoRA while using roughly 80% fewer trainable parameters as Table 1 shows. Finally, it keeps English WER stable at 12.56% after adaptation, preventing the forgetting seen under FFT.

3.1. Key challenges.

DAMA prevents forgetting in a single, static adaptation step, only for one new language or domain. However, for lifelong use raises three problems it does not face. (i) Applied sequentially, gradient interference between steps lets catastrophic forgetting re-accumulate (RQ2). (ii) Existing coreset/data-selection algorithms are text-centric, relying on lexical semantics and discarding the paralinguistic information that distinguishes speech (RQ3.1, data). (iii) Patching a frozen backbone is limited, since the transformer architecture has no native mechanism for retaining knowledge over time motivating retention built into the model itself (RQ3.2, architecure).

4. Plans for the Future (RQ2 & RQ3)

Continual Tuning (RQ2) The current DAMA framework has been extended into a *Continual Depth-Aware Model Adaptation* framework. By freezing the Semantic Valley while applying dynamic subspace projection and orthogonal gradient updates, the system aims to incrementally absorb new acoustic-linguistic knowledge without erasing prior knowledge. This will be stress-tested on extremely low-resource Pacific Indigenous languages [18] to simulate the large distribution shifts encountered in multi-user, real-world deployment.

Speech-Native Data Selection & Benchmarking (RQ3.1) Rehearsal based methods have been widely used in speech continual learning [19, 20, 21, 22]. A coreset selection algorithm will score "acoustic salience" and "semantic density" to decide, in real time, which raw audio segments carry knowledge worth retaining, summarizing, or discarding. I will intergrate it with the reinforcement learning [23] and treat the anti-forgetting as the reward. A comprehensive benchmark for speech knowledge retention will also be constructed alongside it to estimate the performance [24, 25, 26].

Architectural Innovations (RQ3.2) As a longer-term direction, I will investigate whether retention can instead be built in: a native *memory layer*, in the spirit of intrinsic-memory architectures, that consolidates incoming acoustic-linguistic knowledge into a parameterized store updated as new data arrive [27, 28, 29]. The Semantic-Valley findings from RQ1 would guide its design, where such memory should be placed across depth, and which representational subspaces it should read from and write to distinguishing it from text-centric models. This recasts continual learning from an external constraint into an architectural property, positioned as the aspirational endpoint of the thesis rather than a near-term deliverable.

5. Main Contributions

This thesis reframes adaptable, long-context speech modeling as a *continual learning* problem solved from inside the model. Its contributions are: (i) a representation-level account of where linguistic knowledge resides in a speech foundation model, identifying the *Semantic Valley* in middle decoder layers that hold language-agnostic representations; (ii) Depth-Aware Model Adaptation (DAMA), a localization-guided, anti-forgetting method that absorbs a new language in a single step while preventing catastrophic forgetting at a fraction of the trainable parameters of standard LoRA; and (iii) a roadmap from single-step to lifelong adaptation with a continual extension of DAMA (RQ2), speech-native data selection with a retention benchmark (RQ3.1), and, as the aspirational endpoint, architecture-native retention guided by the Semantic Valley (RQ3.2). Together these chart a scalable route to multilingual speech foundation models that can be continually extended without forgetting what they already do.

6. Generative AI Use Disclosure

We use generative AI tools for polishing the manuscript, e.g., correcting grammar.

7. 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 *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[2] A. Défossez, L. Mazaré, M. Orsini, A. Royer, P. Pérez, H. Jégou, E. Grave, and N. Zeghidour, "Moshi: a speech-text foundation model for real-time dialogue," *arXiv preprint arXiv:2410.00037*, 2024.

[3] Z. Li, X. Xing, J. Xing, H. Hu, H. Lu, and X. Xu, "Long-context speech synthesis with context-aware memory," 2025.

[4] X. Zhang, H. Liu, Q. Zhang, B. Ahmed, and J. Epps, "Speechtrag: Reliable depression detection in llms with retrieval-augmented generation using speech timing information," in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025, pp. 10 019–10 030.

[5] K. Mundnich, A. Lapastora, E. Soltanmohammadi, S. Ronanki, K. Han *et al.*, "Speech retrieval-augmented generation without automatic speech recognition," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[6] L. Della Libera, P. Mousavi, S. Zaiem, C. Subakan, and M. Ravanelli, "Cl-masr: A continual learning benchmark for multilingual asr," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[7] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 914–921.

[8] A. Pasad, B. Shi, and K. Livescu, "Comparative layer-wise analysis of self-supervised speech models," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[9] Y. Xiao, E.-J. Holden, and T. Dang, "Adapting where it matters: Depth-aware adaptation for efficient multilingual speech recognition in low-resource languages," *arXiv preprint arXiv:2602.01008*, 2026.

