ISCA STUDENT ADVISORY COMMITTEE

11th Doctoral Consortium

Interspeech 2025

TU Delft · The Netherlands · 16 August 2025

Abstract Book

16 Oral Presentations · 5 Poster Presentations · 21 Doctoral Projects

Contents

ORAL PRESENTATIONS

POSTER PRESENTATIONS

Menstrual Cycle Tracking through Voice Features

Anika A. Spiesberger

Technical University of Munich, Germany

Menstrual Cycle Tracking through Voice Features

Anika A. Spiesberger[1,2]

[1]CHI – Chair of Health Informatics, Technical University of Munich University Hospital, Germany
[2]MCML – Munich Center for Machine Learning, Germany

anika.spiesberger@tum.de

Abstract

As personal health monitoring gains traction, understanding the menstrual cycle and its relationship with general health becomes increasingly important. Despite its relevance, female health research remains underrepresented, highlighting the need for more studies and the development of an accurate, accessible, low-effort, low-cost, and non-invasive tracking method. Comparative statistical research shows that voice characteristics change throughout the menstrual cycle. Building on this, this thesis explores whether machine learning methods can identify patterns in these changes that indicate menstrual cycle phases and whether they are consistent across individuals or vary from person to person. To investigate this, two studies will be conducted: a controlled laboratory study and a real-world longitudinal study. Together, these studies will deepen our understanding of how hormonal fluctuations affect the voice and assess the feasibility of voice-based menstrual cycle tracking.

Index Terms: Menstrual Cycle Tracking, Machine Learning, Computational Paralinguistics

1. Motivation and Research Question

Menstrual cycle tracking, as a subcategory of health monitoring, is a part of everyday life for many menstruating people. The reasons for this are as diverse as the methods used. While some want to monitor their general health or estimate the onset of their next period, others use cycle tracking to determine their fertile days – for example, for contraception or when trying to conceive. Still others adjust their diet, physical activity, or work tasks according to the different phases of their cycle.

Depending on the objective, different methods are used for menstrual cycle tracking. The most well-known is the calendar method, which estimates cycle phases either by counting forward, based on past cycles, or retrospectively backwards. This method assumes that the menstrual cycle is regular; however, for many menstruating people, this is not true. Although this approach may be sufficient for a rough estimate of the next period's onset, its predictive power is limited.

More precise methods are needed if additional information is required – for example, the time of ovulation for pregnancy planning or prevention or the specific cycle phase (follicular/luteal). These methods are based on hormonal and physical changes that occur during the cycle, such as the rise in basal body temperature at ovulation (which can only be confirmed retrospectively), changes in the consistency of the cervical mucus, and the detection of the 'luteinizing hormone (LH) surge' (which triggers ovulation). The LH surge can be assessed using test strips, but these are costly and produce considerable waste.

Given the limitations of the current methods – such as irregular cycles, retrospective insight, material costs, and waste generation – there is a need for alternative approaches that are accurate, accessible, low-effort, low-cost, and non-invasive. One such method, which is mostly unexplored, is using voice as a biomarker for cycle tracking. Various studies have shown that changes in vocal characteristics occur throughout the cycle. For example, Bryant and Haselton [1] and Fischer *et al.* [2] found an increase in the fundamental frequency (F0) before ovulation. Fischer *et al.* [2] additionally reported an increase in the ratio of unvoiced parts and the noise-to-harmonic ratio (NHR) during menstruation. Arruda *et al.* [3] observed a correlation between low oestrogen levels and reduced F0, as well as increased vocal tension, instability, and jitter. Shoup-Knox *et al.* [4] found a significant decrease in shimmer during high fertility phases.

However, not all findings align, which could be due to small sample sizes, differences in study design (e. g., how phases were determined), variations in feature extraction methods, and speaker idiosyncrasies. For instance, Karthikeyan and Locke [5] reported a lower F0 during the high fertility phase. Similarly, Tatar *et al.* [6] found lower F0 and higher NHR before menstruation than during and after. They also observed increased jitter and shimmer before menstruation than after. Further, Pavela Banai [7] found that F0 minimum was lower during menstruation than in the late follicular phase and that vocal intensity was reduced in the luteal phase compared to menstruation and the late follicular phase.

While these studies provide valuable insights into how voice changes over the menstrual cycle, they have primarily used comparative analyses to investigate differences between the phases rather than predictive modelling. Little attention has been given to the potential of machine learning (ML) methods for menstrual cycle tracking. Unlike traditional statistical analysis, which examines differences between the cycle phases, ML can detect complex, non-linear patterns in vocal features. This could enable the classification of cycle phases, the prediction of ovulation, and even hormone level estimation.

This doctoral thesis aims to explore whether ML algorithms can be used to track and predict menstrual cycle phases. The **Research Questions** are:

RQ1 Are there vocal features that show consistent changes across the menstrual cycle for all individuals? Can ML methods use these features to classify the cycle phase accurately?

RQ2 Are there individual-specific patterns in voice changes throughout the cycle? Can personalisation methods improve the accuracy of the ML models?

RQ3 Can ML models predict the time until ovulation and/or the onset of the next period based on voice features?

RQ4 Can ML models estimate hormone levels from voice features?

2. Progress Report

We conducted an initial pilot study using a dataset from Pavela Banai [7], which included 44 naturally cycling women, each recorded in three phases of her cycle. Eight vocal features were provided for each recording. We trained several ML classifiers using these features to distinguish between the menstrual and late follicular phases. Despite the limited feature set, we achieved a mean accuracy of 60 %, suggesting that voice-based menstrual cycle phase recognition has potential and should be further investigated. For a full account, see [8].

3. Workplan

The next step is to collect new data that will allow us to work directly with the audio recordings and extract additional features. We plan two studies to investigate the research questions.

First Study – Laboratory: This study aims to establish a baseline of vocal changes throughout the menstrual cycle under controlled conditions. For this, we will recruit 44 participants between 18 and 45 years who have regular cycles, do not use hormonal contraceptives, and do not show menopausal symptoms. Recordings will be made in three distinct phases of their menstrual cycle – early follicular (menstruation), late follicular (ovulation), and late luteal. These phases were selected because they represent the most hormonally distinct times in the menstrual cycle. The phases will be determined using the calendar method and ovulation tests.

Participants will visit the laboratory on the appropriate dates and perform several tasks, including sustained vowel production, text reading, and free speech. Recordings will be analysed using acoustic features extracted with Praat as well as various feature sets (e. g., EGEMAPS [9], LIWC [10]) and a combination of statistical significance analysis and ML algorithms (e. g., SVM, XGBOOST). We will also explore possible improvements through personalisation techniques. The study has been approved by the responsible ethics committee and will be carried out in Mid 2025.

Second Study – Field: This study will investigate whether voice-based cycle tracking is feasible in real-world settings and explore individual-specific changes. For this, we plan a longitudinal study in which participants will provide daily voice samples using their smartphones across three consecutive cycles. Saliva tests will be used to assess the phases, as they provide accurate hormone levels and can be stored in the participants' freezers.

4. Challenges

Challenges arise mainly due to the nature of the studies.

Confounding Factors: There are various potential confounding factors such as mood, illness, and substance use. While we will try to reduce their influence, we will also need to account for them during analysis using self-reported data.

Time of Day and Experimenter Effects: In the laboratory study, the presence of different experimenters (especially those of different genders) and variations in recording time could introduce inconsistencies. While we will standardise recording times as much as possible, some fluctuations are unavoidable.

Noise and Recording Conditions: In the field study, recordings will be made in different environments with personal smartphones and, therefore, varying microphones. To control for this, participants will be asked to record in quiet settings and start with a clapping sound to assess room acoustics. We will also collect metadata on the recording devices to investigate hardware-related variability.

Funding for Saliva Tests: For the second study, we plan to do daily saliva sampling over 90 days. If funding is unavailable, we will use basal body temperature to estimate ovulation instead. While this is more affordable, it will not allow us to determine precise hormone levels to investigate *RQ4*.

Participant Recruitment and Compliance: Recruiting participants and ensuring their compliance is challenging, particularly in the longitudinal study. Interest seems high, especially if we provide hormone test results. If saliva testing is not possible, we will offer an alternative incentive. We also plan to conduct speaker identification tests to ensure that all audio recordings are from the same person.

Inclusivity vs Sample Size: Menstruation affects not only cis women. We aim for inclusivity while maintaining a reasonable sample size. We will exclude individuals using hormonal contraceptives, as these alter natural hormonal fluctuations.

Ethical and Privacy Concerns: Ethical approval has been obtained for the first study and will be obtained for the second. Confidentiality and informed consent are priorities, ensuring participants fully understand how their data will be used.

5. Discussion of (Expected) Results

Investigating whether voice recordings can be used to track and predict menstrual cycle phases will provide new insights into how hormonal fluctuations affect voice features. This research will provide insight into the changes in vocal features across the menstrual cycle and whether these are consistent across individuals or vary from person to person. This could give rise to a whole new way of menstrual cycle and fertility tracking. However, while we see many opportunities for using voice to track hormonal fluctuations, further research will be necessary. First, we will focus on naturally cycling individuals, as hormonal contraceptives alter hormone levels and suppress the natural fluctuations that influence voice features. Therefore, our research will not account for the changes in hormone levels due to hormonal contraceptives. Second, while we will aim for diversity in our sample, some criteria, such as language, vocal training, and smoking behaviour, are necessary to reduce confounding variables and keep the sample size reasonable. Third, we focus our research on tracking and predicting cycle phases. While we will determine them by assessing the levels of oestrogen and progesterone, we will not record other hormones. Therefore, the next step could be to expand the research to include further sex hormones, to investigate hormonal imbalances which can be indicative of a variety of health issues.

6. Contribution

In line with this year's Interspeech theme – 'Fair and Inclusive Speech Science and Technology' – we believe that investigating voice changes throughout the menstrual cycle is essential, as these changes affect a group that is often overlooked or neglected. Research in this area can help raise awareness of these changes and highlight the importance of their consideration for other fields, such as personalised applications and voice recognition tools. This could also help make cycle tracking more precise, benefiting both individuals and researchers by providing an accurate, accessible, low-effort, low-cost, and non-invasive method to determine and predict cycle phases. Additionally, we plan to make the data from the second study available (on request), enabling other researchers to advance this field further.

7. References

[1] G. A. Bryant and M. G. Haselton, "Vocal cues of ovulation in human females," *Biology Letters*, vol. 5, no. 1, pp. 12–15, 2009.

[2] J. Fischer, S. Semple, G. Fickenscher, R. Jürgens, E. Kruse, M. Heistermann, and O. Amir, "Do women's voices provide cues of the likelihood of ovulation? the importance of sampling regime," *PLOS ONE*, vol. 6, no. 9, pp. 1–8, Sep. 2011.

[3] P. Arruda, M. R. D. da Rosa, L. N. A. Almeida, L. de Araujo Pernambuco, and A. A. Almeida, "Vocal acoustic and auditory-perceptual characteristics during fluctuations in estradiol levels during the menstrual cycle: A longitudinal study," *Journal of Voice*, vol. 33, no. 4, pp. 536–544, 2019.

[4] M. L. Shoup-Knox, G. M. Ostrander, G. E. Reimann, and R. N. Pipitone, "Fertility-dependent acoustic variation in women's voices previously shown to affect listener physiology and perception," *Evolutionary Psychology*, vol. 17, no. 2, p. 1 474 704 919 843 103, 2019.

[5] S. Karthikeyan and J. L. Locke, "Men's evaluation of women's speech in a simulated dating context: Effects of female fertility on vocal pitch and attractiveness," *Evolutionary Behavioral Sciences*, vol. 9, no. 1, p. 55, 2015.

[6] E. C. Tatar, M. Sahin, D. Demiral, O. Bayir, G. Saylam, A. Ozdek, and M. H. Korkmaz, "Normative values of voice analysis parameters with respect to menstrual cycle in healthy adult turkish women," *Journal of Voice*, vol. 30, no. 3, pp. 322–328, 2016.

[7] I. Pavela Banai, "Voice in different phases of menstrual cycle among naturally cycling women and users of hormonal contraceptives," *PLOS ONE*, vol. 12, no. 8, pp. 1–13, 2017.

[8] A. A. Spiesberger, A. Mallol-Ragolta, A. Triantafyllopoulos, and B. W. Schuller, "Towards predicting menstrual cycle phases exploiting paralinguistic features," in *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2024, pp. 1–4.

[9] 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.

[10] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of Language and Social Psychology*, vol. 29, no. 1, pp. 24–54, 2010.

The Role of Multimodality in Predictive Turn-taking

Sam O'Connor Russell

Trinity College Dublin, Ireland

The Role of Multimodality in Predictive Turn-taking

Sam O'Connor Russell

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

russelsa@tcd.ie

1. Background and Motivation

1.1. Predictive turn-taking

Turn-taking is *predictive* as listeners plan their next turn while the speaker is still speaking [1, 2]. However, most turn-taking algorithms for human-robot interaction *react* to the silence after a turn [3]. This leads to human-robot interaction that is less fluid than human interaction [4, 5]. Predictive turn-taking models (PTTMs) overcome this limitation by learning human-like turn-taking from corpora of human interaction. They continuously predict shifts between speakers [6, 7, 8, 4], akin to human turn-taking [9].

1.2. Gaps in the literature

We have identified a number of gaps in the turn-taking literature as part of this thesis, currently in its final year.

1.2.1. The use of manual transcription in narrow domains

PTTMs require annotations of each speech segment for training [7, 10, 11]. This typically consists of utterance boundaries derived from manual transcription [7] which include nonlexical aspects of the speech signal e.g. filled pauses [12]. Much of the work in the turn-taking literature has hence focussed on telephone speech using the manually transcribed Switchboard corpus [13, 14, 10, 11]. Automatic speech recognition (ASR) can vastly increase the amount of data available for PTTM training, enabling new domains to be considered. However, ASR transcription PTTM development has not been explored and the literature lacks detail about ASR transcript quality for natural conversational interaction.

1.2.2. The restriction of models to a single modality

When interlocutors can see one another, multimodal cues including syntax, prosody and gaze support turn-taking [15]. However, state-of-the-art PTTMs rely exclusively on either audio [11, 10] or text alone [16, 17]. It is unclear how including multiple modalities impacts PTTM performance.

1.2.3. No analysis of performance in noise

PTTMs will be deployed *in-the-wild*, where they will inevitably encounter diverse sources and levels of noise. PTTMs have to date only been tested on interactions free from background noise [3, 18]. The impact of noise on PTTM performance is therefore unknown. Access to visual cues assists human language comprehension in noise [19]. It is unknown whether a multimodal PTTM would be able to similarly exploit the visual modality.

1.2.4. Limited analysis of PTTM sensitivity to multimodal cues

Humans make use of important cues from multiple modalities in turn-taking [20]. Prosody enables a human listener to determine the end of a question with multiple points of completion: "are you a student / here / at this university?". PTTMs have been shown to be sensitive to prosodic cues [8]. However, there are many cues a PTTM might exploit which remain unexplored. Notably, nonverbal cues from the listener remain unexplored, as most PTTMs do not use any visual features.

1.3. Research Questions

We have therefore proposed the following research questions:

- **RQ.1** Can turn-taking models be trained using automatically transcribed data?
- **RQ.2** Can visual cues improve the performance of predictive turn-taking models?
- **RQ.3** Can predictive turn-taking models effectively use multimodal cues in a human-like manner, e.g. to improve performance in noise and to resolve lexical ambiguity?

2. Results to Date

2.1. Conversational ASR

We conducted an evaluation of 3 commercial and 3 open source ASRs on conversational speech [21]. Commercial ASR had half the WER of open-source ASR 7.9-9.4% vs 19.4-34.1 % WER). The most frequent errors were non-lexical words (uhm, ehm). Commercial ASR preserved 80% of these words, compared with only 20% in open-source. We published our analysis [21].

2.2. Audio-only turn-taking

We focus on predicting turn-taking in dyadic human-human interaction. We re-implemented and re-trained the VAP model, a state-of-the-art audio-only turn-taking model [10, 11]. It uses the timings of speech segments extracted from a time-aligned transcription. Table 1 shows the model performance on distinguishing holds (no speaker change) from shifts (speaker change) on Switchboard (260 hr telephone) and Candor corpora (870 hr videoconference) corpora [22]. The impact of ASR is small, validating its use to train and test PTTMs.

2.3. Including the visual modality in turn-taking

Next, we conducted a cross-corpus analysis. When training on Candor and deployed on Switchboard, we found minimal change in F_1 hold (0.89 to 0.88), but F_1 shift score dropped considerably (0.71 to 0.27). We analysed facial action units and found enhanced listener lip, jaw and brow movement immediately before

Table 1: *Average shift/hold prediction performance of the VAP model on Candor (CND, videoconference) and Switchboard (SWB, telephone) evaluated during silence between turns.*

Corpus	Alignment	F_1 (Weighted)	F_1 (Hold)	F_1 (Shift)	Accuracy (Balanced %)
SWB	ground-truth	0.82	0.89	0.47	67
SWB	ASR	0.81	0.89	0.45	65
CND	ASR	0.83	0.89	0.71	79

a turn. Our hypothesis is that non-verbal behaviours form vital turn-taking cues when interlocutors can see one another. This hence changes the way interlocutors use the audio channel, making a PTTM trained on videoconferencing speech insufficiently accurate in a telephone speech deployment.

We proposed MM-VAP, a novel multimodal turn-taking model which incorporates speech as well and non-verbal facial expression, gaze and head pose features [23]. Each speaker's audio and visual features are modelled by transformer blocks with self- and cross-attention. We found the inclusion of visual cues led to a significant performance increase above the audio-only model across a range of metrics (hold/shift prediction in Table 2). Notably, shift prediction benefited the most from the inclusion of visual cues. We conducted an ablation study which shows that facial action unit features were consistently the largest contributors to performance across all types of turn shift. We published MM-VAP and our analysis [24, 23].

Table 2: *Average shift/hold prediction across 5 folds on the Candor corpus (best in bold). a = audio only, v = video-only, and a+v = multimodal; e = early fusion, l = late fusion. Percentage change relative to the audio-only model is provided.*

Model	F_1 (Weighted)		F_1 (Hold)		F_1 (Shift)		Accuracy (Balanced %)	
a	0.83		0.88		0.70		79	
v	0.72	↓14%	0.82	↓8%	0.47	↓33%	68	↓15%
a+v (e)	0.84	-	0.90	-	0.74	↑5%	82	↑3%
a+v (l)	**0.86**	↑3%	**0.90**	↑2%	**0.74**	↑6%	**83**	↑4%

2.4. Robust turn-taking in noise

Next, we tested both VAP and MM-VAP by adding babble, music and speech (random overdubbed utterances) noise to the Candor corpus. Our results showed that both VAP and MM-VAP were surprisingly brittle, with hold/shift prediction accuracy dropping to just 51% accuracy in +10dB of music noise. Adjusting training to include examples of noisy data improved performance significantly to 64-75% (VAP and MM-VAP, respectively). The performance of MM-VAP remained consistently higher, equivalent to a +10 dB gain in signal-to-noise ratio, illustrating the use of multimodal features to overcome noise. Performance of models trained exclusively on babble and music did not generalise to other sources e.g. 51% on speech and music. However, models trained on speech noise performed well on types of noise unseen during training (70-76% at 0dB SNR). Mixing in random speakers appears to 'force' MM-VAP to use visual information, rather than relying on the acoustic properties of the noise. We published this work [25].

We show VAP and MM-VAP model outputs for a shift between two speakers in Figure 1. Note how the model favours speaker 1 (red) before the culmination of speaker 0's turn (grey). Note the strong performance of MM-VAP in 0dB babble noise, when compared with the audio-only VAP model.

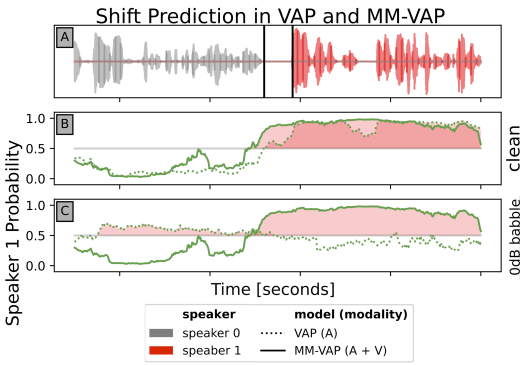
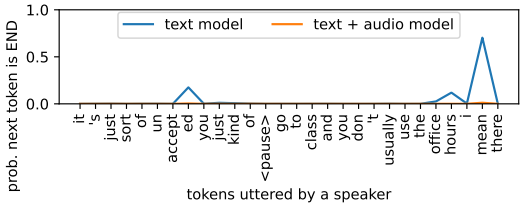


Figure 1: *Model output during a shift in the Candor corpus (unseen during training). Speaker 0 (grey): "So uhm would you feel comfortable telling me about what you do for a living?", Speaker 1 (red) "Uhm yeah well for the most part I test and tweak algorithms". VAP (audio) and MM-VAP (audio+video) model output is shown in clean speech (B) and 0 dB babble speech (C)*

3. Work in Progress

We are currently working on LLM-based turn-taking models. The below figure shows the outputs of two GPT-2 turn-taking models on Switchboard. The figure shows a state-of-the-art text-only model [26] incorrectly predicting the end of turn at points of syntactic completion mid-utterance.



We have proposed a GPT-2 based LLM model which incorporates an acoustic front-end [27], and shows the output in orange. The inclusion of the acoustic features resolves the lexical ambiguity, as no turn-shift is predicted. We plan to further extend this model to incorporate a visual front-end.

4. Challenges and Future Work

A challenge in our work is determining precisely which multimodal cues our PTTMs is sensitive to. We are conducting an analysis of known turn-taking cues (F0, vowel lengthening, listener gaze, etc.), and assessing whether the LLM-based models use them to resolve lexical ambiguity. We plan to submit this work in August, and the thesis will be written Sept - Dec 2025 with the defence in Feb 2026 (approximate timeline).

5. Contributions

Our main contributions and publications are briefly:

- First comprehensive analysis of ASR for conversational speech transcription [21]
- Validation of the suitability of ASR for PTTM training
- Extension of PTTM to include visual features [23, 24]
- First analysis of PTTM performance in noise [25]
- *Analysis of multimodal PTTM senstivity to known human turn-taking cues* (in progress, planned publication)

6. Acknowledgements

This research was conducted with the financial support of Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2 at the ADAPT SFI Research Centre at Trinity College Dublin.

7. References

[1] S. C. Levinson and F. Torreira, "Timing in turn-taking and its implications for processing models of language," *Frontiers in psychology*, vol. 6, p. 731, 2015.

[2] S. Garrod and M. J. Pickering, "The use of content and timing to predict turn transitions," *Frontiers in psychology*, vol. 6, no. 751, pp. 1–12, 2015.

[3] G. Skantze, "Turn-taking in conversational systems and human-robot interaction: a review," *Computer Speech & Language*, vol. 67, p. 101178, 2021.

[4] S. Li, A. Paranjape, and C. D. Manning, "When can i speak? predicting initiation points for spoken dialogue agents," *arXiv preprint arXiv:2208.03812*, 2022.

[5] A. Woodruff and P. M. Aoki, "How push-to-talk makes talk less pushy," in *Proceedings of the 2003 ACM International Conference on Supporting Group Work*, 2003, pp. 170–179.

[6] G. Skantze, "Towards a general, continuous model of turn-taking in spoken dialogue using lstm recurrent neural networks," in *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, 2017, pp. 220–230.

[7] M. Roddy, G. Skantze, and N. Harte, "Investigating speech features for continuous turn-taking prediction using lstms," in *Interspeech 2018*, 2018, pp. 586–590.

[8] E. Ekstedt and G. Skantze, "How much does prosody help turn-taking? investigations using voice activity projection models," in *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2022, pp. 541–551.