[10] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the twelfth language resources and evaluation conference*, 2020, pp. 4218–4222.

[11] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "Fleurs: Few-shot learning evaluation of universal representations of speech," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 798–805.

[12] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.

[13] S.-Y. Liu, C.-Y. Wang, H. Yin, P. Molchanov, Y.-C. F. Wang, K.-T. Cheng, and M.-H. Chen, "Dora: Weight-decomposed low-rank adaptation," in *Forty-first International Conference on Machine Learning*, 2024.

[14] L. Zhang, L. Zhang, S. Shi, X. Chu, and B. Li, "Lora-fa: Memory-efficient low-rank adaptation for large language models fine-tuning," *arXiv preprint arXiv:2308.03303*, 2023.

[15] K. Bałazy, M. Banaei, K. Aberer, and J. Tabor, "LoRA-XS: Low-Rank Adaptation with Extremely Small Number of Parameters," *arXiv preprint arXiv:2405.17604*, 2024.

[16] Y. Li, S. Han, and S. Ji, "Vb-lora: Extreme parameter efficient fine-tuning with vector banks," *Advances in Neural Information Processing Systems*, vol. 37, pp. 16 724–16 751, 2024.

[17] Q. Zhang, M. Chen, A. Bukharin, P. He, Y. Cheng, W. Chen, and T. Zhao, "Adaptive budget allocation for parameter-efficient fine-tuning," in *The Eleventh International Conference on Learning Representations*, 2023.

[18] Y. Xiao, A. Mahmudi, N. Thieberger, E. Ambikairajah, E.-J. Holden, and T. Dang, "Continual adaptation for pacific indigenous speech recognition," in *INTERSPEECH 2026*, 2026.

[19] Y. Xiao, X. Liu, J. King, A. Singh, E. S. Chng, M. D. Plumbley, and W. Wang, "Continual learning for on-device environmental sound classification," in *DCASE 2022 Workshop on Detection and Classification of Acoustic Scenes and Events*, 2022.

[20] Y. Xiao, N. Hou, and E. S. Chng, "Rainbow keywords: Efficient incremental learning for online spoken keyword spotting," in *INTERSPEECH 2022*, 2022.

[21] T. Peng and Y. Xiao, "Dark experience for incremental keyword spotting," *arXiv preprint:2409.08153*, 2024.

[22] Y. Xiao and R. K. Das, "UCIL: An Unsupervised Class Incremental Learning Approach for Sound Event Detection," *arXiv:2407.03657*, 2024.

[23] P. Agrawal, I. Shenfeld, and J. Pari, "Rl's razor: Why online reinforcement learning forgets less," 2025. [Online]. Available: https://arxiv.org/abs/2509.04259

[24] Y. Chen, S. Ji, Z. Wang, H. Wang, and Z. Zhao, "InteractSpeech: A speech dialogue interaction corpus for spoken dialogue model," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, Eds. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 8024–8033.

[25] H. Kim, C. H. Lee, S. Park, J. Yeom, N. Park, S. Yu, and S. Yoon, "Does your voice assistant remember? analyzing conversational context recall and utilization in voice interaction models," in *Findings of the Association for Computational Linguistics: ACL 2025*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 8984–9014.

[26] A. Gosai, T. Vuong, U. Tyagi, S. Li, W. You, M. Bavare, A. Uçar, Z. Fang, B. Jang, B. Liu, and Y. He, "Audio multichallenge: A multi-turn evaluation of spoken dialogue systems on natural human interaction," 2025. [Online]. Available: https://arxiv.org/abs/2512.14865

[27] J. Lin, L. Zettlemoyer, G. Ghosh, W.-T. Yih, A. Markosyan, V.-P. Berges, and B. Oğuz, "Continual learning via sparse memory finetuning," *arXiv preprint arXiv:2510.15103*, 2025.

[28] A. Behrouz, P. Zhong, and V. Mirrokni, "Titans: Learning to memorize at test time," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[29] D. Zhang, W. Li, K. Song, J. Lu, G. Li, L. Yang, and S. Li, "Memory in large language models: Mechanisms, evaluation and evolution," *arXiv preprint arXiv:2509.18868*, 2025.

Shreeram Suresh Chandra

Johns Hopkins University, United States

Title

Natural Language as the Interface to Speaking Style: Describing and Controlling How Speech Is Delivered

Self-Introduction

I am a fourth-year Ph.D. candidate advised by Prof. Berrak Sisman at Johns Hopkins University's Center for Speech and Language Processing. My research focuses on the understanding, generation, and evaluation of expressive and emotional speech. I aim to develop natural, flexible, and human-centered speech generation models, while extending these capabilities to assistive communication technologies such as brain-to-speech interfaces.