[9] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *language*, vol. 50, no. 4, pp. 696–735, 1974.

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

[11] K. Inoue, B. Jiang, E. Ekstedt, T. Kawahara, and G. Skantze, "Multilingual turn-taking prediction using voice activity projection," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia, May 2024.

[12] B. Jiang, E. Ekstedt, and G. Skantze, "What makes a good pause? investigating the turn-holding effects of fillers," in *20th International Congress of Phonetic Sciences (ICPhS). August 7-11 2023, Prague, Czech Republic*. International Phonetic Association, 2023, pp. 3512–3516.

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

[14] M. Roddy, G. Skantze, and N. Harte, "Multimodal continuous turn-taking prediction using multiscale rnns," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018, pp. 186–190.

[15] J. Holler, K. H. Kendrick, M. Casillas, and S. C. Levinson, *Turn-taking in human communicative interaction*. Frontiers Media SA, 2016.

[16] E. Ekstedt and G. Skantze, "Turngpt: a transformer-based language model for predicting turn-taking in spoken dialog," *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.

[17] S. Leishman, P. Bell, and S. Wallbridge, "Pairwiseturngpt: a multi-stream turn prediction model for spoken dialogue," in *Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue*, 2024.

[18] K. Onishi, H. Tanaka, and S. Nakamura, "Multimodal voice activity prediction: Turn-taking events detection in expert-novice conversation," in *Proceedings of the 11th International Conference on Human-Agent Interaction*, 2023, pp. 13–21.

[19] A. MacLeod and Q. Summerfield, "Quantifying the contribution of vision to speech perception in noise," *British journal of audiology*, vol. 21, no. 2, pp. 131–141, 1987.

[20] J. Holler and S. C. Levinson, "Multimodal language processing in human communication," *Trends in Cognitive Sciences*, vol. 23, no. 8, pp. 639–652, 2019.

[21] S. O. Russell, I. Gessinger, A. Krason, G. Vigliocco, and N. Harte, "What automatic speech recognition can and cannot do for conversational speech transcription," *Research Methods in Applied Linguistics*, vol. 3, no. 3, p. 100163, 2024.

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

[23] S. O. Russell and N. Harte, "Visual cues enhance predictive turn-taking for two-party human interaction," *Findings of the Association for Computational Linguistics: ACL*, 2025.

[24] ——, "Towards multimodal turn-taking for naturalistic human-robot interaction," *Multimodal Symposium on Human Interaction, Frankfurt, Germany*, 2024.

[25] ——, "Visual cues support robust turn-taking prediction in noise," *Proceedings of Interspeech*, 2025.

[26] J. Lei Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *ArXiv e-prints*, pp. arXiv–1607, 2016.

[27] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

Anonymizing with Empathy: Voice Conversion That Preserves What Matters

Suhita Ghosh

Otto von Guericke University, Germany

Anonymising with Empathy: Voice Conversion That Preserves What Matters

Suhita Ghosh

Artificial Intelligence Lab, Otto-von-Guericke-University, Germany

suhita.ghosh@ovgu.de

Abstract

My doctoral research focuses on advancing emotion-preserving voice conversion (VC) for speech anonymisation, with particular emphasis on non-standard speech such as elderly and pathological utterances. Standard anonymisation pipelines often discard expressive or clinically relevant vocal traits, limiting their applicability in real-world domains like digital healthcare and affective computing. To address this, I propose novel VC frameworks and loss formulations that disentangle speaker identity, emotion, and pathology-related speech properties, enabling effective anonymisation even under noisy, low-resource, and non-parallel data conditions.

Index Terms: speech anonymisation, voice conversion, DDSP

1. Introduction

The increasing use of cloud-based speech devices, such as smart speakers and microphones, raises significant concerns regarding the protection and confidentiality of sensitive user data [1, 2]. In cases of data compromise, recorded speech can be exploited to bypass speaker verification systems or impersonate authorized users, posing severe privacy and security threats [3, 4]. Therefore, anonymising speech data before sharing it across systems has become imperative. Voice conversion (VC) achieves anonymisation by altering the speech characteristics of a source speaker to sound like a different target speaker while preserving linguistic content.

Preserving emotional content and domain-specific vocal traits during anonymisation is crucial, especially when digital assistants or healthcare monitoring systems respond based on emotional states or clinical vocal markers. Current VC approaches [5] either rely heavily on parallel datasets, which are expensive and impractical to obtain, or utilize non-parallel data, which is more representative of real-world conditions [6].

Existing VC approaches [5], while varied, present key limitations. Traditional VC methods [7] based on parallel data, such as parametric statistical modelling, exemplar-based techniques, and neural network approaches, are impractical due to their dependency on paired source-target recordings, which are costly and labour-intensive to obtain. Recent methods employing non-parallel data, including encoder-decoder models with phonetic posteriorgrams (PPGs) [8, 9] and variational autoencoders (VAEs) [10, 11], often produce conversions characterized by unnatural prosody and emotional flattening due to spectrum smoothing and alignment inaccuracies. GAN-based frameworks, notably StarGANv2-VC [12], have improved naturalness and intelligibility through adversarial training and cycle-consistency constraints. However, even these state-of-the-art models struggle significantly in preserving nuanced emotional expressions and critical prosodic variations, particularly in scenarios involving diverse emotions, highly variable pitch, and atypical speech patterns such as those found in elderly or pathological speech [13].

2. Research Questions

My research addresses the limitations of current VC methods that fail to preserve emotional nuance, prosody, and clinically relevant vocal traits. I aim to develop robust VC frameworks that anonymise speech while explicitly retaining emotional and domain-specific cues, even under noisy and limited data. A key focus is on designing models that disentangle speaker identity from paralinguistic features. I also develop evaluation protocols that ensure both linguistic and diagnostic fidelity, critical for applications in digital health and affective computing. The key research problems I address are:

- **RQ1:** How can speech be anonymised while preserving emotional expression?
- **RQ2:** How can perceptual quality and intelligibility be improved without increasing model complexity?
- **RQ3:** How can clinically relevant markers in pathological speech (e.g., jitter, shimmer) be retained after anonymisation?
- **RQ4:** How can effective anonymisation be achieved with limited and noisy data?
- **RQ5:** How can pathological regions (e.g., dysfluencies) be detected and segmented using weak labels for rigorous evaluation?

3. Methods

My research follows an incremental path, beginning with emotion-preserving voice conversion on healthy speech and progressing toward handling pathological data, motivated by practical needs and key gaps in voice anonymisation:

Emotion-Preserving Voice Conversion: To address RQ1, preserving emotional content in anonymised speech, I developed a semi-supervised GAN-based voice conversion framework, Emo-StarGAN [6] that combines direct and indirect emotion supervision. The **direct** pathway uses an adversarial emotion classifier trained to distinguish emotional states from converted speech, guiding the generator to retain emotional expressivity. **Indirect** supervision involves computing losses using handcrafted acoustic descriptors, such as ΔF0, loudness, and spectral centroid, along with deep emotion embeddings from a pre-trained encoder fine-tuned on an emotion conversion task. These losses capture the emotion-related discrepancy between the original and the converted samples. Crucially, the proposed losses are model-agnostic and improve emotion preservation across architectures, including VQ-VAE [14].

Perceptual Quality Enhancement: Despite the gains in emotion preservation achieved through the method addressing RQ1, the converted speech exhibited degraded perceptual quality. To mitigate this, I introduced perceptual loss functions [14] that operate on two complementary fronts: (1) optimal transport-based feature alignment (F0 and formants), which aligns intermediate activations from the generator with the original utterance, and (2) formant

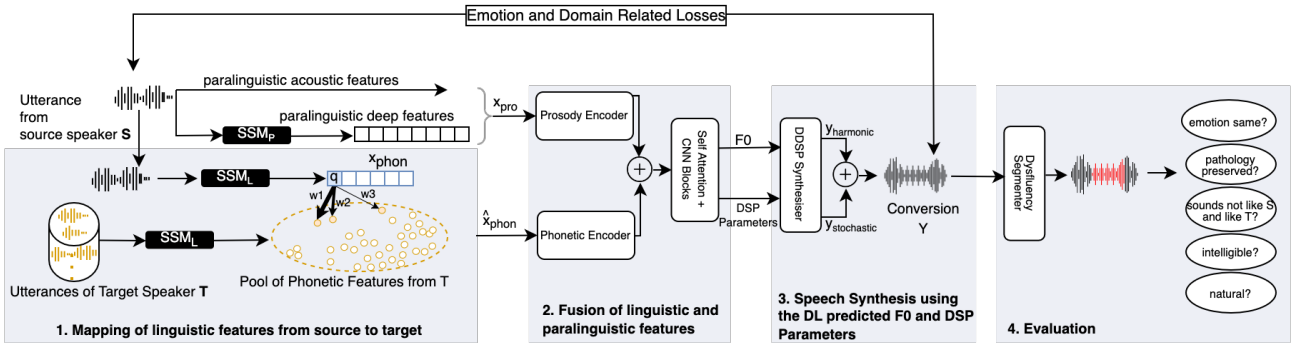


Figure 1: *Overview of the proposed anonymisation and evaluation pipeline.* $\mathrm{SSM}_{\langle\cdot\rangle}$ *represents a layer of the self-supervised model.*

trajectory loss, which explicitly constrains formant dynamics to preserve articulatory structure. These losses significantly reduced artifacts and improved speech intelligibility. Importantly, these enhancements were achieved without adding architectural complexity, thus addressing RQ2 effectively.

Domain-Adaptive Conversion with Limited Data: The GAN-based method developed for healthy emotional speech did not generalize well to pathological speech. This was primarily due to limited data availability, constrained by privacy and ethical considerations, and the presence of significant acoustic variability and noise in pathological speech. The GAN tended to overfit to these artefacts, degrading performance. This motivated the need for a more data-efficient framework capable of capturing relevant representations, such as emotion and phonetic content, while being robust to the non-linguistic noise commonly present in pathological speech.

To address this, I developed a more data-efficient framework, DDSP-QbE [13] based on differentiable digital signal processing (DDSP) [15], designed to extract relevant features, such as emotion and phonetic content, while remaining robust to spurious variability.

The architecture, illustrated in Figure 1, consists of three key components. The first stage requires no training: phonetic embeddings from the source speaker S are mapped to those of a target speaker T using query-by-example (QbE) matching. For every 25 ms frame, embeddings from the l^{th} layer of a self-supervised model (SSM), such as WavLM [16]), are compared to a pool of embeddings from T, and the n nearest neighbours are averaged to form converted phonetic representations. These representations, combined with prosodic features, are fed into a Conformer-based fusion module [17], which predicts DDSP parameters and F_0 required for synthesis [15]. In the third stage, a subtractive synthesiser generates harmonic and stochastic components that are combined to produce the final waveform.

To preserve emotional content and naturalness, I incorporate the emotion- and quality-related losses described earlier (RQ1 and RQ2). Additionally, to retain clinically relevant characteristics in pathological speech (RQ3), I introduce pathology-specific loss functions computed over features such as jitter and shimmer. These losses guide the model to maintain diagnostic markers during conversion.

The resulting system improves intelligibility, prosody, and the preservation of domain-specific traits while achieving effective anonymisation. A summary of objective evaluation results is provided in Table 1.

Fine-Grained Pathological Speech Segmentation: Given the critical importance of precisely evaluating domain-specific characteristics in anonymised speech, I developed a semi-supervised graph-based segmentation method catering to RQ5, StutterCut[1].

[1] accepted at Interspeech 2025

Table 1: *Evaluation Summary: A comparative analysis of the proposed method and a strong baseline across key metrics. Emotion Preservation and Pathology Preservation scores are obtained from expert and user ratings, indicating how well each model retains expressive and clinical vocal traits. Mean Opinion Score (MOS) reflects perceived naturalness on a 5-point scale (1 = robotic/unnatural, 5 = natural/human-like). Character Error Rate (CER) measures intelligibility, while Equal Error Rate (EER) assesses anonymisation effectiveness.*

Method	Pathology Pr. [%] ↑	Emotion Pr. [%] ↑	MOS ↑	CER [%] ↓	EER [%] ↑
Baseline	67.8	68.1	2.41	18.15	**50.16**
Proposed	**78.1**	**77.4**	**3.39**	**1.04**	48.98

This framework partitions speech into dysfluent and fluent regions using embedding similarities refined through classifier-guided uncertainty constraints. The method effectively utilises weak, utterance-level labels and demonstrates high temporal precision in localising dysfluencies. To rigorously benchmark my proposed VC approach, I introduced and publicly released a strongly labelled dataset specifically annotated at the frame-level for various dysfluency types (e.g., blocks, repetitions, prolongations).

These interconnected methodological steps culminate in a comprehensive voice conversion and analysis pipeline, shown in Fig. 1.

4. Conclusion and Future Work

I have developed an initial voice anonymisation framework capable of preserving emotional and diagnostic traits in both typical and pathological speech. Preliminary evaluations and a released pathology-labelled dataset demonstrate its potential for responsible use in clinical and accessibility contexts.

Moving forward, I aim to expand the framework to support additional languages and a broader range of speech disorders. Enhancing the dysfluency segmentation module with adaptive windowing and richer acoustic-prosodic features is also a priority. Additionally, I plan to integrate audio watermarking techniques to enable traceability of converted speech and prevent misuse, paving the way for safe, inclusive, and deployable anonymisation systems.

5. References

[1] C. Wienrich, C. Reitelbach, and A. Carolus, "The trustworthiness of voice assistants in the context of healthcare investigating the effect of perceived expertise on the trustworthiness of voice assistants, providers, data receivers, and automatic speech recognition," *Frontiers in Computer Science*, vol. 3, p. 685250, 2021.

[2] M. Haase, J. Krüger, and I. Siegert, "User perspective on anonymity in voice assistants," in *Proc. of the HCI International*, 2023, p. s.p.

[3] D. Kumar, R. Paccagnella, P. Murley, E. Hennenfent, J. Mason, A. Bates, and M. Bailey, "Skill squatting attacks on Amazon Alexa," in *27th USENIX Security Symposium*, Baltimore, USA, 2018, pp. 33–47.

[4] N. Zhang, X. Mi, X. Feng, X. Wang, Y. Tian, and F. Qian, "Dangerous skills: Understanding and mitigating security risks of voice-controlled third-party functions on virtual personal assistant systems," in *IEEE Symposium on Security and Privacy*, 2019, pp. 1381–1396.

[5] T. Walczyna and Z. Piotrowski, "Overview of voice conversion methods based on deep learning," *Applied Sciences*, vol. 13, no. 5, p. 3100, 2023.

[6] S. Ghosh, A. Das, Y. Sinha, I. Siegert, T. Polzehl, and S. Stober, "Emo-StarGAN: A Semi-Supervised Any-to-Many Non-Parallel Emotion-Preserving Voice Conversion," in *Proc. INTERSPEECH 2023*, 2023, pp. 2093–2097.

[7] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 132–157, 2020.

[8] Z. Lian, Z. Wen, X. Zhou, S. Pu, S. Zhang, and J. Tao, "ARVC: An auto-regressive voice conversion system without parallel training data." in *INTERSPEECH*, 2020, pp. 4706–4710.

[9] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *2016 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2016, pp. 1–6.

[10] D.-Y. Wu, Y.-H. Chen, and H.-y. Lee, "VQVC+: One-shot voice conversion by vector quantization and U-Net architecture," *Proc. Interspeech 2020*, pp. 4691–4695, 2020.

[11] H. Tang, X. Zhang, J. Wang, N. Cheng, and J. Xiao, "Avqvc: One-shot voice conversion by vector quantization with applying contrastive learning," in *Proc. of the IEEE ICASSP*, 2022, pp. 4613–4617.

[12] Y. A. Li, A. Zare, and N. Mesgarani, "StarGANv2-VC: A Diverse, Unsupervised, Non-Parallel Framework for Natural-Sounding Voice Conversion," in *Proc. Interspeech 2021*, 2021, pp. 1349–1353.

[13] S. Ghosh, M. Jouaiti, A. Das, Y. Sinha, T. Polzehl, I. Siegert, and S. Stober, "Anonymising elderly and pathological speech: Voice conversion using DDSP and Query-by-Example," in *Interspeech 2024*, 2024, pp. 4438–4442.

[14] S. Ghosh, F. Dreyer, T. Thiele, F. Lorbeer, and S. Stober, "Improving voice quality in speech anonymization with just perception-informed losses," in *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, 2024.

[15] D.-Y. Wu, W.-Y. Hsiao, F.-R. Yang, O. Friedman, W. Jackson, S. Bruzenak, Y.-W. Liu, and Y.-H. Yang, "DDSP-based singing vocoders: A new subtractive-based synthesizer and a comprehensive evaluation," in *ISMIR 2022*, 2022.

[16] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "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.

[17] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 5036–5040.

Using Speech Prosody to Guide Text Generation by Large Language Models

David Porteš

Masaryk University, Czech Republic

Using Speech Prosody to guide Text Generation by Large Language Models

David Porteš

Faculty of Informatics, Masaryk University, Brno, Czech Republic

xportes@fi.muni.cz

Abstract

The relationship between prosody and text plays a crucial role in the Text to Speech (TTS) task, where prosody is commonly predicted from text in order to improve the naturalness of synthesized speech. However, little if any work has been done on the inverse problem, that is, generating text that fits a given prosody. We believe that modeling this inverse relationship could lead to interesting and diverse applications, especially when coupled with the power of contemporary LLMs. Therefore, we propose a novel task, which we call Prosodical Guiding. The goal of the task is to steer the output of an LLM to fit the prosody of a given speech recording. As a first-shot architecture aimed at solving the task, we propose Meligner (Melodic Aligner), which works by rescoring each token during the LLM's decoding phase. Each token is rescored based on the token's compatibility with the specified prosody, as judged by a separate model trained on (prosody, transcription) pairs. Finally, we showcase various use cases of the Prosodical Guiding task, and describe our current progress.

Index Terms: prosody, VQ-VAE, Fundamental frequency, F0, Energy, embeddings, rescoring, LLM

1. Introduction

We propose a novel task, which we call Prosodical Guiding. The goal of Prosodical Guiding is to extend the prompt-based method of interaction with LLMs by allowing the user to upload a custom recording of speech, in addition to the textual prompt. The LLM would then generate text based on the prompt, that at the same time fits the prosody of the speech recording. This set up would allow the user to freely vary both the LLM prompt and the desired prosody, leading to a degree of control that is impossible with current text-based models. The use cases of such system are highly varied, ranging from creative writing aids to unintelligible speech decoding, or even emotional speech translation. They are described in greater detail in Section 3.

To enable these applications, we propose the Meligner architecture. Meligner works by rescoring each token in the LLM's decoding phase based on the token's compatibility with the specified prosody. As a judge of compatibility, we propose using a model trained to predict text from prosody. Since, to the extent of our knowledge, no such model has been published so far, this thesis aims to train a first-of-its-kind model that learns this relationship.

2. Methods

As a first-shot architecture aimed at solving the Prosodical Guiding task, we propose Meligner (Melodic Aligner). The Meligner architecture (See Figure 1.) works by employing a separate rescoring model; a Sequence-to-sequence model trained on (F0+Energy, transcription) pairs. The textual prompt is processed in the standard way by the LLM, while the F0 and Energy features are first jointly embedded into vectors, and then fed into the rescoring model. The rescoring model is then used during the decoding stage of the LLM, where it assigns each candidate token a score representing the 'goodness of fit' between the token and the prosody. This score is merged with the LLM's next token probabilities, steering the decoding process towards prosody-compatible tokens.

2.1. Embeddings generation

To convert the F0 and Energy trajectories into vector embeddings, we make use of a Vector Quantized Variational Autoencoder (VQ-VAE) [1]. The VQ-VAE consists of an encoder, a bottleneck layer containing a learnable codebook, and a decoder. To obtain our vector embeddings, we feed the trajectories into the encoder and extract the latent vectors from the bottleneck layer.

2.2. Rescoring model

To teach our rescoring model the correspondence between prosody and text, we cast the problem into a translation task. We consider the F0+Energy embeddings to be a sort of language, and we try to translate this 'prosodic embedding language' into the transcription of each recording. While any seq2seq model can be used, we use the standard Transformer architecture conforming to [2]. We train our models on the LibriTTS [3] dataset.

2.3. Evaluation

As a performance benchmark for development, we make use of the recording transcription, since it trivially fits the prosody of the recording. We propose a scheme whereby we let the LLM generate text freely with a generic prompt, such as "Generate an English sentence.", and try to steer its output toward the transcription of the provided speech recording, using only the prosodical features of the speech as input. However, since we do not expect any model to be able to recover the full transcription, due to the problem being one-to-many in nature, we instead measure the rank of each word from the transcription at each decoding stage. The full system with non-generic prompts will be evaluated using human evaluation.

3. Use cases

Preserving the standard textual prompt of the LLM grants the Prosodical Guiding task broad applicability. Any prompt that could benefit from the ability to specify a desired prosody can be considered a use case of Prosodical Guiding. In this Section,

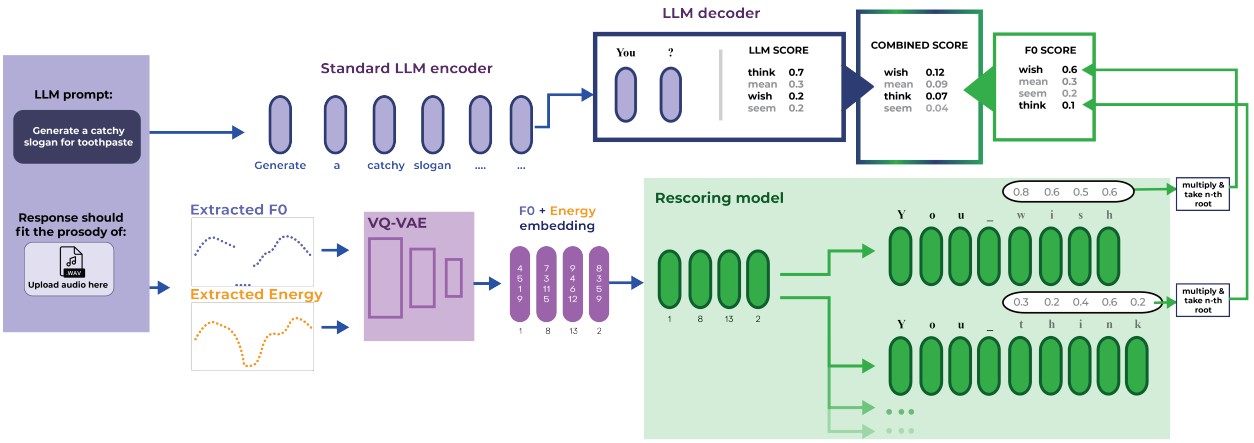


Figure 1: *Illustration of the Meligner architecture.*

we propose just a few of the possible use cases.

3.1. Creative writing aid

Being able to guide text generation by prosody could allow authors to take existing dialogues and write alternative ones with identical emotional content. One could take, for example, a dialogue about topic A and turn it into a dialogue about topic B, while exactly preserving the emotional content of the original.

Additionally, when writing advertisement slogans or similar, Prosodical Guiding could make it possible to automatically generate a fitting text to a custom, pre-crafted prosody. This could allow copywriters to focus their efforts on crafting the ideal, catchy melody for their slogans, leaving it to the Prosodical Guiding model to fill in the words after the fact. One could even imagine writers crafting 'prosodical templates' optimized for maximum emotional impact, which then get filled with content as needed. This could also allow, for example, rap artists, to mumble incoherent speech with their desired prosody, and have it automatically converted into lyrics on the topic of their choosing.