PhD Project Source / Funding

JHU startup grant

Natural Language as the Interface to Speaking Style: Describing and Controlling *How* Speech Is Delivered

Shreeram Suresh Chandra [1,2]

[1] Johns Hopkins University, USA
[2] The University of Texas at Dallas, USA

schand30@jh.edu, shreeramsureshchandra@gmail.com

Abstract

The expression and perception of human speaking styles are inherently continuous and unbounded in nature. Yet conventional approaches collapse this continuum into discrete labels or rigid taxonomies that are costly to annotate and too coarse to capture how speech actually varies. Natural language offers a compelling alternative: it is open-ended and compositional enough to describe speaking styles at the resolution at which they occur, and it provides a shared medium for both perceiving and producing them. In my work, I explore natural language as an interface to build paralinguistic-informed speech frameworks that can improve speech understanding and expressive speech generation. In the scope of this paper, we present the research questions that motivate us, our contributions and a brief outline on future work.

Index Terms: computational paralinguistics, speaking style captioning, emotion, speech generation

1. Motivation and research questions

Until recently, computational models of speech paralinguistics represented emotion as a small set of discrete categories (happiness, sadness, neutral, anger etc.), grounded in the categorical theories of Ekman [1] and Plutchik [2]. This paradigm is convenient but fundamentally lossy: human expressivity is continuous, compositional, and high-dimensional, and collapsing it into a handful of labels limits how faithfully we can either understand or generate expressive speech.

To move beyond this, a growing body of work has adopted natural language as the medium for paralinguistic information, given that it provides a human-interpretable, open-ended, and continuous representation capable of describing *how* speech is delivered at the resolution at which it actually varies. My thesis investigates natural language as an *interface* to paralinguistic speech: a single medium through which expressive speech can be aligned with language, described, generated under linguistic control, and rigorously evaluated.

Developing this interface is currently held back by two broad, cross-cutting challenges.

(C1) Data scarcity: The lack of diverse, scalable speech corpora with natural-language paralinguistic annotations

(C2) Evaluation: Standard metrics fail to assess open-ended outputs (e.g., free-form captions, style-conditioned audio), lacking reliable measures for quality and faithfulness.

These challenges span our research: data scarcity is addressed by introducing new datasets (RQ1, RQ3) and developing an annotation-free method (RQ2), while open-ended outputs necessitate novel evaluation paradigms (RQ1, RQ3, RQ4).

Research Questions (RQs) :

RQ1 (Speech-natural language alignment). How can we learn joint speech–language embedding spaces that align emotional speech with natural-language style descriptions while preserving the continuous, ordinal structure of emotion, and use them for speech-emotion retrieval?

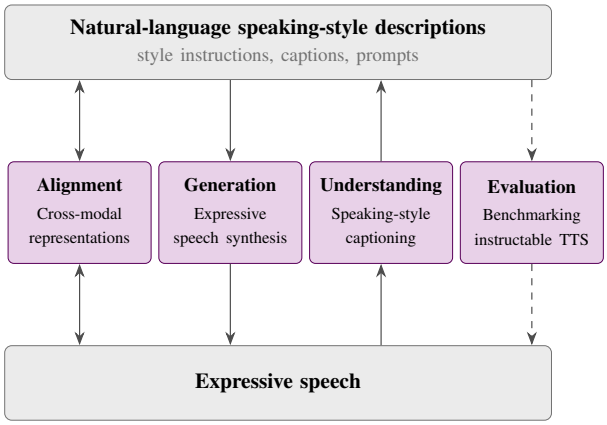


Figure 1: *Natural language as an interface to expressive speech, explored across four dimensions - cross modal alignment, expressive speech synthesis, speaking style captioning, and the evaluation of instruction-guided TTS*

RQ2 (Natural language guided speech generation). How can natural language serve as a control medium for expressive speech generation without relying on scarce, explicitly labeled style data?

RQ3 (Speech understanding via Natural Language). Speech large language models (SpeechLLMs) take speech in and produce free-form text out. How can we steer them to caption paralinguistic speaking style on demand, and how do we evaluate generations that are open-ended rather than categorical?

RQ4 (Benchmarking Natural language guided speech generation). As natural-language-guided speech generation systems proliferate, how do we reliably benchmark their controllability and faithfulness, measuring not just whether instructions are followed, but whether outputs are grounded in the signal?

2. Key contributions

My thesis explores the connection between natural-language speaking-style descriptions and expressive speech across several dimensions and tasks as shown in Fig. 1. In each dimension, we pose key questions (KQs) that push the field forward, while the setup remains constant: the models operate over speech and natural-language speaking-style descriptions, and the goal is to tie the two together faithfully.

The first dimension is **alignment**: how to build joint embedding spaces that align the continuums of dimensional speech emotion and natural-language speaking-style descriptions. The second extends this connection to **generation**, using natural language to control expressive synthesis, with a focus on conversational speech. The third turns to **speech understanding**, examining how large audio–language models perform speaking-style captioning. Finally, we shift to **evaluation**, benchmarking the controllability of instructable text-to-speech (TTS) models.

3. Methods

3.1. Cross-modal alignment

KQ: How can we build a joint embedding space for natural-language speaking-style captions and speech emotion that preserves the continuous, ordinal structure of emotion rather than collapsing it into discrete categories?

In the first dimension, we examine contrastive language–audio pre-training (CLAP) as a means of aligning natural-language speaking-style descriptions with speech emotion. Prior methods naïvely align audio samples with their corresponding text prompts using a symmetric cross-entropy loss [3, 4, 5, 6], treating each pair as an isolated positive. This ignores the relative ordering between emotions and yields a wide audio–text modality gap, which in turn limits fine-grained, inter-emotion understanding. To address this, we adopt the Rank-N-Contrast objective [7] to learn an *ordered* representation space, supervising the geometry with dimensional emotion attributes (valence and arousal) so that embeddings are arranged by emotional similarity rather than merely clustered by pair.

The proposed method, EmotionRankCLAP [8], substantially narrows the modality gap: it reduces the Maximum Mean Discrepancy (MMD) [9] between the audio and text spaces to 0.087 and the Wasserstein distance [10] to 0.065, compared to $0.096/0.180$ for CLAP-SCE and $0.444/0.417$ for SupCon-CLAP (lower is better). The same ordinal structure transfers to downstream speech-emotion retrieval, where EmotionRankCLAP achieves the highest arousal and valence ordinal consistency (Kendall's τ of 0.552 and 0.616), outperforming all CLAP baselines including CLAP-SCE ($0.492/0.505$) and prior models such as CLAP4emo ($0.284/0.533$) and ParaCLAP ($0.283/0.217$).

3.2. Expressive speech generation

KQ: In a natural conversational setting, can we utilize the spoken text itself as context to control the speaking style of a text-to-speech (TTS) system?

In the second dimension, we extend the connection from understanding to generation. Prior emotional TTS methods produce well-articulated emotion only by relying on explicit emotion labels or a stylistically matched reference utterance—neither of which is available in conversational use. We develop TEMOTTS [11], a reference-free framework built on global-style-token codebooks [12] and FastSpeech2 [13], premised on the natural correlation between spoken words and affective prosody: the text itself supplies the emotional control signal.

Among reference-free models, TEMOTTS is both more intelligible and more emotionally faithful. It attains the lowest error rates (CER 4.08%, WER 13.11% versus $8.22/18.57$ for VITS [14] and $6.88/18.34$ for FastSpeech2, is chosen best in 80.95% of Best–Worst Scaling comparisons for emotion-faithfulness (worst in only 1.90%), and achieves the highest MOS (3.75 ± 0.06).

3.3. Speaking style captioning

KQ: Can large audio-language models describe speaking style on demand, generating captions that target a requested stylistic dimension and stay grounded in the speech signal?

Next, we focus on speech paralinguistics understanding. Although large audio-language models excel at tasks like automatic speech recognition and Q&A, paralinguistic captioning lacks a unified task formulation or evaluation paradigm. This ambiguity exists because speaking style is subjective and spans interacting factors like speaker identity, affect, prosody, and contextual delivery. To address this, we reformulate speaking-style captioning (SSC) as an instruction-following audio-language task. By using explicit instructions to specify which aspect of style the model should describe, SSC becomes controllable and tailored to downstream intent, rather than producing a single, generic description.

We introduce *StyleInstructCaps*, the first standardized benchmark for controllable SSC. It evaluates captions across five axes: (1) **metadata groundedness**, (2) **hallucination**, (3) **instruction-following ability**, (4) **speaker-style consistency**, and (5) **generalization to unseen speech and prompts**.

To support this benchmark, we built a multi-style corpus expanding on ParaSpeechCaps (PSC) [15]. This includes 2,900 hours of speech where each utterance is paired with six complementary caption types: speaker identity, contextual delivery, emotion, pragmatics, listener perception, and a holistic summary. These parallel annotations allow us to test whether models can isolate individual stylistic dimensions, integrate multiple cues when required, and generalize across diverse conditions.

Evaluating nine SOTA audio-language models reveals that task-aligned, instruction-diverse training is critical for SSC. Beyond generic large-scale pretraining, this targeted training improves metadata groundedness, reduces hallucination, strengthens speaker-style consistency, and enables robust generalization to new speakers, datasets, and prompting conditions.