3.2. Expressive speech translation or dubbing

The Prosodical Guiding framework could help human translators translate expressive speech, such as when dubbing movies into different languages. In contrast to the current state of the machine translation field, which is mainly focused on matching the human level of translation, Prosodical Guiding could instead help humans discover better, more creative translations that are unburdened by the literal meaning of the original sentence.

The user would input translated text leading up to the sentence to be translated into the LLM prompt, and ask the LLM to generate a continuation of the text. As audio input, the user would upload the recording of the sentence to be translated.

Our assumption is that by limiting the space of possible continuations to those that naturally flow from the previously translated text (LLM), and at the same time have the same emotional content as the original sentence (prosody), we can carve out from the vast space of possible continuations a small space where many good translations "live". The user can then navigate this limited space by making adjustments to the LLM prompt, and discover new and original translations in the process.

3.3. Unintelligible speech decoding

When conducting Automatic Speech Recognition under conditions where phonemes are unintelligible, i.e. muffled speech, our approach could lead to generating text that, while not exactly accurate, preserves the intent of the speaker. Coupled with the context provided by the user in the prompt, this approach could help to iteratively decode unintelligible speech.

4. Current state of work

We have conducted a study on generating high-quality F0 embeddings, as well as joint F0 and Energy embeddings, to be used in the rescoring model. Unfortunately, the prosody-text relationship is not a very clear one, and we are struggling to train our rescoring model to learn any meaningful relationships. To help the rescoring model to train, we plan to use much larger datasets (YODAS [4]), as well as experiment with word-level joint F0+Energy embeddings, where we would try to predict individual words at a time, as opposed to whole sentences. In the future, we also plan to experiment with replacing the LLM with the novel language diffusion models [5], which we find to be better suited for the task, due to their non-autoregressive nature.

5. Conclusion

We have proposed Prosodical Guiding, a novel task that involves using prosody of speech, provided as an additional audio input, to guide the text generation process of Large Language Models (LLMs). Since the task is designed to preserve the standard textual prompt of the LLM, its applications are as varied as the number of prompts one can come up with. We have also proposed a rescoring-based architecture, Meligner, as a first-shot attempt at solving the Prosodical Guiding task. Finally, we have showcased various use cases of the proposed task.

6. References

References

[1] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., Dec. 2017, pp. 6309–6318.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[3] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," in *Interspeech 2019*. ISCA, Sep. 2019, pp. 1526–1530.

[4] X. Li, S. Takamichi, T. Saeki, W. Chen, S. Shiota, and S. Watanabe, "Yodas: Youtube-oriented dataset for audio and speech," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.

[5] S. Nie, F. Zhu, Z. You, X. Zhang, J. Ou, J. Hu, J. Zhou, Y. Lin, J.-R. Wen, and C. Li, "Large language diffusion models," *arXiv preprint arXiv:2502.09992*, 2025.

Phonetic Perception and Speech Enhancement for Mandarin Listeners of English in Adverse Listening Conditions

Yunqi C. Zhang

University of Auckland, New Zealand

Phonetic Perception and Speech Enhancement for Mandarin Listeners of English in Adverse Listening Conditions

Yunqi C. Zhang[1]

[1]Mechanical and Mechatronics Engineering, Univeresity of Auckland, New Zealand

clara.zhang@auckland.ac.nz

Abstract

Speech enhancement is a widely explored field and contributes to improving communication efficiency. However, most of the speech enhancement algorithms were designed and tested on native listeners, despite the fact that non-native listeners are facing more challenges when listening in noisy conditions. This paper summarises my PhD research on exploring a speech enhancement method for non-English speakers, specifically, Mandarin listeners, to improve their intelligibility when listening in noise. This research first explored the effect of existing speech enhancement methods on Mandarin listeners, and analysed the results from a phonetic aspect, then designed an algorithm specifically for Mandarin listeners.

Index Terms: non-native perception, speech enhancement, phonetics

1. Introduction and Research Questions

Understanding speech in noisy environment is a fundamental challenge in both everyday communication and speech technology development. Listeners' ability to comprehend speech under such conditions can be influenced by multiple factors, including the type and level of background noise, the listener's first language, and the language being spoken [1, 2]. Numerous studies have demonstrated that non-native (L2) listeners generally experience greater difficulty in noise compared to native (L1) listeners [3]. This is partly because L2 listeners are often less sensitive to the acoustic-phonetic cues and less experienced in utilising such cues in the presence of noise [4].

Despite the development of numerous speech enhancement (SE) algorithms aimed at improving intelligibility in noise, most have been designed and evaluated with L1 users only [5, 6]. As a result, little is known about the effectiveness of these algorithms for L2 users, who may employ different perceptual strategies depending on their linguistic background. For example, speakers of tonal languages such as Mandarin may process speech differently from speakers of non-tonal languages like English [7]. Furthermore, the age of immersion in the target language also plays an important role in L2 speech perception capabilities. Earlier immersion typically results in superior speech intelligibility, particularly under challenging listening conditions [8, 9]. These factors collectively highlight the need to explore the development of SE algorithms specifically for L2 listeners.

This PhD research addresses this gap by focusing on Mandarin-speaking L2 listeners' perception of New Zealand English (NZE) in noise. Four principal research questions (RQ) guide this investigation:

1. How effective are existing single-channel SE algorithms at improving the intelligibility of NZE for Mandarin listeners with different levels of exposure to NZE?
2. How do Mandarin listeners perceive noisy NZE speech at the phonetic level?
3. Can linguistic features from Mandarin be used to help Mandarin listeners better understand NZE?
4. How to design a SE algorithm that is specifically tailored for Mandarin listeners by utilising the phonetic knowledge?

2. Results So Far

This section introduces the investigations undertaken to address the research questions. The study began with a subjective listening test evaluating the effectiveness of existing SE algorithms in improving the intelligibility of NZE for Mandarin listeners. Based on participants' responses, a subsequent phonetic analysis was conducted to identify systematic patterns of phonetic confusion. According to the main vowel errors made by the Mandarin listeners, a pitch modification on vowels was implemented to investigate whether Mandarin listeners can utilise pitch, which is an important factor in intelligibility in their native language, to improve their perception of NZE as well. The findings and potential underlying factors are briefly discussed.

2.1. Research Question 1: SE Algorithm Effectiveness

To address RQ1, a subjective listening test was conducted, evaluating five widely recognised single-channel SE algorithms representative of different stages in SE technology development: *A priori* Wiener filter (WF)[10], generalised subspace (SS)[11], unsupervised Bayesian non-negative matrix factorisation (NMF)[12], convolutional time-domain audio separation network, (Conv)[13], and deep complex U-net (Unet)[14]. In addition, the intelligibility of unprocessed noisy speech was included as a baseline condition for comparison.

Three groups of participants were recruited, categorised by their age of immersion in NZE:

- Early-immersed (NZE) group: 20 NZE listeners who had arrived in New Zealand before the age of 12 years old.
- Late-immersed (NZM) group: 19 Mandarin listeners who had been immersed in New Zealand English for more than one year after the age of 15.
- Non-immersed (CM) group: 19 Mandarin listeners residing in Mainland China who had never been immersed in English-speaking environments.

Participants listened to NZE speech from the SPANZ BKB corpus [15], contaminated by multi-talker babble [16] or speech-shaped noise (SSN) at varying SNRs (babble: 0, -3, -6 dB; SSN: -3, -6, -9 dB).

Figure 1 illustrates the linear prediction of the proportion

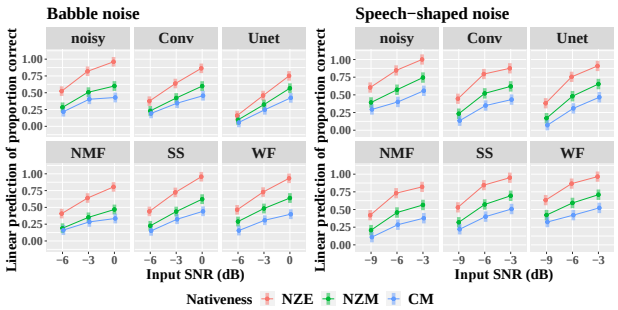


Figure 1: *Predicted proportion correct from linear mixed effect model under different noise types in terms of immersion condition separated by algorithms. The error bars represent the 95% confidence intervals.*

correct of the participant groups separated by SE algorithms. The results showed that the tested SE algorithms failed to improve and rather degraded the speech intelligibility of all participants in most cases. As expected, the NZE group always yields significant higher intelligibility than the Mandarin groups. The data also revealed that the intelligibility decreases as the age of immersion increases, with the NZM group significantly outperforming the CM group at higher SNR levels. However, this advantage disappeared at lower SNR levels (-3 dB for babble noise and -9 dB for SSN). The findings suggest two key issues: First, current SE algorithms do not improve the speech intelligibility for listeners under negative SNR conditions. Second, they fail to narrow the perception gap between L1 and L2 listeners.

2.2. Research Question 2: Phonetic-Level Analysis

Based on the participant responses from Section 2.1, a detailed phonetic analysis was conducted to explore error patterns. Unlike most SE studies, that assess intelligibility solely on word-level correctness, this study investigated phoneme-level confusions. Errors were analysed by comparing the phonetic components of the stimulus keywords and the participants' responses. Key findings include:

- Consonant errors were dominated by dropped codas or onsets, likely due to noise masking where these phonemes were treated as noise and suppressed by SE algorithms.
- The NZE group exhibited fewer phonetic errors compared to the Mandarin groups for both vowels and consonants.
- Both Mandarin groups exhibited vowel confusions attributable to their unfamiliarity with NZE vowels, including the /e/-/ɪ/, /æ/-/e/, and /ɜː/-/uː/ confusions [17, 18].
- NZM participants showed fewer vowel confusions than CM participants, presumably due to immersion effects.

2.3. Research Question 3: Fundamental Frequency Manipulation

Building on the phonetic findings, RQ3 explores phonetic-level perception improvement strategies. Given that consonants were primarily masked by noise and treated as artifacts by SE algorithms, the focus shifted to vowel enhancement. The vowels /ae/, /e/, /ɪ/, and /i/ were selected as targets due to their frequent confusion across all listener groups and their association with NZE characteristics.

Considering that Mandarin is a tonal language and native Mandarin speakers are sensitive to pitch change, while pitch, or fundamental frequency (F0) is also an important factor in speech perception, the research investigated whether pitch modification could enhance Mandarin listeners' vowel identification in noise. Since limited literature has investigated the effect of F0 on Mandarin listeners' perception in English, a subjective test was conducted by modifying the F0 of /ae/ and /e/ with a high-fidelity speech synthesiser, STRAIGHT [19], and added with babble noise at different SNR levels (at 0 and -6 dB SNRs and no noise). Both native NZE listeners and native Mandarin (NM) listeners participated in vowel identification tasks involving /ae/, /e/, /ɪ/, and /i/ vowels in hVd format.

The result showed no significant change in vowel identification accuracy for the NZE group by changing the F0. However, the NM group had a trend of accuracy improvement when the F0 is lowered for /e/ by reducing the /e-i/ confusion when noise is absent. Such a trend is statistically significant when the F0 is lowered by 4 semitones. This improvement was potentially attributable to enhanced intrinsic F0 differences between vowels, allowing Mandarin listeners to better utilise vowel duration cues for identification. The /ae-e/ confusion pair showed no improvement, likely due to similar intrinsic F0 and vowel duration characteristics of the vowels. No significant trend was found for noisy cases, likely due to the masking effect of the noise.

3. Current Work and Challenges

The research has progressed to its final stage: designing an SE algorithm that incorporates Mandarin listeners' perceptual characteristics to make NZE speech more intelligible in noise for them. Based on previous findings, the current approach involves F0 adjustment to exaggerate intrinsic F0 differences.

However, identifying individual phonemes and determining their intrinsic F0 ranges in noisy sentences at negative SNR levels presents significant challenges. This preliminary algorithm estimates the F0 of the whole speech and calculates the average. Values above this average value will be raised and those below will be reduced, thereby expanding the overall F0 range. Pitch shifting is implemented by STRAIGHT used in Section 2.3.The designed algorithm functions as a post-processing stage for any SE algorithm output. However, several technical challenges occurred during algorithm development:

- Reliable F0 estimation from the noisy speech
- Quality of speech resynthesis by STRAIGHT based on the noisy speech
- Unnatural artifact sound caused by dramatic F0 changes, which leads to intelligibility degradation.

To address the F0 estimation challenge, this preliminary design assumes access to original F0 information (values and temporal indices). For the artifact issue, a sigmoid function is applied to smooth and naturalise the speech. The resynthesis quality challenge remains unresolved and requires further investigation.

4. Conclusion

This PhD project investigates the possibility of developing a SE method tailored to Mandarin L2 listeners of English. By systematically analysing the performance of current SE algorithms, phonetic confusion patterns, and the role of pitch, this research contributes to a deeper understanding of L2-specific perceptual challenges and tried to propose a novel, perceptually motivated enhancement method. Findings emphasise the limitations of current SE approaches for L2 listeners and offer directions for more inclusive and linguistically informed technologies.

5. Acknowledgements

The Interspeech 2025 organisers would like to thank ISCA and the organising committees of past Interspeech conferences for their help and for kindly providing the previous version of this template.

6. References

[1] N. R. French and J. C. Steinberg, "Factors Governing the Intelligibility of Speech Sounds," *The Journal of the Acoustical Society of America*, vol. 19, no. 1, pp. 90–119, Jan. 1947, publisher: Acoustical Society of America. [Online]. Available: http://asa.scitation.org/doi/abs/10.1121/1.1916407

[2] J. Mejia, H. Dillon, R. V. Hoesel, E. Beach, H. Glyde, I. Yeend, T. Beechey, M. Mclelland, A. O'Brien, J. Buchholz, M. Sharma, J. Valderrama, and W. Williams, "Loss of speech perception in noise – causes and compensation," *Proceedings of the International Symposium on Auditory and Audiological Research*, vol. 5, pp. 205–216, Dec. 2015.

[3] g.-i. family=LIU, given=CHANG and S.-H. JIN, "Intelligibility of American English vowels of native and non-native speakers in quiet and speech-shaped noise," *Bilingualism*, vol. 16, no. 1, pp. 206–218, 2013.

[4] O. Scharenborg and M. van Os, "Why listening in background noise is harder in a non-native language than in a native language: A review," *Speech Communication*, vol. 108, pp. 53–64, 2019.

[5] M. A. Abd El-Fattah, M. I. Dessouky, A. M. Abbas, S. M. Diab, E.-S. M. El-Rabaie, W. Al-Nuaimy, S. A. Alshebeili, and F. E. Abd El-samie, "Speech enhancement with an adaptive Wiener filter," *International Journal of Speech Technology*, vol. 17, no. 1, pp. 53–64, 2014.

[6] H. Wang and D. Wang, "Cross-Domain Diffusion Based Speech Enhancement for Very Noisy Speech," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[7] J. Yang and R. A. Fox, "Perception of English Vowels by Bilingual Chinese–English and Corresponding Monolingual Listeners," *Language and Speech*, 2013.

[8] L. H. Mayo, M. Florentine, and S. Buus, "Age of Second-Language Acquisition and Perception of Speech in Noise," *Journal of Speech, Language, and Hearing Research*, vol. 40, no. 3, pp. 686–693, 1997.

[9] L. Mi, S. Tao, W. Wang, Q. Dong, S.-H. Jin, and C. Liu, "English vowel identification in long-term speech-shaped noise and multi-talker babble for English and Chinese listeners," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. EL391–EL397, 2013.

[10] P. Scalart and J. Filho, "Speech enhancement based on a priori signal to noise estimation," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 2, May 1996, pp. 629–632 vol. 2, iSSN: 1520-6149.

[11] Y. Hu and P. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 4, pp. 334–341, Jul. 2003, conference Name: IEEE Transactions on Speech and Audio Processing.

[12] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and Unsupervised Speech Enhancement Using Nonnegative Matrix Factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, 2013.

[13] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[14] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex U-Net," *ICLR*, p. 20, 2019.

[15] J.-H. Kim and S. Purdy, "Speech Perception Assessments New Zealand (SPANZ)," *New Zealand Audiological Society Bulletin*, vol. 24, pp. 9–16, 2014.

[16] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[17] C. I. Watson, M. Maclagan, and J. Harrington, "Acoustic evidence for vowel change in New Zealand English," *Language Variation and Change*, vol. 12, no. 1, pp. 51–68, 2000.

[18] M. Maclagan, C. I. Watson, R. Harlow, J. King, and P. Keegan, "Investigating the Sound Change in the New Zealand English Nurse Vowel /ɜː/," *Australian Journal of Linguistics*, vol. 37, no. 4, pp. 465–485, 2017.

[19] K. Hideki. HidekiKawahara/legacy_STRAIGHT: A vocoder framework which had been widely used in research community since 1999. [Online]. Available: https://github.com/HidekiKawahara/legacy_STRAIGHT

Quantifying Anatomical Variation in the Vocal Tract in Typical and Atypical Speakers Using 3D MRI

Emily Kiff

University of York, United Kingdom

Quantifying anatomical variation in the vocal tract in typical and atypical speakers using 3D MRI

Emily Kiff

University of York, UK
emily.kiff@york.ac.uk

Abstract

Three-dimentional volumetric magnetic resonance imaging (3D MRI) can be utilised to gain insight into individual anatomical vocal tract variation and between speaker variation. There is a lack of understanding of anatomical variation between populations of typical speakers and no understanding of atypical anatomical variation. With advances such as 3D MRI technology, we are able to apply geometric morphometric methods to data that goes beyond simplistic estimations of vocal tract shape, length and acoustic output amongst speakers. Understanding anatomical vocal tract variation and formally linking anatomy to acoustic measures is thus vital for forensic speech science, speaker identification and speech technologies and to ensure atypical speakers are not disadvantaged by current, modern-day technologies.

Index Terms: volumetric magnetic resonance imaging (MRI), vocal tract anatomy, acoustic output, individual variation, atypical speakers, geometric morphometrics

1. Research questions and motivation

Acoustics of speech are fundamentally linked to the overall shape of the vocal tract, this is determined by both the structural shape of the vocal tract and by controlled articulatory movements [1]. With empirical studies representing the vocal tract as simplistic tubes [2], and others treating both the source and filter as separate entities [3], there is still much more to investigate when it comes to understanding how acoustics of speech are generated by the complex, three-dimensional vocal tract structure [4]. Whilst much work has previously focussed on two-dimensional real-time MRI - which does provide a rich source of information about vocal tract shaping and articulatory movements [5] - to truly understand the vocal tract anatomy, we need to observe individual anatomical differences using 3D MRI, since the vocal tract in itself is a 3D structure. By understanding anatomical variation across populations of typical speakers and mapping these anatomical structures to acoustic cues, it provides us with a foundation to explore atypical oral anatomies of individuals with speech disorders such as apraxia of speech (AOS). AOS is a speech disorder that is estimated to affect up to 4.4 per 100,000 people [6]. Speakers with AOS have difficulties with motor planning and programming of speech movements, resulting in sound distortions, segmental inaccuracy and vowel errors [7]. With ever-emerging advances in speech technologies, struggles that atypical speakers may face are unknown and overlooked. Typical speakers can behave poorly in automatic speaker recognition (ASR) systems [8], so we can infer that atypical speakers will be even more problematic for forensic speaker identification. Consequently, atypical speakers need to be better understood for forensic casework, due to such subjects being part of the general population as a whole.

With this in mind, the project aims to address the following research questions:

RQ1: How the vocal tract shape varies within the population of typical speakers?

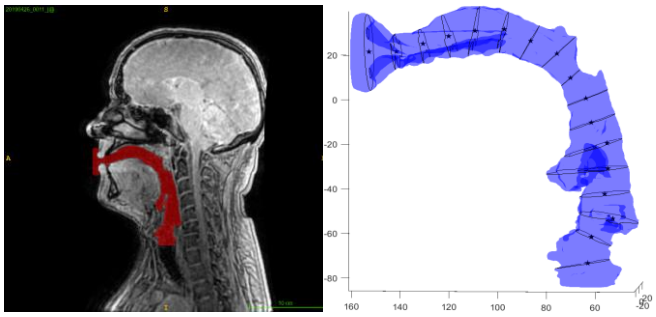
RQ2: How does vocal tract shape differ in specific populations of atypical speakers?

RQ3: What are the implications of these findings for forensic speaker identification?

2. Data and method

The data obtained is 3D volumetric MRI scans of adult vocal tracts articulating different sustained vowel qualities with synchronous audio. The previously collected data comprises approximately 100 mixed-sex subjects, with the vast majority being obtained from the USC 75-Speaker Speech MRI Database and other public datasets [4, 9]. Other data of typical speakers includes unpublished MRI collected by researchers at the University of York. Exploration into atypical vocal tracts will involve approximately 66 adult speakers with AOS. This database of atypical speakers is currently being collected for a wider project by researchers at Macquarie University, Sydney.

To quantify the vocal tract shape using MRI, two methods are utilised. The first method, which has already been undertaken for this project, details the vocal tract anatomy and length using a segmentation process in ITK-SNAP [10]. The vocal tract airway was manually segmented from the lips to the larynx, and an iterative bisection algorithm used to calculate the vocal tract length for each MRI observation, as displayed in Figure 1. An interpolated, smooth midline that represents airflow was then calculated from the anterior extent of the lips to the midline of the glottis for each of the segmentations.



(a) vocal tract segmentation *(b) segmentation midpoints*

Figure 1: *Vocal tract airway segmentation of the neutral schwa vowel in ITK-SNAP (a), and with the addition of midpoints in which the vocal tract measurement is obtained (b)*

The second methodological approach, which this project will soon explore, is based on geometric morphometrics (GMM). This is a method that is well established in other fields such as biology but has only recently been used to investigate vocal tract shape [11]. GMM uses landmarks to quantify and compare anatomical shapes. GMM is a much more suitable method to apply to larger datasets due to the segmentation process in ITK-SNAP being a time-consuming and lengthy process.

For analysis of any audio, acoustic measurements are obtained using Praat [12]. F1-F4 measurements are extracted from synchronous, denoised audio captured in the MRI scanner, or from synchronous audio captured in an anechoic chamber under 'MRI like' conditions [13]. So far, two methods of extracting formant measurements have been exploited. Firstly, manual midpoint measurements have been taken for F1-F4. Secondly, a Praat script that extracts formant information for each vowel has been run on each token. Settings are manually adjusted for each vowel and each vowel requires an estimated inputted number for each formant value. The manual measurements previously obtained are rounded to the nearest 100 Hz and inputted into the Praat script as a rough guide as to where the expected formants will be. Measurements from literature were also used as an estimate to ensure accuracy, but this had very little effect on the output. Future work will involve the exploration of the Praat plugin FastTrack [14] to extract formant measurements. Using three formant measurement techniques will aid in finding the most accurate, reproducible and reliable method.

3. Results so far

Results that have been obtained so far are vocal tract length and acoustic measurements from two British English professional phoneticians (one male, one female). Subjects articulated six, held, isolated vowels [i, a, ɑ, ɔ, u, ə] in three conditions: neutral, lowered larynx and with additional lip rounding. These subjects form part of the population of typical speakers. By investigating vocal tract length in speakers in differing supralaryngeal settings and mapping these to acoustic resonances, it aids in the investigation of vocal tract shape and length variation amongst individuals. We can assume that rounding the lips and lowering the larynx will increase vocal tract length, which consequently lowers the resonant frequencies [2]. Despite this, to what extent the resonant frequencies decrease in each condition is yet to be explored. Average results are displayed in Table 1, demonstrating the average vocal tract length and acoustic output per subject and condition. By using 3D MRI, we can map these lengths – and other dimensions – directly to resonant frequencies in the speech signal.