4. Current and future research directions

4.1. Evaluating instruction-guided TTS

KQ: As natural-language prompts become the dominant control interface for expressive TTS, how do we reliably evaluate whether generated speech actually realizes the emotion described in the prompt?

In the final dimension, we turn from building natural language-guided systems to *measuring* them. Natural-language prompts offer a continuous, free-form control surface that is well suited to emotional speech, but their open-endedness makes evaluation hard: there is no fixed label to check against, and subjective listening tests are costly and slow. I plan to develop *InstructEmotionEval*, a benchmark that quantifies cross-modal emotion control in instruction-guided TTS along four complementary axes. The benchmark spans four axes: **cross-modal alignment** (does generated speech match the prompted emotion, via CLAP-based retrieval and cosine similarity), **emotion monotonicity** (do ordered prompts yield monotonically ordered acoustic responses), **disentanglement** (does changing one emotion axis leak into others), and **emotion coverage** (how much of the valence–arousal–dominance space is reachable via prompts). Together, these probe whether language is not merely an expressive control interface, but also a *faithful* and *complete* one, thereby extending the theme of groundedness in this thesis.

5. Generative AI Use Disclosure

Generative AI tools were employed solely for language polishing of text written by the authors. These tools were not used to generate scientific content, results, experimental designs, analyses, or conclusions. All authors are responsible for the full content of this paper and consent to its submission.

6. References

[1] P. Ekman, "Are there basic emotions?" 1992.

[2] R. Plutchik, *The emotions*. Bloomsbury Publishing PLC, 1991.

[3] Y. Pan, Y. Hu, Y. Yang, W. Fei, J. Yao, H. Lu, L. Ma, and J. Zhao, "Gemo-clap: Gender-attribute-enhanced contrastive language-audio pretraining for accurate speech emotion recognition," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 021–10 025.

[4] H. Dhamyal, B. Elizalde, S. Deshmukh, H. Wang, B. Raj, and R. Singh, "Prompting audios using acoustic properties for emotion representation," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 936–11 940.

[5] W.-C. Lin, S. Ghaffarzadegan, L. Bondi, A. Kumar, S. Das, and H.-H. Wu, "Clap4emo: Chatgpt-assisted speech emotion retrieval with natural language supervision," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 791–11 795.

[6] X. Jing, A. Triantafyllopoulos, and B. Schuller, "Paraclap–towards a general language-audio model for computational paralinguistic tasks," *arXiv preprint arXiv:2406.07203*, 2024.

[7] K. Zha, P. Cao, J. Son, Y. Yang, and D. Katabi, "Rank-n-contrast: Learning continuous representations for regression," *Advances in Neural Information Processing Systems*, vol. 36, pp. 17 882–17 903, 2023.

[8] S. S. Chandra, L. Goncalves, J. Lu, C. Busso, and B. Sisman, "EmotionRankCLAP: Bridging Natural Language Speaking Styles and Ordinal Speech Emotion via Rank-N-Contrast," in *Interspeech 2025*, 2025, pp. 3000–3004.

[9] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The journal of machine learning research*, vol. 13, no. 1, pp. 723–773, 2012.

[10] G. Peyré and M. Cuturi, *Computational optimal transport: With applications to data science*. Now Foundations and Trends, 2019.

[11] S. S. Chandra, Z. Du, and B. Sisman, "Exploring speech style spaces with language models: Emotional tts without emotion labels," in *Proc. odyssey 2024*, 2024, pp. 194–200.

[12] 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.

[13] 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*.

[14] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International conference on machine learning*. PMLR, 2021, pp. 5530–5540.

[15] A. Diwan, Z. Zheng, D. Harwath, and E. Choi, "Scaling rich style-prompted text-to-speech datasets," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 3639–3659.

Debasish Ray Mohapatra

University of British Columbia, Canada

Title

Physics-Based Simulation of Vowel Utterances using a Biomechanical-Acoustic Model

Self-Introduction

Deb is a final-year PhD student in Electrical and Computer Engineering Department at the University of British Columbia, Canada, where he works in the Human Communication Technologies Lab under the supervision of Prof. Sidney Fels. His doctoral work focuses on biomechanical tongue modeling, articulatory speech synthesis, and developing controllable speech synthesizers as vocal instruments. His broader research interests include speech acoustics, musical acoustics, and human–computer interaction, with an emphasis on exploring new ways of interacting with and controlling vocal sound.