Table 1: *Average vocal tract length (VTL) and manual formant measurements*

Subject	Condition	VTL (cm)	F1 (Hz)	F2 (Hz)	F3 (Hz)	F4 (Hz)
R5252 (male)	Neutral	19.02	550	1283	2522	3167
	Lip Rounding	20.69	488	1084	2357	3101
	Lowered Larynx	20.05	467	1089	2423	3071
R5338 (female)	Neutral	15.76	659	1628	2993	4196
	Lip Rounding	16.30	544	1314	2867	3984
	Lowered Larynx	16.42	571	1315	2916	4044

As expected, the average results across the six vowels for each condition display an increase in vocal tract length in the lengthening conditions. Oppositely, the formant measurements display a decrease in the lengthening conditions. Differences can be observed in Table 2. The percentage change indicates the change from the neutral condition to the lengthening condition for each subject.

Table 2*: Average percentage change from neutral condition to lengthening conditions*

Subject	Condition	VTL (%)	F1 (%)	F2 (%)	F3 (%)	F4 (%)
R5252 (male)	Lip Rounding	+9	-11	-16	-7	-2
	Lowered Larynx	+5	-15	-15	-4	-3
R5338 (female)	Lip Rounding	+3	-17	-19	-4	-5
	Lowered larynx	+4	-13	-19	-3	-4

4. Future plans

Immediate future work will entail using FastTrack as another method to obtain formant measurements to ensure accuracy and reliability. Future work will then involve applying GMM to the published and unpublished 3D MRI datasets of typical speakers. Upon exhausting the exploration of vocal tract variation in the population of typical speakers alongside the corresponding acoustic signal, atypical speakers with AOS will then be explored using the same method. Once the methods have been applied, variation between and within both populations can be investigated.

5. Challenges

With MRI data being expensive and time-consuming to obtain and collect, this project needs to exploit already published and unpublished datasets. Whilst the vast majority of data is consistent and comprises of overlapping vowel qualities in English, the data as a whole varies in which vowel qualities are used and consists of mostly British English, American English and Australian English speakers. Ideally, all data would consist of speakers with the same accent and include all the same vowel qualities. Despite this, due to the expense of MRI data collection, we can only utilise what data is accessible to us currently.

6. References

[1] Lammert, A., Proctor, M., & Narayanan, S. (2013). Interspeaker Variability in Hard Palate Morphology and Vowel Production. *Journal of Speech, Language, and Hearing Research*, *56*(6), 1924–1933.

[2] Stevens, K. N. (1998). *Acoustic phonetics.* Boston, MA: MIT Press.

[3] Fant, G. (1970). *Acoustic Theory of Speech Production: With Calculations Based on X-Ray Studies of Russian Articulations*. Mouton.

[4] Birkholz, P., Kürbis, S., Stone, S., Häsner, P., Blandin, R., & Fleischer, M. (2020). Printable 3D vocal tract shapes from MRI data and their acoustic

and aerodynamic properties. *Scientific Data*, *7*(1), 255.

[5] Yue, Y., Proctor, M., Zhou, L., Gupta, R., Piyadasa, T., Gully, A., Ballard, K., & Jin, C. (2024). Towards Speech Classification from Acoustic and Vocal Tract data in Real-time MRI. *Interspeech 2024*, 1345–1349.

[6] Duffy, J. R., Utianski, R. L., & Josephs, K. A. (2021). Primary progressive apraxia of speech: From recognition to diagnosis and care. *Aphasiology*, *35*(4), 560–591.

[7] Allison, K. M., Cordella, C., Iuzzini-Seigel, J., & Green, J. R. (2020). Differential diagnosis of apraxia of speech in children and adults: A scoping review. *Journal of Speech, Language, and Hearing Research: JSLHR*, *63*(9), 2952–2994.

[8] Hughes, V., Wormald, J., Foulkes, P., Harrison, P., Kelly, F., Vloed, D. van der, Welch, P., & Xu, C. (2023). Automatic speaker recognition with variation across vocal conditions: a controlled experiment with implications for forensics. *Interspeech 2023*, 591–595.

[9] Lim, Y., Toutios, A., Bliesener, Y., Tian, Y., Lingala, S. G., Vaz, C., Sorensen, T., Oh, M., Harper, S., Chen, W., Lee, Y., Töger, J., Monteserin, M. L., Smith, C., Godinez, B., Goldstein, L., Byrd, D., Nayak, K. S., & Narayanan, S. S. (2021). A multispeaker dataset of raw and reconstructed speech production real-time MRI video and 3D volumetric images. *Scientific Data*, *8*(1), 187.

[10] Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C., & Gerig, G. (2006). User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *NeuroImage*, *31*(3), 1116–1128.

[11] Gully, A. J. (2021). Quantifying vocal tract shape variation and its acoustic impact: A geometric morphometric approach. *Interspeech 2021*.

[12] Boersma, P., & Weenink, D. (2025). *Praat: doing Phonetics by Computer (version 6.4.31) [Computer Program]*. https://www.fon.hum.uva.nl/praat/

[13] Gully, A. J., Foulkes, P., French, P., Harrison, P., & Hughes, V. (2019). The Lombard effect in MRI noise. *Proc. of the 19th International Congress of Phonetic Sciences*, 800–804.

[14] Barreda, S. (2021). Fast Track: fast (nearly) automatic formant-tracking using Praat. *Linguistics Vanguard: Multimodal Online Journal*, *7*(1). https://doi.org/10.1515/lingvan-2020-0051

Towards Automatic Voice Disorders Screening: Robustness and Generalisation of Speech-Based Methods in German-Language

Monica Gonzalez-Machorro

audEERING GmbH / Technical University of Munich, Germany

Towards Automatic Voice Disorders Screening: Robustness and Generalisation of Speech-Based Methods in German-Language

Monica Gonzalez-Machorro[1,2]

[1]CHI – Chair of Health Informatics, Technical University of Munich, Germany
[2]audEERING GmbH, Germany,

monica.gonzalez@tum.de

Abstract

This on-going doctoral thesis aims to advance voice disorders screening in the context of German-language. To do so, it focuses on three key areas: 1) collecting three datasets with a total of 175 patients for various voice disorders (organic and functional disorders), 2) evaluating acoustic features robustness, 3) quantifying speech symptoms such as breathiness and roughness obtained from the GRBAS scale, a widely used scale for assessing voice impairment. This technology could potentially assist otolaryngologists to automatically screen for specific disorder manifestations during initial patient consultations.
Index Terms: voice disorders, speech technology, German language, speech symptoms, feature robustness

1. Introduction

Voice disorders have a negative impact on the quality of life, as the voice is essential for human communication and social interaction [1]. These disorders are classified into two categories based on their underlying condition: organic and functional. Organic voice disorders correspond to actual pathological changes in the larynx or vocal folds [2], such as vocal cords atrophy, chronic laryngitis, spasmodic dysphonia, or laryngeal cancer, which leads to structural changes in the larynx. These disorders may also stem from neurological conditions [3]. Examples are Multiple Sclerosis (MS), Parkinson's Disease (PD), and Amyothrophic Lateral Sclerosis (ALS). Notably, voice changes are one of the first symptoms to develop in neurological conditions [4], making otolaryngologists the first clinicians to encounter patients with these disorders. Functional voice disorders arise from the misuse of the larynx without the presence of any structural pathology [5]. Examples are muscle tension dysphonia or vocal fatigue.

1.1. Motivation

In recent years, there has been an increasing interest in the potential of speech-based machine learning (ML) to aid clinicians, such as otolaryngologists, in screening for voice disorders [6, 7]. There have been numerous works exploring ML systems to distinguish voice disorders–whether organic or functional–from healthy speech, with promising results [8]. For instance, some report accuracies up to 90% for detecting PD [9] or around 80% for Alzheimer's Disease (AD) [10], approximately 80% for predicting chronic laryngitis [11], and around 70% to tell apart dysphonia from a healthy control speech [12].

However, most studies frame this as a binary classification task, distinguishing a specific voice disorder from a healthy control group. Fewer studies have addressed this task as three-class or multi-class classification tasks, for example, aiming to identify between Mild Cognitive Impairment (MCI), AD and healthy speech [13] or analysing multiple diseases, as attempted with the Saarbrücken Voice Database, which is a large-scale speech dataset, containing data from over 2000 speakers and 71 different pathologies [14]. These multi-class studies have shown less conclusive results as the ones focused on binary classification tasks [15]. Additionally, many acoustic features used in voice disorder detection, such as eGeMAPS [16], were originally developed for other purposes and lack sufficient repeatability. Their robustness across diverse recording conditions remains unclear, which is a critical measurement impacting reliability of results [17]. Consequently, despite recent advancements, these approaches have not yet been translated into practical tools that otolaryngologists can utilise during initial patient consultations to identify voice disorder manifestations [8].

To contribute in the development of these methods towards a real-life application, we identify three key challenges: **1) Data:** Existing datasets are limited in both the number of speakers and the range of disorders represented. Cross-dataset comparisons are challenging due to variations in recording conditions and speech tasks. Standardising recording protocols is critical for obtaining larger sample sizes and more disorders, enabling more robust applications. **2) Features:** Current features used in voice disorders detection were primarily developed for tasks like speech emotion recognition, not specifically for voice disorders. Their robustness need thorough evaluation to ensure reliability in clinical applications. **3) Generalisation:** Most modelling approaches focus on diagnosing specific disorders, mostly through binary classification (disorder versus healthy), overlooking multi-morbidity and the limitations of using speech alone for accurate differential diagnosis. We propose shifting toward symptom monitoring, such as predicting each component from the GRBAS scale, to enhance the effectiveness and generalisation of the modelling approaches. This approach would also save time required to annotate the GRBAS scale.

1.2. Research Questions

This doctoral thesis aims to advance voice disorder detection using ML methods by addressing these challenges in the field. The research questions are:

1) Which speech tasks are the most effective for collecting speech data for voice disorders detection?
2) Which acoustic features are robust across varying recording conditions for voice disorder assessment?
3) Can robust acoustic features predict symptom detection obtained from the GRBAS scale and the Voice Handicap Index for voice disorders without relying on definitive diagnostic classifications?

2. Work Plan

The following outlines the key phases of the research project:

1) **Data Collection** Speech data from 55 German-speaking ALS patients have already been collected at the Department of Neurology, TUM University Hospital, Munich, Germany and 50 Swiss-German-speaking MS patients were collected at the Department of Neurology, Inselspital, Bern, Switzerland. We also plan to collect speech data from 70 German-speaking patients visiting the Otolaryngology department at the TUM University Hospital, Munich, Germany for routine check-ups related to any voice impairment symptoms. These new patients may present with a wide range of voice disorders. Ethics approval has been submitted, with data collection scheduled to begin in September 2025. To ensure consistency across data collection studies, we use standardised speech tasks such a read speech, sustained utterances and free speech tasks, and similar recording procedures in the three data collection set ups. Microphones vary for each dataset. Further, each patient completes the Voice Handicap Index and other relevant questionnaires assessing the impact of the speech impairment in their quality of life. An example is the ALS-specific Quality of Life Instrument-Revised. Speech samples will be rated on the GRBAS scale by a minimum of three expert evaluators, including speech-language pathologists and otolaryngologists. This includes five components: Grade – overall grade of hoarseness, Roughness, Breathiness, Aesthenia – weakness, and Strain.
2) **Feature Robustness Evaluation** This phase involves evaluating the robustness of a huge variety of acoustic features extracted under different recording conditions (e.g., varying noise levels, and microphones). We also plan to artificially introduce noise using augmentation and pulse analysis. The goal is to ensure the reliability and generalisability of the features across diverse real-world settings.
3) **Modelling Speech Features for Symptom Monitoring** In this phase, we will employ regression and classification to predict individual GRBAS scale components (grade, roughness, breathiness, aesthenia and strain) to assess speech symptoms. Additionally, we will model the Voice Handicap Index and other similar questionnaires. Instead of focusing on definitive diagnoses (e.g., "patient has X"), our approach emphasises supporting the detection of speech symptoms in voice disorders (e.g., "the patient exhibits a high degree of breathiness").

2.1. Challenges

The following methodological challenges arise from this work: **Limited dataset size and diversity:** Even though we plan to collect data from all patients coming to the otolaryngologists department, this approach may not capture a wide range of disorders, as common disorders are likely to predominate while rare ones remain under-represented. This might limit the clinical utility of this work. However, we try to tackle this challenge by focusing on predicting the GRBAS components and questionnaires assessing the impact of the speech impairment in their quality of life. This project is meant to take a step toward voice disorder symptom detection. Future work should expand data collection to include other languages and a broader spectrum of disorders. **Data variability:** Since we are working with impaired cohorts, some patients might not be able to fully complete the speech tasks, which might lead to data inconsistency. Unfortunately, this is part of the data collection process and we might not be able to avoid all of these cases, rigorous quality control during data acquisition may minimise such cases. **Modelling complexity:** Many disorders have overlapping symptoms, making symptom monitoring a more suitable approach than definitive diagnosis. However, we rely on clinical scales that have a subjective nature, for example, the GRBAS assessments are annotated by experts and the Voice Handicap Index is self-assessed. This subjectivity may impact correlation and modelling efforts. Finding more objective measurements would need to be covered in future work.

3. Results and Discussion

Preliminary analyses have been conducted on the individual datasets for MS and ALS speech. Initial findings for MS were presented at [18], and for ALS can be found in [19]. For MS, we achieved an Unweighted Average Recall (UAR) of 71% and identified the Picture Description Task to contain the most information for MS detection compared to read speech tasks. For ALS, we reported an UAR of 88% using sustained utterances. These studies evaluated the efficacy of standard speech features in detecting these disorders through binary classification tasks. We also evaluated the performance of various standard speech tasks for classifying these disorders by training separate models for each task. We now aim to initiate cross-dataset and symptom detection analyses, encompassing feature robustness evaluation and modelling using speech data, questionnaires, and annotated scores.

3.1. Main Contributions

Assessing the robustness of acoustic features in the context of voice disorders is essential for the successful implementation of this technology. Without robust features, generalisation efforts are likely to fail. Furthermore, shifting from binary classification to symptom detection will provide better insights in real-world settings, moving beyond controlled, artificial studies. Given that speech alone may not suffice for definitive diagnoses, this research aims to contribute towards offering more actionable information, supporting practical applications in real-life scenarios. Additionally, the three datasets collected and analysed in this doctoral thesis will provide new insights on voice disorders for German-language and due to the standardised speech tasks, information on the effectiveness of each speech task will also be available. Due to the sensitive nature of the data and restrictions imposed by the ethics committee, these datasets will unfortunately not be publicly available.

4. Conclusion

This paper presents an on-going doctoral thesis aimed at advancing voice disorder screening in German-language contexts. The research focuses on three key areas: collecting three datasets in German-language for various voice disorders, including both organic and functional types, evaluating acoustic features robustness, and prioritising symptom detection over definitive disorder diagnosis.

5. References

[1] S. M. Cohen, J. Kim, N. Roy, C. Asche, and M. Courey, "Direct health care costs of laryngeal diseases and disorders," *Laryngoscope*, vol. 122, no. 7, pp. 1582–1588, 2012. [Online]. Available: https://doi.org/10.1002/lary.23189

[2] C. A. Rosen and T. Murry, "Nomenclature of voice disorders

and vocal pathology," *Otolaryngologic Clinics of North America*, vol. 33, no. 5, pp. 1035–1046, 2000. [Online]. Available: https://doi.org/10.1016/s0030-6665(05)70262-0

[3] T. V. Wang and P. C. Song, "Neurological voice disorders: A review," *International Journal of Head and Neck Surgery*, vol. 13, no. 1, pp. 32–40, 2022. [Online]. Available: https://www.ijhns.com/abstractArticleContentBrowse/IJHNS/8/13/1/28089/abstractArticle/Article

[4] A. L. Merati, Y. D. Heman-Ackah, M. Abaza *et al.*, "Common movement disorders affecting the larynx: a report from the neurolaryngology committee of the aao-hns," *Otolaryngology–Head and Neck Surgery*, vol. 133, no. 5, pp. 654–665, 2005. [Online]. Available: https://doi.org/10.1016/j.otohns.2005.05.003

[5] D. S. Chung, C. Wettroth, M. Hallett, and C. W. Maurer, "Functional speech and voice disorders: Case series and literature review," *Movement Disorders Clinical Practice*, vol. 5, no. 3, pp. 312–316, 2018. [Online]. Available: https://doi.org/10.1002/mdc3.12596

[6] M. Milling, F. B. Pokorny, K. D. Bartl-Pokorny, and B. W. Schuller, "Is speech the new blood? recent progress in ai-based disease detection from audio in a nutshell," *Frontiers in Digital Health*, vol. 4, p. 886615, 2022. [Online]. Available: https://doi.org/10.3389/fdgth.2022.886615

[7] P. Hecker, N. Steckhan, F. Eyben, B. W. Schuller, and B. Arnrich, "Voice analysis for neurological disorder recognition–a systematic review and perspective on emerging trends," *Front. Digit. Health*, vol. 4, p. 842301, 2022.

[8] M. Pedersen, "Artificial intelligence for screening voice disorders: Aspects of risk factors," *AJMCRR*, vol. 4, no. 2, pp. 1–8, 2025.

[9] A. Iyer, A. Kemp, Y. Rahmatallah, P. Lakshim, A. Glover, F. Prior, L. Larson-Prior, and T. Virmani, "A machine learning method to process voice samples for identification of parkinson's disease," *Scientific Reports*, vol. 13, p. 20615, 2023. [Online]. Available: https://doi.org/10.1038/s41598-023-47568-w

[10] S. Noto, Y. Sekiyama, R. Nagata, G. Yamamoto, and T. Tamura, "Analysis of speech features in alzheimer's disease with machine learning: A case-control study," *Healthcare (Basel)*, vol. 12, no. 21, p. 2194, Nov 2024. [Online]. Available: https://doi.org/10.3390/healthcare12212194

[11] G. S. Liu, N. Jovanovic, C. K. Sung, and P. C. Doyle, "A scoping review of artificial intelligence detection of voice pathology: Challenges and opportunities," *Otolaryngology–Head and Neck Surgery*, vol. 171, no. 3, pp. 658–666, 2024. [Online]. Available: https://doi.org/10.1002/ohn.809

[12] M. Chaiani, S. Selouani, and M. Boudraa, "Voice disorder detection using enhanced auditory perception-scaled spectrograms," in *2022 45th International Conference on Telecommunications and Signal Processing (TSP)*. IEEE, 2022, pp. 49–54.

[13] K. Ding, M. Chetty, A. Noori Hoshyar *et al.*, "Speech based detection of alzheimer's disease: a survey of ai techniques, datasets and challenges," *Artificial Intelligence Review*, vol. 57, p. 325, 2024. [Online]. Available: https://doi.org/10.1007/s10462-024-10961-6

[14] M. Pützer and W. Barry, "Saarbrücken Voice Database," http://www.stimmdatenbank.coli.uni-saarland.de/, 2025, institute of Phonetics, University of Saarland. Accessed April 2025.

[15] M. Huckvale and C. Buciuleac, "Automated detection of voice disorder in the saarbrücken voice database: Effects of pathology subset and audio materials," in *Proceedings of Interspeech 2021*, 2021, pp. 1399–1403. [Online]. Available: https://doi.org/10.21437/Interspeech.2021-1507

[16] 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. [Online]. Available: https://doi.org/10.1109/TAFFC.2015.2457417

[17] G. M. Stegmann, S. Hahn, J. Liss, J. Shefner, S. B. Rutkove, K. Kawabata, S. Bhandari, K. Shelton, C. J. Duncan, and V. Berisha, "Repeatability of commonly used speech and language features for clinical applications," 2020, dataset. [Online]. Available: https://doi.org/10.6084/m9.figshare.13317290

[18] M. Gonzalez-Machorro, P. Hecker, U. D. Reichel, H. N. Hammer, R. Hoepner, L. Pedrotti, A. Zmutt, H. Sagha, J. van Beek, F. Eyben, D. M. Schuller, B. W. Schuller, and B. Arnrich, "Towards Supporting an Early Diagnosis of Multiple Sclerosis using Vocal Features," in *Proc. INTERSPEECH 2023*, 2023, pp. 1518–1522.

[19] A. Mallol-Ragolta, M. Gonzalez-Machorro, R. von Heynitz, K. Scherzer, I. Cordts, and B. W. Schuller, "Early detection of als in absence of speech impairments with computer audition," in *Artificial Intelligence in Medicine. AIME 2025*, ser. Lecture Notes in Computer Science, R. Bellazzi, J. M. Juarez Herrero, L. Sacchi, and B. Zupan, Eds., vol. 15734. Springer, Cham, 2025. [Online]. Available: https://doi.org/10.1007/978-3-031-95838-0_27

Understanding and Exploiting Speech Foundation Models for Dialectal ASR: A Case Study on South Tyrolean

Domenico De Cristofaro

Free University of Bozen-Bolzano, Italy

Understanding and Exploiting Speech Foundation Models for Dialectal ASR: A Case Study on South Tyrolean

Domenico De Cristofaro[1]

[1]Faculty of Education, Free University of Bozen, Italy

ddecristofaro@unibz.it

Abstract

This PhD project investigates how speech foundation models can be adapted and interpreted for automatic speech recognition (ASR) of South Tyrolean, a low-resource and non-standardized regional dialect. The research is structured around three core stages: (1) analyzing available dialectal resources and expanding them through large-scale data collection, including web scraping; (2) developing unsupervised phoneme-level alignment strategies using CTC decoding and recursive correction; and (3) probing internal representations of self-supervised models to better understand their phonetic and structural knowledge. The project also explores reference-less evaluation metrics for scenarios where transcriptions are unavailable. Together, these components aim to build linguistically informed, interpretable ASR pipelines and contribute to inclusive, phoneme-level speech processing for under-resourced varieties.

Index Terms: speech recognition, forced alignment, model interpretability, low-resourced languages

1. Motivation and Research Question

This research addresses the growing need to extend automatic speech recognition (ASR) technologies to low-resource linguistic varieties, with a specific focus on regional dialects such as South Tyrolean [1, 2]. South Tyrolean belongs to the Southern Bavarian dialect continuum and exhibits a high degree of phonetic and lexical variation, shaped by both geographic fragmentation and sustained multilingual contact [3]. These varieties challenge conventional ASR models not only due to their internal variability, but also because they lack standardized orthographies and sufficient annotated training data. These constraints call into question the assumptions behind dominant ASR approaches—namely, the availability of orthographic transcriptions and large labeled corpora—and motivate a shift toward phoneme-based modeling. Phonemes offer a more stable unit for analysis and align more directly with the representations learned by modern self-supervised speech models. The central research question guiding this work is: *How can we adapt and interpret speech foundation models to support phoneme-level ASR for under-resourced dialects in the absence of standard orthography and large annotated corpora?* More broadly, this project aims to investigate how linguistic variation is encoded in multilingual speech models and how this knowledge can be harnessed to build inclusive, interpretable ASR systems for dialectal speech.

2. Key Challenges

Working with South Tyrolean and other low-resourced languages presents a number of interconnected challenges, both methodological and theoretical. The absence of a standardized orthography makes it impractical to rely on conventional ASR pipelines that assume word- or character-level transcriptions. This necessitates a shift toward phoneme-level modeling; however, phonemic annotations are often scarce, noisy, and inconsistent, which complicates both training and evaluation in supervised settings. A second major challenge concerns the interpretability of speech foundation models: understanding what type of phonetic or linguistic knowledge is implicitly encoded across different layers and architectures remains difficult. This task is particularly constrained by the lack of reliable time-aligned phoneme boundaries in dialectal corpora, which are essential for conducting fine-grained analysis of internal representations. Without such alignment, probing model behavior at the segmental level is infeasible, leaving foundational models largely opaque—especially when applied to dialectal input that diverges significantly from the training data. Finally, the lack of curated dialectal corpora introduces a practical limitation: high-quality, representative speech data in South Tyrolean is difficult to obtain, and existing resources are limited in size, coverage, and speaker diversity. This scarcity affects both the ability to adapt models effectively and to evaluate them in a linguistically meaningful way.

3. Methods and Work So Far

To address these challenges, we adopt a phoneme-based approach that bypasses the need for standardized orthographic transcriptions using the multilingual model `facebook/wav2vec2-large-xlsr-53`[1] [4]. While collecting data for South Tyrolean, we simultaneously focus on addressing stages 2 and 3 using available annotated corpora.

3.1. Data Collection and Preparation

While some existing resources are available, they were not specifically designed for ASR task. Nevertheless, we leverage these resources and supplement them with additional data. The AlpinLink corpus includes approximately three hours of read sentences in Tyrolean dialects [5], while the Vinko dataset provides isolated words in dialect [6]. To complement these resources with more spontaneous speech, we conducted large-scale web scraping of `podcast.de`, resulting in a collection of approximately 340 hours of South Tyrolean audio. This dataset significantly broadens the range of speakers and topics, providing real-world, naturally occurring dialectal input.

[1]`https://huggingface.co/facebook/wav2vec2-large-xlsr-53`

3.2. Forced Alignment and Evaluation

Precise temporal alignment is essential for probing the internal representations of speech models, yet such annotations are rarely available in low-resource settings. While high-resource languages benefit from tools like the Montreal Forced Aligner (MFA) or `torchaudio.functional.forced_align` [7, 8], these require transcriptions—often unavailable for dialects like South Tyrolean. To overcome this, we propose unsupervised alignment strategies based on CTC outputs from `wav2vec2-large-xlsr-53-espeak-cv-ft`, a model fine-tuned on phoneme-level transcriptions from Common Voice using eSpeak. We introduce recursive correction mechanisms that replace blank predictions using top-k phoneme context, enabling phoneme-level alignment without textual supervision. We evaluate our approach on TIMIT, which provides gold-standard alignments. While it does not yet match [7, 8] in accuracy, our method performs competitively without relying on transcripts. Figure 1 illustrates the alignment improvements achieved through recursive refinement.

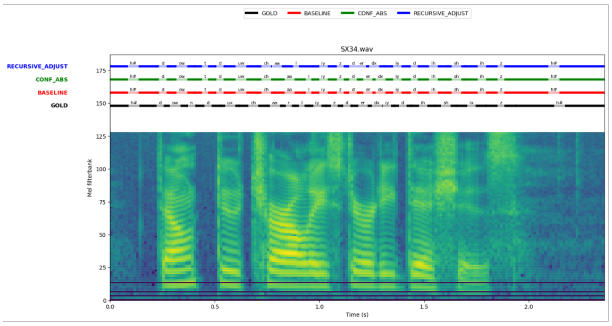


Figure 1: *Comparison of phoneme alignments from different decoding strategies on file* `SX34.wav`.

3.3. Model Probing and Representation Analysis

To investigate how Wav2Vec2 encodes phonetic structure, we conduct a series of probing experiments on its convolutional feature encoder [9]. Specifically, we compare CNN layer activations with MFCC-based features by training SVM classifiers to distinguish phonological categories, such as vowel frontness. As shown in Figure 2 and Table 3, classification accuracy improves across layers, peaking at Layer 5 (95.7%), and exceeding performance obtained with MFCCs alone. To further quantify the relationship between learned representations and traditional acoustic features, we compute mutual information (MI) between MFCCs and CNN activations at each layer (Table 1). The MI scores suggest that lower layers retain more phonetic detail, with information degrading in deeper layers, indicating a shift from acoustic to more abstract representations.

Layer	0	1	2	3	4	5	6
MI	0.0819	0.1040	0.0934	0.0390	0.0522	0.0078	0.0257

Table 1: *Mutual Information between MFCCs and CNN activations at each layer.*

Complementarily, we apply a Logit Lens [10, 11] to Whisper and experiment with removing upper encoder layers across six low-resource languages in the FLEURS dataset [12]. Due to the lack of phonetic transcriptions, Whisper was used as it is character-based model. The results show that both CER and WER improve in several languages when deeper layers are removed, suggesting that lower or mid-level representations may be more effective for dialectal generalization [13].

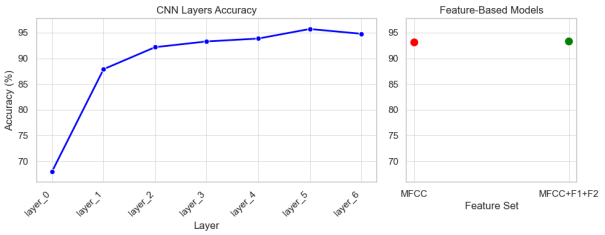


Figure 2: *Accuracy trend*

Feature	**C**	**Kernel**	**Gamma**	**Decision**	**Accuracy**
Layer 0	5.0	rbf	scale	ovr	0.6797
Layer 1	4.0	rbf	scale	ovr	0.8790
Layer 2	3.0	rbf	scale	ovr	0.9218
Layer 3	1.0	poly	scale	ovr	0.9330
Layer 4	1.5	poly	scale	ovr	0.9385
Layer 5	2.5	rbf	scale	ovr	**0.9572**
Layer 6	3.0	rbf	scale	ovr	0.9479
MFCC	0.5	rbf	scale	ovr	0.9311
MFCC + F1 + F2	3.5	rbf	scale	ovr	0.9330

Figure 3: *Best hyperparameters and accuracy scores for SVM classifiers using different feature sets*

CER results			
Language	**Best Layer**	**Last Layer**	**CER Decrease**
ky	**100.2** % (24)	157.0 %	56.8 %
ceb	**16.2** % (29)	36.7 %	20.5 %
ga	**75.2** % (29)	89.2 %	14 %
jv	**24.7** % (29)	35.0 %	10.3 %
kea	**35.0** % (29)	35.7 %	0.7 %
ast	**17.5** % (31)	17.6 %	0.1 %
ckb	**50.9** % (32)	**50.9** %	0.0 %
WER results			
Language	**Best Layer**	**Last Layer**	**WER Decrease**
ky	**112.0** % (24)	210.0 %	98 %
ga	**99.9** % (24)	121.4 %	21.5 %
ceb	**51.7** % (29)	67.4 %	15.7 %
jv	**75.3** % (29)	82.5 %	7.2 %
ckb	**1.05** % (24)	1.13 %	0.8 %
kea	**92.2** % (29)	92.6 %	0.4 %
ast	**63.7** % (31)	64.0 %	0.3 %

Table 2: *Early exiting results on FLEURS test set of low-resource languages. Numbers in parentheses indicate best layer as selected on dev set.*

4. Conclusion and Next Steps

At this stage, we have laid the foundation for phoneme-based, interpretable ASR in low-resource dialects by combining unsupervised alignment, representation probing, and large-scale dialectal data collection. Experiments on TIMIT, MFCC-based probing, and logit lens analysis highlight both the potential and the limits of current speech foundation models for dialectal input. A key future direction is the development of reference-less evaluation metrics, as transcriptions are often unavailable in low-resource settings. We will explore confidence-based and model-internal alternatives—such as logit-based scoring, CTC entropy, and mutual information with phonetic features—building on recent work in ASR quality estimation without ground truth [14, 15, 16]. Moving forward, we aim to scale the alignment pipeline to the full South Tyrolean dataset, refine these metrics, and use insights from model probing to support linguistically informed fine-tuning of pretrained models.

5. References

[1] M. Bartelds, N. San, B. McDonnell, D. Jurafsky, and M. Wieling, "Making more of little data: Improving low-resource automatic speech recognition using data augmentation," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023, pp. 715–729.

[2] J. Mainzinger and G.-A. Levow, "Fine-tuning asr models for very low-resource languages: A study on mvskoke," in *Proceedings of ACL Student Research Workshop*, 2024.

[3] S. Rabanus, "Tedesco," in *Lo spazio comunicativo dell'Italia e delle varietà italiane (Korpus im Text 7)*, R. Bauer and T. Krefeld, Eds., versione 91. [Online]. Available: https://www.kit.gwi.uni-muenchen.de/?p=13187&v=3

[4] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," in *Proceedings of Interspeech*, 2020.

[5] S. Rabanus, A. Kruijt, B. Alber, E. Bidese, L. Gaeta, G. Raimondi, P. B. Mas, S. Bertollo, J. Casalicchio, R. Cioffi, P. Cordin, M. Cosentino, S. Dal Negro, A. Glück, J. Kokkelmans, A. Murelli, A. Padovan, A. Pons, M. Rivoira, M. Tagliani, C. Saracco, E. Siviero, A. Tomaselli, R. Videsott, A. Vietti, and B. Vogt, "AlpiLinK Corpus 1.0.0," https://doi.org/10.5281/zenodo.8360170, 2023, in collaboration with the AlpiLinK team.

[6] P. Cordin, S. Rabanus, B. Alber, A. Mattei, J. Casalicchio, A. Tomaselli, E. Bidese, and A. Padovan, "Vinko. 2," in *Lo spazio comunicativo dell'Italia e delle varietà italiane*, ser. Korpus im Text, T. Krefeld and R. Bauer, Eds. Bayerische Akademie der Wissenschaften, 2019, vol. 67, version 67.

[7] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Proc. Interspeech 2017*, 2017, pp. 498–502.

[8] R. Huang, X. Zhang, Z. Ni, L. Sun, M. Hira, J. Hwang, V. Manohar, V. Pratap, M. Wiesner, S. Watanabe, D. Povey, and S. Khudanpur, "Less peaky and more accurate ctc forced alignment by label priors," 2024. [Online]. Available: https://arxiv.org/abs/2406.02560

[9] D. De Cristofaro and A. Vietti, "Evaluating the representation of vowels in wav2vec feature extractor: A layer-wise analysis using mfccs," 2025, manuscript submitted for publication. In Proceedings of the 2025 International Conference on Natural Language and Speech Processing (ICNLSP 2025).

[10] A. Langedijk, H. Mohebbi, G. Sarti, W. Zuidema, and J. Jumelet, "DecoderLens: Layerwise interpretation of encoder-decoder transformers," in *Findings of the Association for Computational Linguistics: NAACL 2024*, K. Duh, H. Gomez, and S. Bethard, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 4764–4780. [Online]. Available: https://aclanthology.org/2024.findings-naacl.296/

[11] nostalgebraist, "Interpreting gpt: The logit lens," https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens, 2020, accessed: 2025-05-19.

[12] 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," 2022. [Online]. Available: https://arxiv.org/abs/2205.12446

[13] R. S.-E. Shim, D. De Cristofaro, C. M. Hu, A. Vietti, and B. Plank, "Languages in multilingual speech foundation models align both phonetically and semantically," 2025, manuscript submitted for publication. In Proceedings of EMNLP 2025.

[14] K. A. Yuksel, T. Ferreira, A. Gunduz, M. Al-Badrashiny, and G. Javadi, "A reference-less quality metric for automatic speech recognition via contrastive-learning of a multi-language model with self-supervision," 2023. [Online]. Available: https://arxiv.org/abs/2306.13114

[15] M. Negri, M. Turchi, J. G. C. de Souza, and D. Falavigna, "Quality estimation for automatic speech recognition," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, 2014, pp. 1813–1823. [Online]. Available: https://aclanthology.org/C14-1171

[16] K. A. Yuksel, T. C. Ferreira, G. Javadi, M. Al-Badrashiny, and A. Gunduz, "Norefer: a referenceless quality metric for automatic speech recognition via semi-supervised language model fine-tuning with contrastive learning," in *Interspeech 2023*, 2023, pp. 466–470.

Domain Generalization with Bayesian Learning for Speaker Verification and Anti-Spoofing

Jin Li

The Hong Kong Polytechnic University, Hong Kong SAR

Domain Generalization with Bayesian Learning for Speaker Verification and Anti-Spoofing

Jin Li[1,2]

[1]Dept. of Electronical Electronic and Engineering, The Hong Kong Polytechnic University, Hong Kong SAR

[2]Speech@FIT, Brno University of Technology, Czechia

jin666.li@connect.polyu.hk

Abstract

Domain generalization is essential for ensuring robust performance on unseen domains while preserving strong discriminative capability on known ones. In real-world applications, automatic speaker verification (ASV) and anti-spoofing systems often suffer performance degradation under domain mismatch conditions. Prior work has introduced a relaxed instance frequency-wise normalization (RFN) module that combines two normalization techniques and is applied along the frequency axis to mitigate instance-specific domain variations in audio features. However, these approaches are typically built upon standard convolutional neural networks that learn a single, deterministic set of weights, thereby failing to account for the uncertainty inherent in out-of-distribution scenarios, particularly when training data from the target domain is limited or unavailable. In contrast, this work proposes a novel probabilistic framework based on variational Bayesian inference, which incorporates uncertainty estimation into a subset of model parameters. By jointly modeling aleatoric and epistemic uncertainties, the proposed approach enables more effective adaptation to distributional shifts in speaker verification. This leads to reduced overconfidence and better-calibrated predictions, both of which are crucial for improving domain generalization in real-world deployment scenarios. The proposed method is evaluated within two distinct frameworks: a ResNet-based architecture and a self-supervised learning (SSL) setup. Within the ResNet framework, we introduce a Bayesian neural network to model the frequency-wise normalization (RFN) weights along the spectral dimension. This yields a novel, plug-and-play explicit normalization module, which we refer to as Bayesian Weighted Relaxing Frequency-wise Normalization (BWRFN).

Index Terms: domain generalization, Bayesian learning, speaker verification, anti-spoofing

1. Introduction

Automatic speaker verification (ASV) seeks to authenticate a speaker's identity by analyzing their voice [1]. It is widely applied in a variety of domains, particularly in biometric authentication systems deployed on personal smart devices [2]. Recently, motivated by the strong feature extraction capabilities of deep neural networks (DNNs), a variety of deep learning-based approaches for speaker recognition have been proposed [3].

However, ASV systems deployed in real-world scenarios often suffer from performance degradation when domain mismatches arise between training and test conditions. These mismatches can be caused by various factors, such as differences in recording devices, acoustic environments, audio channels, and languages encountered in unknown domains. Moreover, ASV systems remain vulnerable to unseen spoofing attacks, which are frequently generated using novel algorithms not present during the training phase [4, 5]. To mitigate these issues, domain adaptation (DA) and domain generalization (DG) have emerged as two primary strategies for addressing domain mismatch in ASV [6].

Domain adaptation (DA) has been developed to address the domain shift problem by transferring knowledge from a well-labeled, resource-rich source domain to the target domain [7]. For instance, domain adversarial training aims to reduce domain mismatch by projecting features from different domains into a shared latent subspace [8], thereby mitigating domain variations across seen domains. Another approach, probabilistic linear discriminant analysis (PLDA) adaptation, leverages unlabeled in-domain data by first clustering the data and then refining the parameters of out-of-domain models [9]. Correlation Alignment (CORAL), a widely used backend method, further mitigates domain mismatch by whitening source-domain features and recoloring them using the covariance structure of the target domain [10, 11].

In contrast, domain generalization focuses on learning domain-invariant representations that remain robust to domain shift without requiring adaptation to the target domains [12]. In ASV, various approaches have been proposed in recent years to enhance generalization to unseen domains [13, 14]. More recently, a relaxed Instance Frequency-wise Normalization (RFN) has been explicitly integrated into the model to eliminate instance-specific domain discrepancies in acoustic features while minimizing the loss of useful discriminative information [15].

However, previous methods remain susceptible to overfitting and exhibit limited generalization performance, as the use of deterministic model parameters during inference neglects model uncertainty, which may lead to overfitting to the source domain characteristics [16–18].

2. Research Question and Motivation

For many years, the majority of research in speech processing has centered on enhancing model performance in matched domain settings, where training and evaluation data follow similar distributions. Although such approaches have driven significant advances in controlled environments, they often suffer from a critical limitation: models tend to overfit to the source domain and generalize poorly to unseen or mismatched conditions.

To address this limitation, the field has increasingly shifted its focus toward *domain generalization*, which aims to improve system robustness not only on seen (source) domains but also on previously unseen (target) domains. A promising avenue for achieving such generalization is Bayesian learning, a framework that enables models to estimate uncertainty and exploit

it during both training and inference. These uncertainty estimates, in particular, can help models detect domain shifts and adapt more reliably to novel or mismatched input conditions.

Given this context, the central research questions of this thesis are as follows:

1. Can we design a system that performs well on both seen and unseen domains without access to target-domain data during training?
2. Can we leverage uncertainty estimates to assess domain mismatch and improve generalization to unseen domains?
3. Can we develop a Bayesian learning framework that exploits posterior uncertainty under limited domain coverage, and apply it to enhance domain generalization for speaker verification and anti-spoofing?

3. Methodologies

3.1. Relaxed Instance Frequency-wise Normalization

One previous approach to reducing this variability involves normalizing the frequency components in a per-instance manner using global statistics computed from the entire audio signal. This method, introduced in [15], is referred to as "Instance Frequency-wise Normalization (INF)".

Given a mini-batch of N multichannel audio signals, its audio characteristics in the frequency domain can be represented by a tensor $\boldsymbol{x} \in \mathbb{R}^{N\times C\times F\times T}$, where N, C, F, and T are the mini-batch size, numbers of channels, frequency bins, and frames, respectively. The Relaxed instance Frequency-wise Normalization (RFN) [15] is calculated as follows:

$$F(\boldsymbol{x}) = \lambda \text{LN}(\boldsymbol{x}) + (1-\lambda)\text{IFN}(\boldsymbol{x}) \in \mathbb{R}^{N\times C\times F\times T}, \quad (1)$$

where the scalar $\lambda \in [0, 1]$ denotes the degree of relaxation.

3.2. Bayesian Relaxed Instance Frequency-wise Normalization

Eq. 1 interpolates between IFN and LN style normalization. However, the degree of relaxation, λ, of individual frequency bins are fixed and equal. We hypothesize that different frequency bins have different levels of domain dependence and should be weighted differently. We therefore extend RFN to weighted RFN (or WRFN), as follows

$$\begin{aligned} F(\boldsymbol{x}) =& \lambda \text{LN}(\boldsymbol{x}) \odot \sigma(\boldsymbol{w}_1) + (1-\lambda)\text{IFN}(\boldsymbol{x}) \odot \sigma(\boldsymbol{w}_2) \\ & \in \mathbb{R}^{N\times C\times F\times T}, \end{aligned} \quad (2)$$

where $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^F$ are broadcast along the mini-batch, time and channel axes, $\sigma(\cdot)$ is a sigmoid function that squashes the weights to values between 0 and 1, WLN is weighted LN, and WIFN is weighted IFN. We define $\boldsymbol{w} = [\boldsymbol{w}_1^\mathsf{T}, \boldsymbol{w}_2^\mathsf{T}]^\mathsf{T}$, where T is the transpose. Note that the frequency specific weights, $\lambda\sigma(\boldsymbol{w}_1)$ and $\lambda\sigma(\boldsymbol{w}_2)$ generally does not sum to one. However, this can to a large extent be compensated by the subsequent convolutionary layer.

The weights in WRFN, $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$, are deterministic and learned from training data. Since the purpose of these weights is to address domain variations and since there number of domains in the training data is scarce, we hypothesize that the weights will be prone to overfitting. To mitigate this, we improve the robustness of the weighted RFN (Eq. 2) by modeling the uncertainty in $\boldsymbol{w}$ with Bayesian techniques, which results in Bayesian Weighted Relaxed Instance Frequency-wise Normalization (BWRFN).

To account for uncertainty in $\boldsymbol{w}$, we assume that the weight vector is drawn from a prior distribution $p(\boldsymbol{w})$. Let $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{s}_i)\}_{i=1}^N$ denote the *training set*, considering of N utterances, each spoken by one of the S speakers. The loss can be designed by maximizing the variational lower bound $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ as described in [19]:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\sigma}) =& \overbrace{\sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{w}\sim q(\boldsymbol{w})} \log p_{\boldsymbol{\theta}}(\boldsymbol{s}_i|\boldsymbol{x}_i, \boldsymbol{w})}^{\mathcal{L}_1} \\ & - D_{KL}(q(\boldsymbol{w})||p(\boldsymbol{w})). \end{aligned} \quad (3)$$

where $p_{\boldsymbol{\theta}}(\boldsymbol{s}_i|\boldsymbol{x}_i, \boldsymbol{w})$ is the speaker probabilities given by softmax layer of the neural network and $\boldsymbol{\theta}$ are the parameters of the neural network (other than $\boldsymbol{w}$). This loss can be optimized and solved following [19].

During inference, the posterior predictive distribution:

$$\begin{aligned} \hat{\boldsymbol{y}} &= \int_{\boldsymbol{w}} p_{\boldsymbol{\theta}^*}(\hat{y}(\boldsymbol{z})|\boldsymbol{w}) p(\boldsymbol{w}|\mathcal{D})\, \mathrm{d}\boldsymbol{w} \\ &= \mathbb{E}_{\boldsymbol{w}\sim p(\boldsymbol{w}|\mathcal{D})}\left[p_{\boldsymbol{\theta}^*}(\hat{y}(\boldsymbol{z})|\boldsymbol{w})\right], \end{aligned} \quad (4)$$

where $\hat{y}(\cdot)$ represents the embedding layer output, $\boldsymbol{\theta}^*$ is an optimal parameters of the neural network (other than $\boldsymbol{w}$), and $\boldsymbol{z}$ is the Mel-spectrogram of a new utterance.

4. Future Research Directions

In the next phase of my research, I plan to extend uncertainty modeling to large language model (LLM)-based architectures for speaker verification and anti-spoofing tasks. Specifically, I will investigate how uncertainty estimation can be effectively integrated into LLM-driven systems to enhance their robustness under domain shift and out-of-distribution (OOD) conditions.

In parallel, I aim to explore strategies for augmenting neural representations by introducing controlled perturbations or noise during training. The objective is to increase the diversity of the feature space and improve the model's generalization capabilities across varying domains—particularly in scenarios where the target domain is unknown or significantly different from the source.

Collectively, these directions are expected to contribute to the development of more reliable and adaptable speaker verification systems capable of handling real-world distribution shifts.

5. Main Contributions of My Research

The primary contributions of this research are summarized as follows:

1. To the best of my knowledge, this is the first work applying a Bayesian learning framework to model uncertainty in neural networks for both speaker verification and anti-spoofing, improving robustness under domain shift through confidence-aware predictions.
2. Unlike methods requiring additional modules or parameters, the proposed framework enhances performance without introducing extra parameters, maintaining a lightweight design suitable for real-world deployment.
3. This work advances domain generalization research in speech processing under realistic mismatched conditions, offering practical insights and techniques for uncertainty-aware, robust system design.

6. Acknowledgements

This work was supported by The Hong Kong Polytechnic University's Research Student Attachment Programme.

7. References

[1] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65–99, 2021.

[2] K. A. Lee, B. Ma, and H. Li, "Speaker verification makes its debut in smartphone," *SLTC Newsletter*, February 2013.

[3] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Interspeech*, 2020, pp. 3830–3834.

[4] N. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification," in *Proc. Interspeech*, 2013, pp. 925–929.

[5] C. Zeng, X. Miao, X. Wang, E. Cooper, and J. Yamagishi, "Spoofing-aware speaker verification robust against domain and channel mismatches," in *Proc. STL*, 2024, pp. 1150–1157.

[6] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 4396–4415, 2022.

[7] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," *Proc. NIPS*, pp. 137–144, 2006.

[8] Q. Wang, W. Rao, S. Sun, L. Xie, E. S. Chng, and H. Li, "Unsupervised domain adaptation via domain adversarial training for speaker recognition," in *Proc. ICASSP*, 2018, pp. 4889–4893.