PhD Project Source / Funding

This PhD project is upported by the Natural Sciences and Engineering Research Council of Canada (NSERC)

Physics-Based Simulation of Vowel Utterances using a Biomechanical-Acoustic Model

Debasish Ray Mohapatra [1]

[1] Department of Electrical and Computer Engineering, University of British Columbia, Canada

debasishray@ece.ubc.ca

Abstract

Human speech production is a complex sensorimotor process involving coordinated neural activity in the brain, neuromuscular control of the articulators, and aeroacoustic interactions within the vocal tract. Although several biomechanically driven articulatory–acoustic models have been proposed for speech synthesis, their high computational cost and limited geometric fidelity remain major obstacles to real-time applications. In this dissertation, we first develop a lightweight 2.5D acoustic wave solver that efficiently approximates wave propagation in semi-realistic vocal tract geometries. We then demonstrate its parallel implementation on modern GPUs to achieve real-time performance. Finally, we couple the acoustic solver with a biomechanical tongue model, enabling speech to be synthesized directly from tongue muscle activations and the resulting vocal tract configurations. Collectively, these contributions advance biomechanically driven articulatory modeling toward geometrically realistic, computationally efficient speech synthesis.

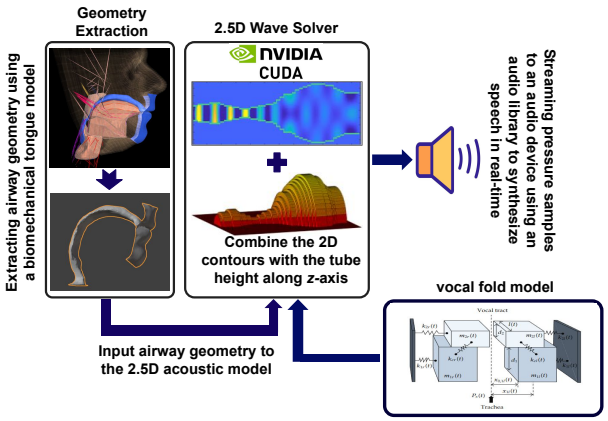


Figure 1: *Modeling pipeline of a biomechanical-acoustic model for real-time speech synthesis.*

1. Introduction

1.1. Motivation and Key Challenges

Vocal tract acoustic models synthesize speech from the geometry of the upper airway, which can be reconstructed from MRIs [1] or CT [2] data. To reduce computational complexity, many low-dimensional vocal tract models [3, 4, 5] represent the airway using a one-dimensional (1D) area function that describes the variation of cross-sectional area along the tract centerline. While such geometric representations are effective for acoustic simulation, directly controlling the shape of a 3D airway or its corresponding area function for speech synthesis remains challenging due to their high DOFs and the absence of biomechanical constraints. As a result, the generated tract configurations may not correspond to physiologically plausible articulatory movements. These limitations motivate the use of biomechanical models [6, 7, 8] that generate vocal tract geometries from physically constrained tongue deformations. However, synthesizing speech using such geometries requires an acoustic model that has low computational cost and high geometric fidelity.

Developing such an acoustic model remains challenging. For example, 3D wave solvers [9, 10] can approximate wave propagation in realistic tract geometries but incur high computational cost. In contrast, 1D solvers [3] enable real-time speech synthesis but offer limited geometric fidelity. They can only model vocal tracts as straight cylindrical tubes. Furthermore, the accuracy of their acoustic response is generally restricted to the plane-wave propagation regime (i.e., up to 5 kHz). Consequently, there remains a need for an acoustic model that balances computational efficiency with geometric realism.

1.2. Research Objective

Figure 1 illustrates the proposed biomechanical–acoustic modeling pipeline. The research is organized around three primary objectives as follows:

- [Completed] **Developing a lightweight acoustic model with high geometric fidelity.** The objective is to achieve a balance between the computational efficiency of 1D models [3] and the geometric realism of 3D solvers [10]. To this end, we propose a 2.5D acoustic model that approximates wave propagation in both the longitudinal and transverse directions while accommodating mirror-symmetric vocal tract geometries.
- [Completed] **Accelerating the 2.5D wave solver using modern GPUs.** The objective is to demonstrate that the proposed 2.5D acoustic model can be efficiently parallelized on modern GPUs. To this end, the time-domain wave equations are implemented using CUDA-based parallel computing, enabling real-time simulation of vocal tract acoustics. The resulting acoustic pressure samples are rendered through an audio device in real time to support speech synthesis.
- [In Progress] **Integrating the 2.5D model with a biomechanical tongue model for speech synthesis.** The objective is to establish a direct link between tongue biomechanics and vocal tract acoustics. To this end, muscle activations generate physically constrained tongue deformations [11] and the corresponding tract geometry. Once the tongue reaches a target configuration, the area function is extracted and supplied to the 2.5D acoustic solver to simulate the resulting acoustic response. This framework enables speech synthesis directly from biomechanically generated tongue configurations.