[9] D. Garcia-Romero, A. McCree, S. Shum, N. Brummer, and C. Vaquero, "Unsupervised domain adaptation for i-vector speaker recognition," in *Proc. Odyssey*, 2014, pp. 260–264.

[10] B. Sun, J. Feng, and K. Saenko, "Correlation alignment for unsupervised domain adaptation," *Domain Adaptation in Computer Vision Applications*, pp. 153–171, 2017.

[11] K. A. Lee, Q. Wang, and T. Koshinaka, "The CORAL+ algorithm for unsupervised domain adaptation of PLDA," in *Proc. ICASSP*, 2019, pp. 5821–5825.

[12] D. Li, Y. Yang, Y.-Z. Song, and T. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *Proce. the AAAI*, 2018, pp. 3490–3497.

[13] H. Zhang, L. Wang, K. A. Lee, M. Liu, J. Dang, and H. Meng, "Meta-generalization for domain-invariant speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1024–1036, 2023.

[14] J. Li, J. Han, S. Deng, T. Zheng, Y. He, and G. Zheng, "Mutual information-based embedding decoupling for generalizable speaker verification," in *Proc. Interspeech*, 2023, pp. 3147–3151.

[15] B. Kim, S. Yang, J. Kim, H. Park, J. Lee, and S. Chang, "Domain generalization with relaxed instance frequency-wise normalization for multi-device acoustic scene classification," in *Proc. Interspeech*, 2022, pp. 2393–2397.

[16] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya *et al.*, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information Fusion*, vol. 76, pp. 243–297, 2021.

[17] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher *et al.*, "A survey of uncertainty in deep neural networks," *Artificial Intelligence Review*, vol. 56, pp. 1513–1589, 2023.

[18] P. Moure, L. Cheng, J. Ott, Z. Wang, and S.-C. Liu, "Regularized parameter uncertainty for improving generalization in reinforcement learning," in *Proc. CVPR*, 2024.

[19] D. P. Kingma, T. Salimans, and M. Welling, "Variational dropout and the local reparameterization trick," *Proc. NIPS*, pp. 2575–2583, 2015.

Expressive Text-to-Speech System for Children in Low-Resource Settings

Shaimaa Alwaisi

Budapest University of Technology and Economics, Hungary

Expressive Text-to-Speech System for Children in Low-Resource Settings

Shaimaa Alwaisi

Department of Telecommunication and Media Informatics
Budapest University of Technology and Economics, Budapest, Hungary
Shaima.alwaisi@edu.bme.hu

Abstract

Modern expressive TTS systems for adult speech can generate highly expressive speech when provided with a sufficient amount of training data. However, expressive TTS systems for children are often restricted and have been under-explored to date. Child TTS systems' challenges arise from the scarcity of training datasets, difficulties in accessing them publicly, and various issues associated with these datasets, including noisy data and indiscernible speech.

This paper introduces an ongoing PhD thesis in which an expressive multi languages text-to-speech system for children is developed. This research first, introduces a high-quality expressive English child dataset. The dataset comprises 3 hours of speech collected from 25 children in grades ranging from third to fourth. grade. Next, we introduced Hungarian child speech corpus. The dataset comprises 8.3 hours of high-quality child speech from 62 speakers (4.5–5.5 years). Both datasets were developed to fill the gap caused by the lack of data and are available for research purposes.

Finally, we propose an expressive Multi-Style Child TTS in low-resource settings. Our model stands as the first expressive TTS model designed specifically for children. It is a style-based generative model based on StyleTTS architecture.

Index Terms: **Index Terms**: Text-To-Speech, Child TTS system, Expressive Child speech synthesis, dataset

1. Introduction

Although the success stories in style Text-To-Speech (TTS) and speech synthesis models for adults [1], [2], [3]], there remains a limitation in developing child TTS systems. Children represent a crucial user group for voice technology, a study has highlighted that, in the U.S. alone, approximately 12% of the users of AI speech assistants are under the age of 12 [4]. This significant portion of users underscores the importance of developing more expressive child-level TTS synthesis; however, this remains a challenge due to the limited availability of suitable training child datasets.

Early efforts in child-TTS began several decades ago, primarily using concatenative and parametric approaches [5], [6]. Yet, these approaches often produced audio that sounded robotic. TTS-based deep learning algorithms have improved the quality of generated speech [7]. Their work is based on Tacotron2 [8] and the WaveRNN [9] vocoder. To improve the quality of synthesized child speech, the WaveRNN vocoder is fine-tuned on a child speech dataset. Despite these efforts, the synthesized audio still suffers from poor quality. A few studies on emotional child TTS have been dedicated to examining the emotions expressed in children's narratives [10], [11].

1.1. Motivation and research goals

Expressive Child speech TTS systems have received limited attention in current studies, leading to a relative neglect of their development. Most of the TTS research focuses on synthesizing speech for adult people. Moreover, one of the biggest challenges in child speech synthesis domain is the scarcity of speech datasets.

To this end, we propose two datasets available for research purposes, The first is an expressive child dataset, encompassing 2 hours of speech gathered from 25 students in grades spanning from third to fourth grade. The new dataset containing 1200 audio samples has been transcribed at the word level, comprising 4 classes of speech style. The second dataset is Hungarian child speech corpus. The dataset comprises 8.3 hours of high-quality child speech from 62 speakers (4.5–5.5 years).

Our primary goal is to develop an expressive child TTS system for children with minimal training data. For this, we propose expressive text-to-speech (TTS) model for children. addressing a critical gap in the current landscape of TTS technologies. Our model stands as the first expressive TTS model designed specifically for children. It is a style-based generative model based on StyleTTS architecture.

2. Results so far

2.1. A Benchmark Dataset and Baseline for Expressive Child Synthesis

For expressive child TTS, the first dataset in expressive-style English child speech was created. This dataset is a high-quality expressive speech dataset that covers four classes of expressive styles (sadness, excitement, happiness, and neutral). The dataset is comprising 3 hours of speech collected from 25 children in grades ranging from third to fourth grade. All audio files were decoded to and down sampled to mono wav format (sampling rate: 22.05 kHz, lin, 16 bit PCM quantization). We performed down sampling of the source signals of the audio channel of MP4 44.1 kHz in FLAC encoding. We removed all such silence regions, leaving only a small fraction at the beginning and end of the audio samples roughly 4 seconds.

Table 1: *Expressive English child dataset description.*

Dataset	Statistics
Number of POI	**25**
Number of Utterances	**1200**
Number of hours	**2**
Number of filtered Videos	**100**

Number of videos per POI	**3**
Avg Number of utterances per POI	**48**
Avg length of utterances [s]	**6.71**

The key statistics of our dataset are described in Table 1. The gender distribution for our dataset is 52% and 48% for boys and girls respectively. The language is only American English. The percentage of styles is 30% sad, 35% excited, 15% happiness and 20%. neutral. The dataset is available for research purposes. To verify the effectiveness of our dataset, AutoVocoder models were trained and synthesized samples were generated. The models were trained on both LJSspeech large scale dataset and our dataset. Figure 1. shows a MUSHRA scores for each model.

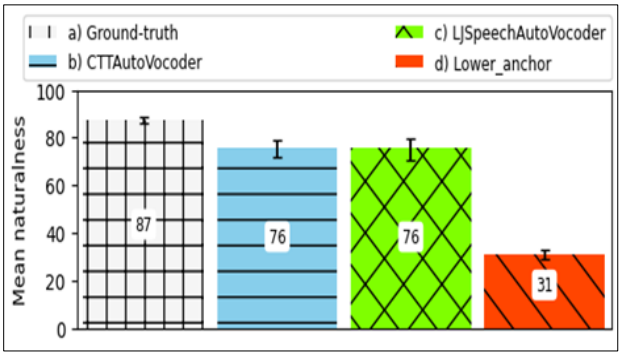


Figure 1: *The MUSHRA scores for the Mean naturalness are presented for (a) Ground truth (b) Our model CTT_ AutoVocoder (c) LJSpeech_ AutoVocoder (d) Lower anchor, with the average results shown. A higher value indicates better overall quality.*

2.2. A Hybrid Pipeline for Constructing the HUKILC-CO Hungarian Child Speech Dataset

we introduce HUKILC-CO, derived from the original HUKILC dataset, as the first Hungarian child speech corpus. We address key challenges including noisy environments, overlapping speech, and CHAT formatting—via a hybrid alignment pipeline integrating Rev AI's ASR[12] and Batchalign2's forced alignment [13]. We combined fuzzy text matching and temporal buffering to synchronize transcripts with audio, achieving precise utterance- and word-level alignment. The dataset comprises 8.35 hours of high-quality child speech from 62 speakers (4.5–5.5years). Table 2 shows summary of our dataset

Table 2: *Summary of HUKILC-CO Dataset Splits.*

Split	# Spk.	# Utt.	Dur. (hrs)	Avg. (s)
Train	50	6,414	6.51	3.66
Valid	5	687	0.69	3.61
Test	7	1,029	1.08	3.80
Sum	62	8,130	8.29	3.67

2.3. ChildStyleTTS: An Expressive Text-to-Speech System for Children in Low Resource Settings

We proposed ChildStyleTTS built upon StyleTTS[14], the first expressive text-to-speech (TTS) system for children. Our framework model is a non-autoregressive TTS model that produces speech with expressive tones. It leverages adaptive normalization as a style conditioning method to derive style embeddings from references.

ChildStyleTTS trained on our Expressive dataset. The training process is divided into two stages: the first stage, where the model reconstructs the mel-spectrogram by learning to generate it from the given text, and the second stage, where all modules remain fixed except for duration and prosody predictors. These predictors are trained to estimate duration, pitch, and energy from the input text. We conditioned the proposed TTS model on a speaker encoder H/ASP[15] , which generates speaker embeddings that encode speaker identity information extracted from reference utterances. We compared our model with the Multi-Speaker Child TTS model [7] and CoquiTTS/Vits [16]. A transfer learning method is used to finetune the baseline models CoquiTTS /Vits and Multi-Speaker Child TTS. In this study, we randomly selected 30 text samples from the test set for evaluation. A total of 19 listeners, aged between 23 and 54 years. According to the results, the preferences of listeners indicated that among the systems evaluated, the ChildStyleTTS model achieved a notably high mean similarity score of 68% relative to the ground truth. Figure 2. illustrates a MUSHRA scores for each model.

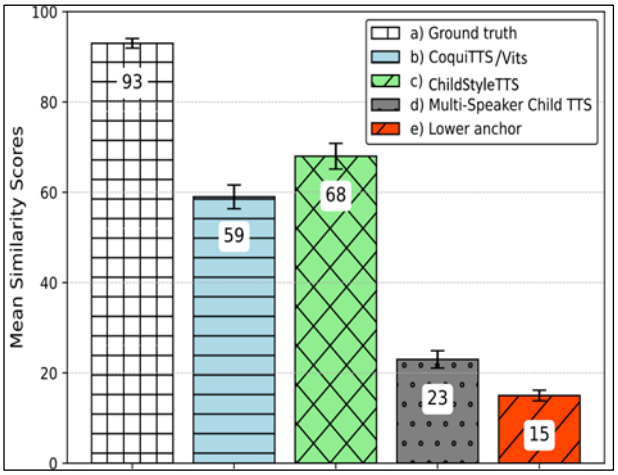


Figure 2: *The MUSHRA scores for the Mean naturalness are presented for (a) Ground truth (b) CoquiTTS /Vits (c) ChildStyleTTS (d) Multi-speaker ChildTTS and (e) Lower anchor, with the average results*

3. Future work

In future work, we further develop the ChildStyleTTS, we plan to expand the model's capabilities to support child speech in other languages (Hungrain and Turkish). This will help extend the applicability of ChildStyleTTS for global use, particularly in multilingual educational and assistive communication technologies.We will explore the utility of HUKILC-CO for training and evaluating TTS , We plan to benchmark state-of-the-art models such as StyleTTS2[17] , XTTS [18], YourTTS[19] including both training from scratch and fine-tuning approaches, to establish robust baselines for both English and Hungarian child TTS.

4. References

[1] Z. Shaheen, T. Sadekova, Y. Matveeva, A. Shirshova, and M. Kudinov, "Exploiting Emotion Information in Speaker Embeddings for Expressive Text-to-Speech," in *INTERSPEECH*, 2023, pp. 2038–2042.

[2] W. Zhao and Z. Yang, "An emotion speech synthesis method based on vits," *Applied Sciences*, vol. 13, no. 4, p. 2225, 2023.

[3] Y. Meng *et al.*, "CALM: Contrastive Cross-modal Speaking Style Modeling for Expressive Text-to-Speech Synthesis," *arXiv preprint arXiv:2308.16021*, 2023.

[4] R. Paul, ", "Voice search statistics 2024 (usage & emerging trends".," 2024.

[5] O. Watts, J. Yamagishi, S. King, and K. Berkling, "Synthesis of child speech with HMM adaptation and voice conversion," *IEEE Trans Audio Speech Lang Process*, vol. 18, no. 5, pp. 1005–1016, 2010, doi: 10.1109/TASL.2009.2035029.

[6] H. Zen, A. Senior, and M. S. Google, "STATISTICAL PARAMETRIC SPEECH SYNTHESIS USING DEEP NEURAL NETWORKS."

[7] R. Jain, M. Y. Yiwere, D. Bigioi, P. Corcoran, and H. Cucu, "A Text-to-Speech Pipeline, Evaluation Methodology, and Initial Fine-Tuning Results for Child Speech Synthesis," *IEEE Access*, vol. 10, pp. 47628–47642, 2022, doi: 10.1109/ACCESS.2022.3170836.

[8] Y. Wang *et al.*, "Tacotron: Towards End-to-End Speech Synthesis," Mar. 2017, [Online]. Available: http://arxiv.org/abs/1703.10135

[9] N. Kalchbrenner *et al.*, "Efficient neural audio synthesis," in *International Conference on Machine Learning*, PMLR, 2018, pp. 2410–2419.

[10] C. O. Alm and R. Sproat, "Perceptions of emotions in expressive storytelling," in *9th European Conference on Speech Communication and Technology*, 2005, pp. 533–536. doi: 10.21437/interspeech.2005-334.

[11] Harikrishna D M, Gurunath Reddy M, and K. S. Rao, "Multi-stage children story speech synthesis for Hindi," in *2015 Eighth International Conference on Contemporary Computing (IC3)*, IEEE, Aug. 2015, pp. 220–224. doi: 10.1109/IC3.2015.7346682.

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

[13] H. Liu and B. MacWhinney, "Morphosyntactic analysis for CHILDES," *arXiv preprint arXiv:2407.12389*, 2024.

[14] Y. A. Li, C. Han, and N. Mesgarani, "Styletts: A style-based generative model for natural and diverse text-to-speech synthesis," *IEEE J Sel Top Signal Process*, 2025.

[15] H. S. Heo, B.-J. Lee, J. Huh, and J. S. Chung, "Clova baseline system for the voxceleb speaker recognition challenge 2020," *arXiv preprint arXiv:2009.14153*, 2020.

[16] G. Eren and The Coqui TTS Team, ", 'Coqui TTS.' ," 2021.

[17] Y. A. Li, C. Han, V. Raghavan, G. Mischler, and N. Mesgarani, "Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models," *Adv Neural Inf Process Syst*, vol. 36, pp. 19594–19621, 2023.

[18] E. Casanova *et al.*, "Xtts: a massively multilingual zero-shot text-to-speech model," *arXiv preprint arXiv:2406.04904*, 2024.

[19] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *International conference on machine learning*, PMLR, 2022, pp. 2709–2720.

Assessing Quality Dimensions in Speech Foundation Models for Inclusive Dutch ASR

Dragoș Alexandru Bălan

University of Twente, The Netherlands

Assessing Quality Dimensions in Speech Foundation Models for Inclusive Dutch ASR

Dragoș Alexandru Bălan[1]

[1]Human-Media Interaction, University of Twente, the Netherlands

d.a.balan@utwente.nl

1. Motivation

State-of-the-art Automatic Speech Recognition (ASR) systems excel in transcribing standard read speech from native speakers of high-resourced languages like English and Mandarin [1]. Many of these systems have been developed as multilingual speech foundation models capable of handling multiple tasks aside from ASR, such as speech translation, language identification, and even text-to-speech [2, 3]. Despite this, they still struggle with inclusivity, meaning underrepresented speaker groups, such as children, the elderly, or non-natives [4, 5], as well as different speech styles [6] and lower-resourced languages [3], are often overlooked due to data and research limitations. However, how can ASR research further develop to be inclusive and generalizable across different populations, languages, and registers? One facet related to inclusivity and generalisability is quality. Ideally, high quality needs to be achieved for a wide range of use cases. However, defining quality is difficult as it is not a one-dimensional, trivial concept, but rather it encompasses different aspects and can be adjusted according to the intended application of the ASR model.

One aspect is data quality. Data would be considered "of high quality" if it contains complete labels and metadata that can provide a comprehensive overview of the dataset as a whole. Currently, this is not the case for the majority of corpora available, which further hinders ASR research and understanding of its behaviour. Another aspect is the methodology and choosing the right components in the training pipeline to ensure the highest model performance and quality. Various methodological approaches can be observed in the literature, but there are not enough studies that compare them in the context of speech foundation models and inclusivity. Lastly, evaluation is one of the most important aspects as it is the main measure of quality. Crafting the right evaluation protocol can reveal many insights about the choices for the other two aspects, which in turn will ensure the highest quality possible of ASR models.

Therefore, it is vital to define and assess the quality in the context of these facets to gain a better understanding of the behaviour of ASR, particularly of large-scale speech foundation models. The research question I will address during my doctoral programme is the following: **How can the quality of Speech Foundation Models be assessed in the context of Inclusive Dutch Automatic Speech Recognition?** The next chapter will provide the outline of how I will approach this question.

2. Project outline

There are three major concepts that I intend to study in the context of quality and which will represent the chapters of my dissertation:

2.1. Data Quality through Data Selection

In this chapter, I will attempt to understand the effects of data curation on model performance. By data selection, I refer to the process of sampling subsets from a large corpus to extract only the speech segments relevant to the use cases of the ASR system. The use cases I will investigate will be low-resource, such as speech from children, the elderly, non-natives, people with dialects, etc. Data curation is severely understudied and, throughout my PhD, I hope to underline its importance in developing high-quality, inclusive ASR models. The next planned study will be placed within this chapter and will look at three methods of child speech training data selection: random selection, metadata-based selection (selecting data from specific areas of the dataset based on the metadata fields and information available), and using a classifier to identify segments of child speech. These experiments will be conducted on several datasets: a large archive of 81,000 hours with scarce metadata available, as well as a smaller dataset, JASMIN-CGN[7], which contains child speech. The latter corpus will be described within the first study chapter. In terms of the classifier, an off-the-shelf solution will be used, based on the wav2vec 2.0 architecture[1].

2.2. Training Methodology and Model Quality

Within this chapter, I will review several state-of-the-art approaches to training large-scale models in the context of quality. There are different aspects through which this can be approached: the model itself, the pre-processing steps, the feature extractor, the output tokens, or the overarching training pipeline. Through suitable architecture and methodology choices, inefficiency can be greatly reduced and high quality can be guaranteed. Thus, emphasizing correct architectural choices within this chapter will hopefully lead to better designs of high-quality speech foundation models and better-targeted ASR systems.

2.3. Quality in Evaluation

The last chapter will cover evaluation as extensively as possible. Conducting a proper evaluation of speech foundation models will ensure our results are valid, comparable, and reflect the performance of the systems in more realistic scenarios. Much of the development of speech foundation models does not use comprehensive and detailed evaluations, which is further emphasized by later studies that reveal their poor performance under low-resource conditions. To progress the design of robust evaluation protocols, the potential studies under this chapter would include assessing the relevancy of certain metrics, data selection for evaluation, as well as frameworks to standardise

[1]https://huggingface.co/bookbot/wav2vec2-adult-child-cls

the evaluation step of ASR, such as benchmarking. I have conducted my first study within this chapter by benchmarking various open-source, state-of-the-art speech foundation models for different use cases of Dutch, which will be extensively covered in the next section.

3. Discussion of First Study

For my first study, I investigated the performance of the latest open-source speech foundation models available for different Dutch speech conditions. My approach involved a more standardised method of evaluation called benchmarking, through which I ensured transparency of the evaluation and comparability of the results. The investigated models were Whisper [3], XLS-R [8], and MMS [2], with the baseline being Kaldi_NL [9], a Dutch ASR model developed using the Kaldi toolkit [10]. The datasets used are JASMIN-CGN [7], a Dutch corpus of children, elderly, and non-native speech, N-Best (not reported here) [11], a dataset of broadcast news and telephone conversations, and COPAS [12], a corpus which contains both healthy and pathological speech. Metrics used include Word Error Rate (WER), Real-Time Factor (RTF), and memory usage. The latter two are not reported in this paper.

Table 1: *WER scores for read *D*utch and *F*lemish in JASMIN-CGN. C=Children, T=Teenagers, NNC=Non-Native Children, NNA=Non-Native Adults, E=Elderly. Best results are highlighted in* ***bold****.*

Model	DC	DT	NNC	NNA	DE
Kaldi_NL	28.1%	16.2%	43.6%	45.3%	20.9%
Whisper v2	**20.3%**	**11.3%**	**29.9%**	**30.6%**	**13.7%**
Whisper v3	28.1%	25.2%	50.9%	62.6%	27.6%
XLS-R	22.4%	13.3%	33.8%	36.1%	17.2%
MMS-1b-102	31.6%	20.3%	54.2%	55.1%	23.9%
MMS-1b-all	28.9%	20.0%	50.1%	54.0%	28.3%
Model	FC	FT	NNC	NNA	FE
Kaldi_NL	59.2%	33.5%	51.3%	43.3%	24.7%
Whisper v2	**42.4%**	**11.7%**	**19.9%**	**21.0%**	16.7%
Whisper v3	57.2%	30.6%	44.4%	41.1%	38.7%
XLS-R	47.4%	13.3%	30.1%	26.8%	**16.4%**
MMS-1b-102	55.3%	22.4%	43.0%	37.0%	23.0%
MMS-1b-all	49.2%	21.8%	34.9%	35.8%	22.3%

Results for JASMIN-CGN can be seen in Table 1. Looking at the WER scores, large differences can be seen between the different speaker groups, with native teenagers having the lowest score (11.3%) and non-native adults having the highest score (30.6%) for Whisper large-v2. Differences can also be noticed between models, with Whisper large-v3 performing the worst on average and Whisper large-v2 performing the best.

Table 2: *WER scores for COPAS. N=Normal/healthy speech, D=Dysarthria, H=Hearing impairment, L=Laryngectomy, C=Cleft palate, V=Voice disorder. Best results are highlighted in* ***bold****.*

Model	N	D	H	L	C	V
Kaldi_NL	47.4%	78.0%	78.7%	86.8%	94.8%	56.3%
Whisper v2	**15.6%**	**44.7%**	51.8%	**44.8%**	**64.0%**	30.3%
Whisper v3	20.1%	49.4%	**46.3%**	52.7%	69.6%	**29.5%**
XLS-R	38.3%	64.7%	70.5%	73.7%	96.7%	50.4%
MMS-1b-102	39.3%	71.5%	76.4%	80.5%	96.5%	50.4%
MMS-1b-all	34.4%	65.8%	75.3%	77.4%	97.1%	56.7%

The results for COPAS are presented in Table 2. The differences in WER are even larger than in the case of JASMIN-CGN. Pathological speakers suffer from reduced ASR performance with scores ranging from 29.5% to 64%, which far exceed the WER obtained for healthy speakers. Healthy speech scores the lowest WER for Whisper large-v2 with a score of 15.6%.

Overall, these results show that there are biases present for speech foundation models, with some speaker groups exhibiting a stronger bias than others. The results provide a solid starting point for my research that aims to ultimately improve not only quality, but also inclusivity and generalisability of speech foundation models.

4. Potential Challenges

One of the causes of underrepresented speaker and speech groups having low performance is the lack of data available. I expect the lack of audio data to be a potential issue in my research, which might limit the breadth of my studies. Working with speech foundation models, particularly training them from scratch, can also generate problems. Pre-training foundation models would be part of my second chapter about methodology. However, my doctoral programme is funded by a project that works towards creating a speech foundation model for Dutch, which would reduce the potential of issues arising due to a lack of hardware resources, data preprocessing, and so forth.

Another challenge could be the difficulty of designing a sensible evaluation that is as transparent as possible for speech foundation models. Explicitly, tailoring the evaluation to the various conditions that I want to assess could increase the complexity of designing the evaluation. There are many factors to consider when conducting an evaluation of ASR models, which can affect the interpretability and robustness of the results, thus leading to researchers and users being misled in terms of transcription expectations if done improperly. At the same time, a considerable amount of time would have to be spent to carefully design an evaluation that makes sense and is methodologically robust, such that quality can be correctly assessed.

5. Expected Contributions

I expect to contribute towards building higher-quality speech foundation models and ASR systems from the different perspectives mentioned in this paper. Through a smart selection of training data, I expect to generate more inclusive and balanced transcriptions and reduce the practice of utilizing all the data available, which would harm the environment and hinder the development speed. Through a review of suitable methodologies, I expect to contribute towards more conscious choices that are tailored for the goals set by researchers. Improving the evaluation quality would lead to higher transparency and robustness of the obtained results, as well as a better understanding of model behaviour. The outcomes would not be restricted to Dutch only, but would also extend beyond and translate easily to other languages.

6. References

[1] J. R. Green, R. L. MacDonald, P.-P. Jiang, J. Cattiau, R. Heywood, R. Cave *et al.*, "Automatic speech recognition of disordered speech: Personalized models outperforming human listeners on short phrases." in *Interspeech*, vol. 2021, 2021, pp. 4778–4782.

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

[3] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.

[4] M. Fuckner, S. Horsman, P. Wiggers, and I. Janssen, "Uncovering bias in asr systems: Evaluating wav2vec2 and whisper for dutch speakers," in *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, 2023, pp. 146–151.

[5] S. Feng, B. M. Halpern, O. Kudina, and O. Scharenborg, "Towards inclusive automatic speech recognition," *Computer Speech Language*, vol. 84, p. 101567, 2024.

[6] A. Lopez, A. Liesenfeld, and M. Dingemanse, "Evaluation of automatic speech recognition for conversational speech in Dutch, English and German: What goes missing?" in *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, 2022, pp. 135–143.

[7] C. Cucchiarini, J. Driesen, H. Van hamme, and E. Sanders, "Recording speech of children, non-natives and elderly people for HLT applications: the JASMIN-CGN corpus." in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, May 2008.

[8] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal *et al.*, "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale," in *Interspeech 2022*, 2022, pp. 2278–2282.

[9] S. O. Spraaktechnologie, "GitHub - opensource-spraakherkenning-nl/Kaldi_NL," https://github.com/opensource-spraakherkenning-nl/Kaldi_NL, 2023, [Accessed 26-05-2025].

[10] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel *et al.*, "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Dec. 2011.

[11] D. A. van Leeuwen, J. Kessens, E. Sanders, and H. van den Heuvel, "Results of the N-Best 2008 Dutch Speech Recognition Evaluation," in *Interspeech 2009*, 2009, pp. 2571–2574.

[12] C. Middag, Y. Saeys, and J.-P. Martens, "Towards an ASR-free objective analysis of pathological speech," in *INTERSPEECH 2010*, 2010, pp. 294–297.

Speech Data Collection and Disfluency Analysis in Ukrainian: Towards Robust ASR for Low-Resourced Language

Anna Havras

VoiceInteraction / University of Lisbon / INESC-ID, Portugal

Speech Data Collection and Disfluency Analysis in Ukrainian: Towards Robust ASR for Low-Resourced Language

Anna Havras[1,2,3]

[1]VoiceInteraction, Lisbon, Portugal
[2]FLUL, University of Lisbon, Lisbon, Portugal
[3]INESC-ID, Lisbon, Portugal
anna.havras@voiceinteraction.ai

1. Research questions and motivation

This project focuses on the approaches to handle code-switching and code-mixing in speech involving Ukrainian and Russian, in the domain of Broadcast News (BN), including political debates, news reports, meetings, street interviews, among others. Ukrainian is still an under-researched language in the context of ASR [1, 2, 3, 4]. A key challenge is the frequent code-switching and code-mixing between Ukrainian and Russian [5, 6], with speakers often alternating between the two languages. This multilingual context, along with the presence of dialectal variation in both Ukrainian and Russian, makes acoustic and language modeling more challenging. Despite its widespread occurrence in natural speech, this phenomenon has received limited attention in ASR research [7, 8, 9]. Improving ASR for Ukrainian, especially under these multilingual conditions, could significantly advance language technologies in the region.

The core goals of this research consist of four main tasks: (i) collecting current information and creating linguistic resources; (ii) developing transcription guidelines for Ukrainian speech, including code-switching and code-mixing; (iii) creating validation tests to evaluate ASR performance; and (iv) integrating these elements into ASR systems and conducting benchmarking.

2. Key challenges

Adapting an ASR system for the Ukrainian language is complex due to several linguistic and technological challenges. One major challenge is that Ukrainian speakers frequently switch between Ukrainian and Russian during conversation [5, 6], either by alternating between the two languages (code-switching) or mixing elements of both within a single sentence (code-mixing). For instance, code-mixing can occur in sentences like "Vona pishla na rabotu ranishe," where the Russian word "rabotu" (work) appears in Ukrainian sentence.

These phenomena are particularly common in the Ukrainian language, occurring regularly in contexts such as political debates and news broadcasts. As a result, ASR systems face increased complexity when processing bilingual or mixed-language speech.

Similarly, while ASR systems tend to perform well with planned speech, their accuracy typically declines when dealing with spontaneous speech, which includes disfluencies such as repetitions, pauses, lengthenings, and other speech disruptions [10, 11, 12, 13]. In recent years, researchers have begun to develop ASR models that include disfluencies, which have improved the recognition and processing of spontaneous speech [14, 15, 16]. In the case of Ukrainian, studies have shown that incorporating Filled Pauses (FPs) (e.g., uh, uhm) can enhance ASR performance, as these were reported as being the most common type of disfluency in Ukrainian speech [17, 18]. However, research on FPs in Ukraine remains limited [19, 17, 18], particularly with respect to their acoustic properties and potential integration into ASR systems. Therefore, this research also aims to explore the FPs types present in Ukrainian, to incorporate them into the ASR system, and compare them to those already described and studied for Russian.

A key challenge in adapting ASR systems for Ukrainian is the limited resources of text and speech data to train the required models, specifically data that captures real-world usage. In particular, code-switching and code-mixing between Ukrainian and Russian, common in everyday spoken communication, are poorly represented in existing resources.

2.1. Phonetic challenges

Frequent switching between Ukrainian and Russian, two closely related Slavic languages, presents a major challenge for ASR systems, which must recognize and process both languages, sometimes within a single utterance.

The Russian spoken by Ukrainian speakers differs from native Russian in several phonetic aspects. For example, while all Russian consonant phonemes are present in Ukrainian, six Ukrainian consonant phonemes have no equivalent in Russian [5]: /ɦ/, /dz/, /dʒ/, /ʒʲ/, /tsʲ/, and /tʃ/. Additionally, the Ukrainian voiced glottal fricative /ɦ/ and the Russian voiced velar plosive [g] share the same written form. Ukrainian speakers often replace the Russian [g] with the Ukrainian [ɦ] in pronunciation, which gives their Russian speech a noticeable accent [5].

Finally, a key difference involves the frequency of the front open-mid vowel [ɛ] after a palatalized consonant (e.g., Russian word [ˈlʲɛs] ('forest')). This vowel in that position is quite rare in Ukrainian but common in Russian.

2.2. Filled pauses

Since research has shown that incorporating FPs can improve ASR performance, and they have been described as the most common type of disfluency in Ukrainian [17, 18], we propose including them in the ASR system to increase recognition accuracy. Furthermore, given their frequency and spontaneous nature, FPs may serve as native linguistic markers that assist in the identification or analysis of code-switching and code-mixing phenomena. Our hypothesis is that FPs could function as cues that signal transitions or hesitation points where language alternation is more likely to occur. However, this area remains underexplored in Ukrainian, and there is no established typology of FPs in this language.

Therefore, our initial goal was to explore FPs present in Ukrainian speech, focusing on their occurrence across three domains: Broadcast News (BN), Political Interviews (PI), and Guest Interviews (GI). For this, we used a sample of 4 hours

Table 1: *Types and rate of filled pauses (overall and in BN domain).*

Type	Overall		BN Domain	
	No. of FPs	% of FPs	No. of FPs	% of FPs
%a	265	**54.0**	276	**56.91**
%ɪ	177	**36.0**	164	**33.81**
%m	27	5.5	12	2.47
%e	6	1.2	3	0.62
%am	5	1.0	16	3.3
%u	4	0.8	–	–
%mna	2	0.4	–	–
%na	2	0.4	3	0.62
%au	1	0.2	1	0.21
%mɪ	1	0.2	–	–
%nɪ	1	0.2	–	–
%nʲa	1	0.2	2	0.41
%aj	–	–	2	0.41
%av	–	–	2	0.41
%jaj	–	–	1	0.21
%ej	–	–	1	0.21
%om	–	–	1	0.21
%ma	–	–	1	0.21

of publicly available data. Subsequently, we focused on FPs occurring specifically within the BN domain, using 5 hours of sample data, since this is the target domain for our ASR system. The results are demonstrated in Table 1, using the percentage sign (%) as a marker to indicate FP.

2.3. Strategies for bilingual speech recognition

Another major challenge are phonetic inventories in bilingual speech. Currently, there are multiple strategies for dealing with the phonetic differences between languages that involve code-switching and code-mixing [20, 21, 22]. These include monolingual approaches, in which separate phonetic alphabets, vocabularies, and text corpora are created for each language, and mixed approaches, where resources are integrated into a single unified system. The effectiveness of these strategies remains an open question, particularly for Ukrainian. For example, [20] used a multilingual strategy - Grapheme-to-Phoneme (G2P) models from Russian, Belarusian, German, and English. The approach involved mapping Ukrainian graphemes to those of related languages. Russian and Belarusian languages demonstrated the best results.

Additionally, [5] has investigated the use of interlingual systems to address code-switching using a Large Vocabulary Continuous Speech Recognition (LVCSR) approach. These studies have examined different phonetic inventory configurations and found that even slight phonetic additions, such as incorporating the Russian [ʲɛ], can improve recognition accuracy in speech involving code-switching and code-mixing.

Furthermore, [21] explored the creation of separate ASR models for both Russian and Ukrainian, as well as unified models that shared acoustic, pronunciation, and language models. Their findings indicated that ASR systems based on Ukrainian models consistently outperformed those based on Russian, regardless of whether they were used in separate or unified configurations. This reinforces the importance of understanding the specific phonetic and orthographic characteristics of Ukrainian to develop a robust ASR system for this language.

A similar challenge has been addressed in other languages, such as Mandarin-English code-switching, as shown in the ASRU 2019 Code-Switching Challenge [7]. The study demonstrated that both traditional hybrid and end-to-end ASR systems can benefit from unified modeling approaches, the integration of artificial code-switched text, and data augmentation techniques such as spec-augment. These methods significantly improved recognition accuracy and the strategy may be relevant for Ukrainian as well.

3. Expected results

This research aims to adapt an ASR system capable of processing Ukrainian BN content and handling spontaneous speech, including instances of code-switching and code-mixing with Russian. One of the primary outcomes will be the creation of linguistic resources for Ukrainian, which include phonetic inventories, pronunciation rules, and disfluency patterns, particularly FPs. These resources will serve as the foundation for developing robust acoustic, pronunciation, and language models. Additionally, a manually created golden set of transcribed broadcast news segments will be created to support the evaluation and benchmarking of ASR performance under realistic conditions involving spontaneous and mixed-language speech.

The project will integrate these resources into ASR systems and evaluate performance using standard metrics such as WER, precision, F-measure, and processing time [23].

4. Plans for the future

The future steps of this project are divided into several key phases focused on adapting an ASR system for the Ukrainian language. Initially, linguistic data, covering phonetic, phonological, and grammatical features, as well as code-switching and code-mixing with Russian, will be collected from public corpora, including broadcast news and journalistic content.

To ensure the quality of the training data, detailed transcription guidelines will be developed for Ukrainian speech. These guidelines will outline standardized procedures for transcribing various speech phenomena, including disfluencies, such as FPs, repetitions, prolongations, *inter alia*, and the accurate representation of code-switching and code-mixing instances. By applying these consistent rules, the transcriptions will then serve as an input for the extraction of linguistic patterns to build statisticals models. Text data from articles and newspapers will be used to build a language model, while audio data from news broadcasts will support the development of an acoustic model.

A Russian ASR system will be adapted to Ukrainian using a monolingual approach, with the potential for multilingual integration later. Since this is an adaptation process, the phone set used in the Russian system will be aligned with that of Ukrainian, with necessary adjustments made to account for phonological differences between the two languages.

Finally, we will explore different ASR systems available in the industry to determine which best fits the use case of recognizing code-switching speech in Ukrainian. Although WER will serve as the main performance indicator, additional factors such as processing time will also be considered to provide a comprehensive evaluation. Based on the results, the most appropriate solution will be selected and further refined to improve system accuracy.

5. Main contributions of the research

This research aims to advance the study of ASR for Ukrainian and to contribute to multilingual ASR development. By exploring strategies for handling code-switching and evaluating industry ASR solutions, we aim to improve recognition of linguistically complex Ukrainian speech and to technological advancement for Ukrainian and related languages.

6. References

[1] P. Mozharov, O. Moskaliuk, S. Zaitsev, and M. Vovk, "Experimental comparison ofspeech recognition systems for ukrainian language," in *2017 IEEE First International Conferenceon Data Stream Mining & Processing (DSMP)*, 2017, pp. 45–49.

[2] V. Pratap, A. Sriram, P. Tomasello, A. Hannun, V. Liptchinsky, G. Synnaeve, and R. Collobert, "Massively multilingual asr: 50 languages, 1 model, 1 billion parameters," *arXiv preprint arXiv:2007.03001*, 2020.

[3] M. Sazhok, A. Poltyeva, V. Robeiko, R. Seliukh, and D. Fedoryn, "Punctuation restoration for ukrainian broadcast speech recognition system based on bidirectional recurrent neural network and word embeddings." in *COLINS*, 2021, pp. 300–310.

[4] L. Kobylyukh, Z. Rybchak, and O. Basystiuk, "Analyzing the accuracy of speech-to-text apis in transcribing the ukrainian language." in *COLINS (2)*, 2023, pp. 217–227.

[5] T. Lyudovyk and V. Pylypenko, "Code-switching speech recognition for closely related languages," in *Spoken Language Technologies for Under-Resourced Languages*, 2014.

[6] I. Zbyr, "Contemporary language issues in ukraine: Bilingualism or russification," *Journal of Arts and Humanities*, vol. 4, no. 2, pp. 45–54, 2015.

[7] X. Shi, Q. Feng, and L. Xie, "The asru 2019 mandarin-english code-switching speech recognition challenge: Open datasets, tracks, methods and results," *arXiv preprint arXiv:2007.05916*, 2020.

[8] M. B. Mustafa, M. A. Yusoof, H. K. Khalaf, A. A. Rahman Mahmoud Abushariah, M. L. M. Kiah, H. N. Ting, and S. Muthaiyah, "Code-switching in automatic speech recognition: The issues and future directions," *Applied Sciences*, vol. 12, no. 19, p. 9541, 2022.

[9] S. Sitaram, K. R. Chandu, S. K. Rallabandi, and A. W. Black, "A survey of code-switched speech and language processing," *arXiv preprint arXiv:1904.00784*, 2019.

[10] J. Butzberger, H. Murveit, E. Shriberg, and P. Price, "Spontaneous speech effects in large vocabulary speech recognition applications," in *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.

[11] M. Goto, K. Itou, S. Hayamizu *et al.*, "A real-time filled pause detection system for spontaneous speech recognition." in *Eurospeech*, 1999, pp. 227–230.

[12] E. Shriberg, "To 'errrr'is human: ecology and acoustics of speech disfluencies," *Journal of the international phonetic association*, vol. 31, no. 1, pp. 153–169, 2001.

[13] H. G. S. Moniz, "Processing disfluencies in european portuguese," Ph.D. dissertation, Universidade de Lisboa (Portugal), 2013.

[14] P. A. Heeman and J. F. Allen, "Improving robustness by modeling spontaneous speech events," in *Robustness in language and speech technology*. Springer, 2001, pp. 123–152.

[15] J. Hough and M. Purver, "Processing self-repairs in an incremental type-theoretic dialogue system," in *Proceedings of the 16th SemDial workshop on the semantics and pragmatics of dialogue (SeineDial)*, 2012, pp. 136–144.

[16] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 5, pp. 1526–1540, 2006.

[17] V. V. Pylypenko and O. N. Ladoshko, "Annotatsiya i uchet rechevykh sboyev v zadache avtomaticheskogo raspoznavaniya spontannoy ukrainskoy rechi.translation: Annotation and consideration of speech disruptions in the task of automatic recognition of spontaneous ukrainian speech." *Shtuchnyi intelekt*, 2010.

[18] O. Ladoshko and A. Prodeus, "Markup of spontaneous ukrainian speech," *Electronics and Communications*, vol. 16, no. 1, pp. 97–103, 2011.

[19] N. Verbich, "Pauza v naukovomu movlenni.translation: Pause in scientific speech," *Kultura slova*, no. 79, pp. 192–200, 2013.

[20] T. Schlippe, M. Volovyk, K. Yurchenko, and T. Schultz, "Rapid bootstrapping of a ukrainian large vocabulary continuous speech recognition system," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7329–7333.

[21] M. Sazhok, R. Seliukh, D. Y. Fedoryn, O. Yukhymenko, and V. Robeiko, "Automatic speech recognition for ukrainian broadcast media transcribing," *Control systems & computers*, 2019.

[22] R. Safarik and J. Nouza, "Unified approach to development of asr systems for east slavic languages," in *Statistical Language and Speech Processing: 5th International Conference, SLSP 2017, Le Mans, France, October 23–25, 2017, Proceedings 5*. Springer, 2017, pp. 193–203.

[23] J. Makhoul, F. Kubala, R. Schwartz, R. Weischedel *et al.*, "Performance measures for information extraction," in *Proceedings of DARPA broadcast news workshop*. Herndon, VA, 1999, pp. 249–252.

Who’s Next? The Role of Speech Melody in the Turn-Taking System of Dutch

Ariëlle Reitsema

Leiden University, The Netherlands

Who's Next? The Role of Speech Melody in the Turn-Taking System of Dutch

Ariëlle Reitsema[1]

[1]Leiden University Centre for Linguistics
a.reitsema@hum.leidenuniv.nl

Overview

Turn-taking is a vital part of human conversation, but it is still unclear what role utterance-final speech melody plays in relation to our ability to swiftly and smoothly switch between interlocutors. The PhD-project presented here, which is still at an early stage, sets out to work on this puzzle by first investigating the function of final boundary tones in a corpus of task-oriented, spontaneous Dutch dialogue, and then carrying out follow-up experiments, for example using eye-tracking.

Index Terms: turn-taking, final boundary tones, turn projection, intonation, dialogue, short gaps

1. Background and motivation

1.1. Turn-taking in conversation

Turn-taking is a fundamental and universal [1] aspect of human conversation, where interlocutors typically switch back and forth between speaker and listener roles in a remarkably smooth and accurately timed manner. As is frequently quoted from Sacks and colleagues' seminal paper [2], turn-transitions in conversation are characterised by very little gap and overlap between consecutive speakers, where "[o]verwhelmingly, one party talks at a time" (p. 700). This does not mean that these rules are never violated, but rather that the violations themselves can be communicatively meaningful [3], e.g. a delayed response could communicate reluctance. The speed of human turn-taking is in stark contrast with spoken dialogue systems, which have traditionally relied on silence thresholds to detect turn-ends [4]. Since pauses within speakers' turns are often longer than pauses between turns [see e.g., 1], silence is a rather poor end-of-turn indicator, which advocates for including turn-taking cues in end-of-turn prediction [4].

Successful and swift turn-taking in human conversation relies on a complex interplay of cues, e.g. verbal cues at the syntactic, semantic, as well as pragmatic level, prosodic cues such as pauses, speech melody, loudness and voice quality, but also e.g., breathing, gestures or gaze. The aim of this research is to clarify the precise function and contribution of utterance-final intonation in signalling turn-taking in Dutch, specifically whether – and if so, under which conditions – it can function as an independent cue.

1.2. Utterance-final intonation and the timing problem

The intonation of an utterance consists of the sequence of pitch variations over the course of that utterance. In Dutch, intonation serves primarily to mark information structure: new or important information is highlighted with *pitch accents*, i.e. conspicuous pitch changes that are often peak-shaped, while the beginnings and endings of utterances have *boundary tones*. Final boundary tones can be either low (Transcription of Dutch Intonation (ToDI): L%, characteristic of statements, suggesting finality), high (H%, characteristic of e.g. echo questions), or level (%, signalling incompleteness).

On the one hand, it seems that boundary tones play a role in the turn-taking system of Dutch, especially by signalling turn-keeping. Earlier work on Dutch intonation by Caspers [6], [7] has found support for the idea that the level final boundary tone following a high pitch accent (H* %) is used by speakers to indicate that they want to keep the floor. This turn-keeping cue can be characterised as comma intonation. It has been proposed that the high final boundary tone may similarly be used to signal continuation (in contours such as H*L H% and H* H%) and thus functions as a turn-keeping cue [7], [8], [9], [10]. Since H% also functions as a turn-*yielding* cue when signalling questions (e.g., declarative questions like "You saw him?"), the question of its relevance in the turn-taking system is still open.

On the other hand, it appears that these meanings associated with utterance-final speech melody cannot play a role in helping listeners anticipate the turn-end in order to plan their response, since final boundary tones may be too late in the speech signal. Between-speaker gaps in natural dialogue, for instance in telephone conversations, have been measured to last 0-200 ms on average [1], [11], [12]. This is extremely short, given that laboratory studies have found that it takes about 600-1500 ms to prepare an utterance [13], [14]. To attain such short response latencies, speakers need to plan new utterances in advance, concurrently with the previous speaker's incoming turn. Indeed, Bögels and colleagues, among others, have found support for early planning. They presented participants with questions in which they manipulated the moment at which the gist became clear, which yielded faster responses for those questions where the critical information was given earlier [15].

This raises the question of what role is left for utterance-final intonation, since 200 ms gaps would not be possible if utterance planning would only start at the realisation of final boundary tones, which typically falls on the utterance-final syllable. How can we reconcile these contradicting datapoints? It seems that there are four possible scenarios, that need not be mutually exclusive. First, final boundary tones may not play a role in the turn-taking system, contrary to the findings of e.g. Caspers [7], but in line with e.g. De Ruiter and colleagues [16]. They found that the lexicosyntactic content of an utterance was "necessary (and possibly sufficient) for projecting the moment of its completion," whereas the intonational contour was "perhaps surprisingly … neither necessary nor sufficient for end-of-turn projection" (p. 515). If there is still a role for utterance-final speech melody in turn-taking, it may be limited to cases that are otherwise ambiguous.

Second, rather than aiding the anticipation of a turn-end for the sake of early utterance planning, boundary tones may in fact function as "turn-final go-signals" that help listeners to accurately time their response which they have already planned.