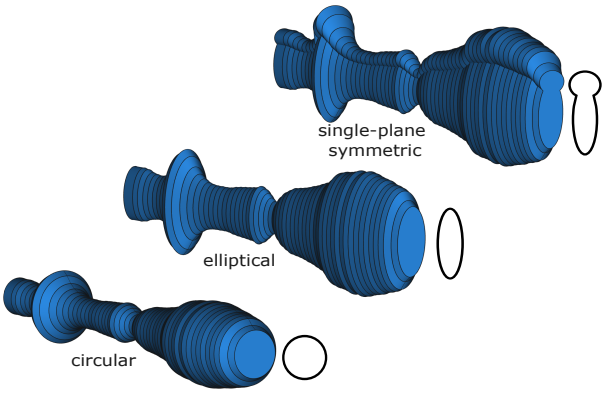


Figure 2: *Modeling pipeline of a biomechanical-acoustic model for real-time speech synthesis.*

1.3. Background Study

In 2D, a vocal tract is represented by a flat contour, typically extracted from the midsagittal slice of a 3D tube built from its area function [5]. However, its practical advantages are limited. As shown by Arnela and Guasch [12] for vowel synthesis, a direct midsagittal representation yields inaccurate formant frequencies and bandwidths compared with 3D simulations. Correcting these discrepancies requires a nonlinear scaling of the tract geometry that substantially narrows constrictions, sometimes by an order of magnitude. These narrow constrictions demand very fine spatiotemporal resolutions, increasing computational cost and preventing real-time performance. Consequently, despite their lower dimensionality, 2D acoustic solvers have not yet provided a practical solution for real-time biomechanically driven speech synthesis. Existing biomechanical–acoustic frameworks instead couple articulatory models with either computationally efficient but acoustically limited 1D solvers or computationally expensive 3D solvers, leaving the integration of biomechanics with real-time 2D acoustics largely unexplored.

2. Methodology

2.1. 2.5D Vocal Tract Model

The 2.5D model builds upon the rationale of 2D vocal tract modeling while overcoming its main limitations. It employs an extended 2D Finite-Difference Time-Domain (2.5D FDTD) solver that simulates acoustic wave propagation in 3D geometries that are symmetric along one dimension. This is achieved by incorporating additional impedance terms derived from the tube depth D, sampled at every grid point to form a depth map. Combined with the midsagittal contour, the depth map enables efficient simulation of 3D vocal tract acoustics using computational resources comparable to those of a conventional 2D solver. Depth maps can be extracted from area functions. Importantly, the 2.5D representation preserves the original vocal tract proportions, eliminating the need for the area function scaling and extremely fine spatiotemporal resolutions required by conventional 2D approaches. At its core, the 2.5D model solves the following wave equations:

$$\frac{\partial p}{\partial t} = -\frac{\rho_0 c^2}{D}\left[\frac{\partial (D v_x)}{\partial x} + \frac{\partial (D v_y)}{\partial y} + \frac{\partial \hat{z}}{\partial t}\right] \qquad (1)$$

$$\beta \frac{\partial \boldsymbol{v}}{\partial t} + (1-\beta)\,\boldsymbol{v} = -\beta^2 \frac{\boldsymbol{\nabla} p}{\rho} + (1-\beta)\,\boldsymbol{v}_b \qquad (2)$$

Here p and $\boldsymbol{v}$ denote the acoustic pressure and velocity, respectively. The term $\hat{z}$ represents the vocal tract surface displacement along the z axis due to yielding walls. The parameters ρ_0 and c represent the air density and speed of sound, respectively. The scalar parameter $\beta = 1$ corresponds to air, while $\beta = 0$ indicates the enforcement of a prescribed particle velocity $\boldsymbol{v}_b$. This velocity can be provided either by an external source, such as an excitation function or through the lossy wall condition. Figure 2 illustrates several vocal tract geometries that can be represented by the 2.5D model, demonstrating its geometric flexibility. The implementation of the 2.5D FDTD wave solver for vocal tract acoustic simulation is publicly available at https://github.com/Debasishray19/VocalSim-2p5D-FDTD

2.2. CUDA Implementation

The CUDA implementation employs a cooperative kernel launch strategy, using grid-level synchronization through CUDA Cooperative Groups to enforce data dependencies between the pressure and velocity update stages within a single kernel launch. To reduce global memory traffic, time-invariant solver coefficients are precomputed and fused during an initialization stage, while shared-memory tiling caches neighboring velocity and depth values required by the finite-difference stencil. Real-time audio output is decoupled from GPU computation through a producer–poller–callback pipeline, in which simulation chunks are written to a pinned-memory ring buffer, transferred to the audio callback via non-blocking CUDA event queries, and downsampled to 44.1 kHz for playback.

3. Results and Discussion

We measured the transfer functions and pressure distributions of vocal tract geometries corresponding to the three cardinal vowels (/a/, /i/ and /u/) using the proposed 2.5D FDTD solver and compared the results with those obtained from a baseline 3D finite-element (FE) model [13]. The results show that the 2.5D solver effectively captures the influence of higher-order modes in mirror-symmetric vocal tract configurations, reproducing resonances and antiresonances beyond the plane-wave propagation regime. The predicted transfer functions and pressure fields closely match those of the 3D model, with only minor discrepancies. These findings demonstrate the geometric flexibility and acoustic accuracy of the 2.5D approach, overcoming key limitations of conventional 1D and 2D vocal tract models. A comprehensive analysis of the acoustic results and computational performance of the proposed 2.5D solver can be found here [4].

4. Future Work

To enable biomechanically driven speech synthesis, the 2.5D acoustic solver will be coupled with the Frank biomechanical tongue model [11]. Muscle activations will be used to generate physically realistic tongue deformations and the corresponding 3D airway geometry. An area-function extraction algorithm will then be applied to the resulting surface airway mesh to derive the vocal tract representation required by the 2.5D model. The extracted geometry will serve as input to the acoustic model, enabling speech synthesis directly from biomechanically generated tongue configurations.

5. Generative AI Use Disclosure

During the preparation of this manuscript, the authors used large language models exclusively for language editing and grammar correction to improve readability and clarity. The authors take full responsibility for all content and confirm that no AI was used for data analysis, results interpretation, or substantive content generation.

6. References

[1] B. H. Story, "Comparison of magnetic resonance imaging-based vocal tract area functions obtained from the same speaker in 1994 and 2002," *The Journal of the Acoustical Society of America*, vol. 123, no. 1, pp. 327–335, 2008.

[2] H. Bakhshaee, C. Moro, K. Kost, and L. Mongeau, "Three-dimensional reconstruction of human vocal folds and standard laryngeal cartilages using computed tomography scan data," *Journal of Voice*, vol. 27, no. 6, pp. 769–777, 2013.

[3] K. van den Doel and U. M. Ascher, "Real-time numerical solution of webster's equation on a nonuniform grid," *IEEE transactions on audio, speech, and language processing*, vol. 16, no. 6, pp. 1163–1172, 2008.

[4] D. R. Mohapatra, V. Zappi, and S. Fels, "A 2.5-dimensional acoustic wave solver: Modeling simplified vocal tract geometries with reduced computational load," *The Journal of the Acoustical Society of America*, vol. 159, no. 4, pp. 3515–3532, 2026.

[5] ——, "A comparative study of two-dimensional vocal tract acoustic modeling based on finite-difference time-domain methods," *arXiv preprint arXiv:2102.04588*, 2021.

[6] S. Dabbaghchian, M. Arnela, O. Engwall, and O. Guasch, "Simulation of vowel-vowel utterances using a 3d biomechanical-acoustic model," *International Journal for Numerical Methods in Biomedical Engineering*, vol. 37, no. 1, p. e3407, 2021.

[7] Y. Payan and P. Perrier, "Synthesis of vv sequences with a 2d biomechanical tongue model controlled by the equilibrium point hypothesis," *Speech communication*, vol. 22, no. 2-3, pp. 185–205, 1997.

[8] S. Buchaillard, P. Perrier, and Y. Payan, "A biomechanical model of cardinal vowel production: Muscle activations and the impact of gravity on tongue positioning," *The Journal of the Acoustical Society of America*, vol. 126, no. 4, pp. 2033–2051, 2009.

[9] D. R. Mohapatra, M. Fleischer, V. Zappi, P. Birkholz, and S. S. Fels, "Three-dimensional finite-difference time-domain acoustic analysis of simplified vocal tract shapes." in *INTERSPEECH*, 2022, pp. 764–768.

[10] T. Vampola, J. Horáček, and J. G. Švec, "Fe modeling of human vocal tract acoustics. part i: Production of czech vowels," *Acta acustica united with Acustica*, vol. 94, no. 3, pp. 433–447, 2008.

[11] P. Anderson, S. Fels, N. M. Harandi, A. Ho, S. Moisik, C. A. Sánchez, I. Stavness, and K. Tang, "Frank: A hybrid 3d biomechanical model of the head and neck," in *Biomechanics of living organs*. Elsevier, 2017, pp. 413–447.

[12] M. Arnela and O. Guasch, "Two-dimensional vocal tracts with three-dimensional behavior in the numerical generation of vowels," *The Journal of the Acoustical Society of America*, vol. 135, no. 1, pp. 369–379, 2014.

[13] P. Birkholz, S. Kürbis, S. Stone, P. Häsner, R. Blandin, and M. Fleischer, "Printable 3d vocal tract shapes from mri data and their acoustic and aerodynamic properties," *Scientific data*, vol. 7, no. 1, p. 255, 2020.