Barthel et al. (2017) support this idea with data from an eye-tracking study where both the presence of early lexical cues and of late intonation cues yielded shorter response times, but the intonation cues did not seem to affect the timing of utterance planning. Then, boundary tones may still be useful to turn-taking simply by marking the location of the turn-end.

Third, there are indications that the reported 200 ms average gap durations paint a skewed picture of the timing of real turn-changes. Corps et al. [17] argue that the short inter-speaker gaps may be "overrated", since the relevant studies did not sufficiently distinguish between real turns (responding to the content of the preceding turn) and other "segments" such as backchannels (e.g. okay, yes, hm-hmm), and parallel talk (continuations of the same speaker's previous turn). Thus, as Heldner and Edlund [11] also argue, a considerable proportion of turn-changes have longer gaps which may allow more space for reaction to prosodic cues such as boundary tones.

A fourth scenario is that boundary tones are not as confined to the final syllable as has often been assumed, such that an impending turn-end would become perceivable earlier on in the speech melody. It is currently unclear which of these interpretations best describe the real functioning of boundary tones, and there is a methodological challenge in reconciling generalisability with the highly varied nature of conversation.

2. Aims, methods and challenges

2.1. Research question and hypotheses

The goal of this PhD-project is to find out of whether utterance-final intonation in Dutch can play an *independent* role in signalling turn-taking in conversation, i.e., whether boundary tones can signal turn-yielding or turn-holding in their own right under some conditions, or whether their role in turn-taking is redundant. To this end at least the following hypotheses will be tested in the corpus study described below, controlling for other expected relevant cues to turn-taking:

H1. There is a strong correlation between the utterance final level boundary tone (%) and turn-keeping

H2. There is no direct relationship between low boundary tones (L%) and turn-keeping or turn-yielding

H3. (a) There is no general relationship between the high boundary tone (H%) and turn-keeping or turn-yielding, but (b) when H% functions as an independent turn-yielding cue, there will be a longer gap (> 200 ms) between consecutive speakers than at other turn-changes.

2.2. Methods and challenges

This research will consist of two complementary methodological phases, starting with a detailed corpus study aimed at the aforementioned hypotheses, which will lead to follow-up experiments To limit the number of possible variables at play, we will investigate only the auditory domain, comparable with telephone calls. The auditory analyses, annotations, and manipulations will be carried out in Praat [18].

A corpus of Dutch-language Map Task dialogues – guided spontaneous conversations – will be annotated at various levels in order to flesh out the role of speech melody in relation to other potential cues to turn-taking. This will allow us to see whether, and when, there is an independent role for intonation. In a Map Task dialogue, participants need to cooperate verbally to reproduce the route given on the one participant's map on the other participant's map, and minor differences between the maps elicit more varied turn-taking behaviour. The eight annotated dialogues analysed by Caspers [7] will be checked and supplemented, and an additional eight dialogues will be added to the corpus and annotated from scratch. The existing annotations comprise inter-pausal units (IPU) – i.e., consecutive stretches of speech by the same speaker, separated by either silent gaps larger than 100 ms, or speaker changes (see also [7], [19]) – as well as transcriptions of utterance-final intonation using ToDI, syntactic completion points, and turn-transition types (change, hold, simultaneous start, interruption, and backchannel). A challenge is how to deal with backchannels and parallel talk in relation to turn-taking.

The main challenge for the corpus study is to find systematic ways to handle the varied and often context-bound nature of spontaneous conversation. I am currently looking into using large language models to perform text classification on the orthographic transcriptions, to label each IPU in terms of its how likely it is followed by either a turn change or hold. Hereby accounting for the possible syntactic, semantic and pragmatic turn-changing cues, the role of intonation should become more clear. Additionally, I am considering analysing the f0-contours in a more data-driven way via cluster analysis [20].

In trying to exclude as many potential other cues to turn-taking as possible, instances where intonation is really the only relevant turn-taking cue may turn out to be sparse due to the limited scale of the corpus. If more corpus data will need to be analysed, there are challenges in streamlining and further automatising the labour-intensive annotation process. Another challenge in the current corpus data is that the recordings were not made with separated speaker channels, limiting possibilities for analysing intonation of overlapping speech.

The precise design of the experiments will depend in part on the findings of the corpus study. Nevertheless, the projected plan is to conduct an eye-tracking and reaction time experiment using programmed avatars and prosodically manipulated speech, inspired by [21] and [22]. In a next-speaker decision task, participants will listen to a dialogue between two avatars on a screen, and click on the speaker they expect to talk next when the audio stops. The timing of gaze-shifts from one to the other avatar can then inform us as to when turn-taking cues become available to listeners, and reaction times can teach us about the prototypicality of certain turn-holding and turn-yielding configurations. The biggest challenge in designing suitable follow-up experiments seems to be ensuring sufficient experimental control, whilst maintaining maximum similarity to natural conversation.

3. Looking ahead

I look forward to being able to present some initial results of the corpus study at the Doctoral Consortium. Although I do not yet have results to present here, I hope to gain a more nuanced and deepened understanding of the possible function of melodic cues in guiding turn-taking in conversation through this PhD-research. Even if utterance-final speech melody only proves to be an independent cue in very limited cases, Skantze [4] argues that it may still prove useful to conversational systems. "Since conversational systems do not have the same computational/cognitive constraints and might not have to prepare the response in advance to the same extent as humans, they could make use of turn-final cues to a larger extent" (p. 10). Uncovering patterns in the use of boundary tones may then be informative for improving the interactive functioning of dialogue systems, i.e., helping robots know who's next.

4. References

[1] T. Stivers *et al.*, 'Universals and Cultural Variation in Turn-Taking in Conversation', *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 26, pp. 10587–10592, 2009.

[2] H. Sacks, E. A. Schegloff, and G. Jefferson, 'A Simplest Systematics for the Organization of Turn-Taking for Conversation', *Language*, vol. 50, no. 4, pp. 696–735, 1974, doi: 10.2307/412243.

[3] J. P. de Ruiter, 'Turn-Taking', in *The Oxford Handbook of Experimental Semantics and Pragmatics*, C. Cummins and N. Katsos, Eds., Oxford University Press, 2019, p. 0. doi: 10.1093/oxfordhb/9780198791768.013.7.

[4] G. Skantze, 'Turn-taking in Conversational Systems and Human-Robot Interaction: A Review', *Comput. Speech Lang.*, vol. 67, p. 101178, May 2021, doi: 10.1016/j.csl.2020.101178.

[5] J. Edlund and M. Heldner, 'Exploring Prosody in Interaction Control', *Phonetica*, vol. 62, no. 2–4, pp. 215–226, Dec. 2005, doi: 10.1159/000090099.

[6] J. Caspers, 'Who's Next? The Melodic Marking of Question vs. Continuation in Dutch', *Lang. Speech*, vol. 41, no. 3, pp. 375–398, Jul. 1998.

[7] J. Caspers, 'Local speech melody as a limiting factor in the turn-taking system in Dutch', *J. Phon.*, vol. 31, no. 2, pp. 251–276, Apr. 2003, doi: 10.1016/S0095-4470(03)00007-X.

[8] J. Caspers, 'Investigating the relationship between high, low and level contour endings and punctuation symbols in Dutch', in *Proceedings of the 16th International Congress of Phonetic Sciences*, B. W. J. Trouvain J, Ed., Saarbruecken, Germany, 2007, pp. 1221–1224. Accessed: Feb. 14, 2025. [Online]. Available: https://hdl.handle.net/1887/14380

[9] J. Caspers, 'Punt, komma of vraagteken? De waarneming van zinsfinale intonatie door eerste- en tweedetaalsprekers van het Nederlands', *Neerlandica Extra Muros*, vol. 45, no. 3, pp. 2–9, 2007.

[10] N. Jansen, '1.7.2.3 Pragmatische functies van intonatie', Algemene Nederlandse Spraakkunst. Accessed: May 27, 2025. [Online]. Available: https://e-ans.ivdnt.org/topics/pid/topic-16050028826383246

[11] M. Heldner and J. Edlund, 'Pauses, gaps and overlaps in conversations', *J. Phon.*, vol. 38, no. 4, pp. 555–568, Oct. 2010, doi: 10.1016/j.wocn.2010.08.002.

[12] S. G. Roberts, F. Torreira, and S. C. Levinson, 'The effects of processing and sequence organization on the timing of turn taking: a corpus study', *Front. Psychol.*, vol. 6, May 2015, doi: 10.3389/fpsyg.2015.00509.

[13] Z. M. Griffin and K. Bock, 'What the Eyes Say about Speaking', *Psychol. Sci.*, vol. 11, no. 4, pp. 274–279, 2000.

[14] P. Indefrey and W. J. M. Levelt, 'The spatial and temporal signatures of word production components', *Cognition*, vol. 92, no. 1–2, pp. 101–144, May 2004, doi: 10.1016/j.cognition.2002.06.001.

[15] S. Bögels, L. Magyari, and S. C. Levinson, 'Neural signatures of response planning occur midway through an incoming question in conversation', *Sci. Rep.*, vol. 5, no. 1, p. 12881, Aug. 2015, doi: 10.1038/srep12881.

[16] J. P. D. Ruiter, H. Mitterer, and N. J. Enfield, 'Projecting the end of a Speaker's Turn: A Cognitive Cornerstone of Conversation', *Language*, vol. 82, no. 3, pp. 515–535, 2006.

[17] R. E. Corps, B. Knudsen, and A. S. Meyer, 'Overrated gaps: Inter-speaker gaps provide limited information about the timing of turns in conversation', *Cognition*, vol. 223, p. 105037, Jun. 2022, doi: 10.1016/j.cognition.2022.105037.

[18] Boersma, Paul & Weenink, David (2024). Praat: doing phonetics by computer [Computer program]. Version 6.4.04, January 6, 2024, http://www.praat.org/.

[19] H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den, 'An Analysis of Turn-Taking and Backchannels Based on Prosodic and Syntactic Features in Japanese Map Task Dialogs', *Lang. Speech*, vol. 41, no. 3–4, pp. 295–321, Jul. 1998, doi: 10.1177/002383099804100404.

[20] C. Kaland, 'Contour clustering: A field-data-driven approach for documenting and analysing prototypical f0 contours', *J. Int. Phon. Assoc.*, vol. 53, no. 1, pp. 159–188, Apr. 2021, doi: 10.1017/S0025100321000049.

[21] M. Rossi, K. Feindt, and M. Zellers, 'Perception of F0 movements towards potential turn boundaries in German and Swedish conversation: background and methods for an eye-tracking study', in *Proceedings of FONETIK 2022*, Stockholm, Sweden, 2022. Accessed: Jul. 07, 2025. [Online]. Available: https://www.researchgate.net/publication/368882042_Perception_of_F0_movements_towards_potential_turn_boundaries_in_German_and_Swedish_conversation_background_and_methods_for_an_eye-tracking_study

[22] K. Feindt, M. Rossi, G. Esfandiari-Baiat, A. G. Ekström, and M. Zellers, 'Cues to next-speaker projection in conversational Swedish: Evidence from reaction times', in *INTERSPEECH 2023*, ISCA, Aug. 2023, pp. 1040–1044. doi: 10.21437/Interspeech.2023-778.

Sociolinguistic Variation in Mandarin: Native and L2 Perspectives on Neutral Tone and Rhotacization

Xiao Dong

Indiana University Bloomington, USA

Sociolinguistic Variation in Mandarin: Native and L2 Perspectives on Neutral Tone and Rhotacization

Xiao Dong[1]

[1]Department of East Asian Languages and Cultures, Indiana University Bloomington, USA

dong1@iu.edu

1. Research Questions and Motivation

This dissertation investigates how two prosodic features in Mandarin Chinese—neutral tone and rhotacization—are produced and perceived by native speakers and second language (L2) learners. Both features are prominent in Beijing Mandarin but are less commonly used in Taiwan Mandarin. This study addresses four main questions: (1) How do native speakers from Beijing use neutral tone in production? (2) How are neutral tone and rhotacization perceived across different listener groups, particularly among Beijing and Taiwan speakers? (3) How do L2 learners produce and interpret these socially meaningful features? (4) How do factors such as language awareness, attitude, and exposure influence learners' acquisition of sociolinguistic variation?

This project is motivated by three main gaps in the current literature. First, sociolinguistic study of variation and change has a long-standing bias towards speech communities in Western and especially Anglophone societies, leaving variation in non-Western societies and languages such as Mandarin relatively underexplored [1, 2, 3, 4]. Second, while previous studies have examined how various social factors influence the use of specific variables in Mandarin, such as neutral tone (NT) [5, 6, 7, 8, 9, 10, 11], the social meanings of these variables, especially in cross-regional contexts, remain largely unexamined. Third, L2 learners' ability to acquire and interpret socially meaningful variation remains poorly understood, even as sociolinguistic competence becomes increasingly central to second language pedagogy and research [12, 13, 14, 15, 16]. Research on L2 learners' sociolinguistic competence in Mandarin Chinese remains especially limited.

2. Methodology

The dissertation comprises three studies. The first examines production patterns of neutral tone among 36 native Beijing speakers (20 female, aged 25–45), using sociolinguistic interviews that included conversational speech, a word list, and a passage-reading task. It is important to note that Mandarin neutral tone words fall into three categories: obligatory, optional, and forbidden. The word list included 48 optional and 44 forbidden neutral tone words, while the passage reading task contained 14 optional and 27 forbidden items. The conversational section included background questions, daily life topics, and 15 questions specifically designed to measure participants' orientation towards Beijing. This study provides empirical evidence of ongoing change in the use of neutral tone in contemporary Beijing and explores how social factors such as speakers' place orientation and speech style shape linguistic behavior.

The second study uses a web-based matched-guise perception experiment involving 96 listeners (40 from Beijing, 56 from Taiwan). Four talkers (1 male 1 female from both northern and southern China) recorded two paragraphs (one targeting neutral tone, the other rhotacization) in three guises: natural speech and two manipulated versions with different realizations of the target variables. Target words were extracted and resynthesized into natural speech using Praat [17], resulting in 16 stimuli. Listeners rated each recording on traits such as femininity, likability, intelligence, and personality, and completed a background questionnaire. This experiment investigates the social meanings indexed by neutral tone and rhotacization, and examines how speaker accent, speaker gender, and listener background shape these perceptions.

The third study, currently being launched, will investigate advanced L2 learners of Mandarin through both production and perception tasks. Participants will complete sociolinguistic interviews, matched-guise perception tasks, and background questionnaires assessing language attitudes, awareness, and exposure. This component will allow me to evaluate how learners acquire not only the linguistic forms, but also their social meanings, providing insights into how learners integrate into new linguistic communities and construct identity through language use.

3. Results So Far

In experiement 1, two native Mandarin-speaking coders labeled neutral tone usage as 0 (neutral tone) and 1 (full tone) across speech types. To assess speakers' place orientation, the same two raters scored their responses to 15 target questions using a 5-point scale (1 = lowest, 5 = highest). Each speaker's final place orientation score was calculated as the average of both raters' scores across all 15 items. Scores ranged from 2.60 to 4.46, with a median of 3.27.

Linear mixed-effects modeling revealed several significant effects. First, speech style has a significant effect: native Beijingers used significantly less neutral tone in passage reading and word list reading than in free or conversational speech ($p < 0.001$). Second, neutral tone type also played a signficant role: neutral tone was realized significantly less often in forbidden neutral tone words compared to optional ones ($p < 0.001$). Obligatory items were excluded from the analysis, as they are expected to be neutralized in nearly all cases. Third, a significant effect of place orientation was observed ($p = 0.017$): speakers with a stronger orientation toward Beijing produced more neutral tone overall. Compared to [10], who reported nearly categorical use of neutral tone, the current study found decreased usage—averaging 70.2% for women and 72.8% for men—indicating a potential sound change in progress. Interestingly, gender had no significant effect on neutral tone usage, contrasting with patterns often found in English sociolinguis-

tic research, where women often lead linguistic change (e.g., [18, 19, 20]).

Preliminary results from Experiment 2 reveal regional differences in NT perception and complex interactions among speaker gender, speaker accent, listener gender, and NT realization. Across all listeners, female talkers were rated as more feminine, gentle, and likable. Among Beijing listeners, non-neutral tone speech was perceived as more accented ($p < 0.001$), and NT use positively influenced perceptions of cuteness and reduced accentedness, especially for northern-accented talkers. In contrast, Taiwan listeners associated non-NT speech with higher intelligence ($p < 0.05$) and marginally higher likability ($p = 0.055$), and rated southern-accented NT as more accented ($p < 0.01$). Male Taiwan listeners showed stronger preferences for non-NT speech, while female listeners did not. Analyses of NT-related judgments such as perceived education and regional origin are ongoing, and analysis of rhotacization is forthcoming.

4. Plans for the Future

Ongoing analysis of the perception data will examine how listeners infer speakers' region, education level, occupation, and other social characteristics based on variation in neutral tone and how they perceive rhotacization. These analyses will deepen our understanding of the social meanings of these variables in Mandarin and how such meanings vary by region and interact with other factors such as listener gender and speaker accent.

The third experiment has recently been finalized and is ready for data collection. We aim to recruit approximately 30 advanced-level L2 learners of Mandarin. Data collection is planned for completion in October, and analysis will begin shortly thereafter.

5. Main Challenges

This research presents both methodological and theoretical challenges. Methodologically, participant recruitment has proven difficult, with L2 participants posing an even greater challenge. A small L2 sample could limit the statistical power of the third experiment. Additionally, it is important to ensure that learner participants have acquired the target features; otherwise, their production data may lack meaningful variation for analysis. Ensuring the familiarity of word list and passage items is another challenge in designing production tasks for L2 speakers.

Theoretically, this project addresses sociolinguistic variation in Mandarin Chinese, a relatively underdeveloped area even among native speakers, offering limited prior research for reference. Moreover, by examining how L2 learners navigate socially meaningful variation, the study challenges traditional models of L2 phonological acquisition that prioritize grammatical accuracy over socially appropriate usage.

6. Main Contributions

This dissertation makes several important contributions. It offers one of the first comprehensive sociophonetic accounts of neutral tone and rhotacization in Mandarin across regional and learner groups, documenting both synchronic variation and possible change in progress. It demonstrates how these features function as socially meaningful variables, with distinct indexical values across listener groups. It extends sociolinguistic inquiry beyond English and Indo-European languages, contributing to a more typologically diverse understanding of language variation and perception. It advances research in L2 sociolinguistics by examining how learners acquire and evaluate variation in a tonal language, including the role of study abroad in shaping sociolinguistic competence. It also provides a replicable methodological model—combining sociolinguistic interviews and perception experiments—for future research on language variation and social meaning across speaker/listener groups.

7. References

[1] N. Nagy and M. Meyerhoff, "Introduction: Social lives in language," in *Social lives in language–Sociolinguistics and multilingual speech communities: Celebrating the work of Gillian Sankoff*. John Benjamins Publishing Company, 2008, pp. 1–16.

[2] D. Smakman, "The westernising mechanisms in sociolinguistics," in *Globalising sociolinguistics*. Routledge, 2015, pp. 16–35.

[3] J. N. Stanford, "A call for more diverse sources of data: Variationist approaches in non-english contexts," *Journal of Sociolinguistics*, vol. 20, no. 4, pp. 525–541, 2016.

[4] A. Adli and G. R. Guy, "Globalising the study of language variation and change: A manifesto on cross-cultural sociolinguistics," *Language and Linguistics Compass*, vol. 16, no. 5-6, p. e12452, 2022.

[5] X. Dong, F. Liu, M. Nesbitt, and C.-J. C. Lin, "Place orientation and language practice: An update on the use of neutral tone among beijing professionals," *U. Penn Working Papers in Linguistics*, vol. 30, no. 2, 2024.

[6] F. Gao, "Full tone to sound feminine: Analyzing the role of tonal variants in identity construction," *U. Penn Working Papers in Linguistics*, 2020.

[7] M. Hu, "Beijinghua chutan [a preliminary study of the peking dialect]," 1987.

[8] H. Zhao, "Language variation and social identity in beijing," Ph.D. dissertation, Queen Mary University of London, 2018.

[9] S. Jin, *Xian dai Han yu qing sheng dong tai yan jiu*. Beijing: Ethnic Publishing House, 2002.

[10] Q. Zhang, "A chinese yuppie in beijing: Phonological variation and the construction of a new professional identity," *Language in society*, vol. 34, no. 3, pp. 431–466, 2005.

[11] C. Zhou, "Beijinghua qingsheng, erhua, qingruzi de bianyi yanjiu [the variance research of un-nuclear tone, erhua and qingruzi in beijing speech]," Ph.D. dissertation, PhD thesis, Beijing Language and Culture University, Beijing, 2006.

[12] H. D. Adamson and V. M. Regan, "The acquisition of community speech norms by asian immigrants learning english as a second language: A preliminary study," *Studies in Second Language Acquisition*, vol. 13, no. 1, pp. 1–22, 1991.

[13] R. Bayley and D. R. Preston, *Second language acquisition and linguistic variation*. John Benjamins Publishing, 1996, vol. 10.

[14] K. L. Geeslin, "A comparison of copula choice: Native spanish speakers and advanced learners," *Language Learning*, vol. 53, no. 4, pp. 703–764, 2003.

[15] K. L. Geeslin, L. García-Amaya, M. Hasler-Barker, N. Henriksen, and J. Killam, "The sla of direct object pronouns in a study abroad immersion environment where use is variable," in *Selected proceedings of the 12th Hispanic linguistics symposium*. Cascadilla Proceedings Project Somerville, MA, 2010, pp. 246–259.

[16] K. L. Geeslin and A. Gudmestad, "The acquisition of variation in second-language spanish," *Journal of Applied Linguistics*, vol. 5, no. 2, pp. 137–157, 2015.

[17] Y. Jadoul, B. Thompson, and B. De Boer, "Introducing parselmouth: A python interface to praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[18] W. Labov, "The social motivation of a sound change," *Word*, vol. 19, no. 3, pp. 273–309, 1963.

[19] P. Trudgill, "Sex, covert prestige and linguistic change in the urban british english of norwich," *Language in society*, vol. 1, no. 2, pp. 179–195, 1972.

[20] L. Milroy and S. Margrain, "Vernacular language loyalty and social network1," *Language in Society*, vol. 9, no. 1, pp. 43–70, 1980.

Minimizing Human Effort in Adaptive Synthetic Speech Quality Assessment via Active Learning

Natacha Miniconi

Le Mans University, France

Minimizing Human Effort in Adaptive Synthetic Speech Quality Assessment via Active Learning

Natacha Miniconi[1]

[1]LIUM, University of Le Mans, France

natacha.miniconi@univ-lemans.fr

Abstract

Synthetic speech has advanced significantly in recent years. However, its quality assessment still relies heavily on subjective human evaluation, which is costly, scarce, and often biased. This thesis aims to address these challenges to minimize human intervention while improving automatic quality prediction of synthetic speech across languages and domains. Specifically, we investigate: (i) methods to minimize human intervention and enhance the automatic prediction of speech quality using active learning, (ii) how to automate the continuous improvement of performance and generalization in quality predictors across diverse languages and domains, (iii) ways to model individual listener perception of quality, and (iv) the identification of synthetic speech issues revealed by challenging text characteristics. Through these research directions, we aim to build more efficient, generalizable, and scalable approaches for synthetic speech quality evaluation.

Index Terms: Synthesis speech quality assessment, MOS prediction, Active learning

1. Motivations & Research Questions

Synthetic speech has made significant progress in recent years. As these systems evolve, the need for its assessment is present. Traditionally, the evaluation of synthetic speech relies on human perception, most commonly through Mean Opinion Score (MOS) ratings. While MOS is widely used due to its simplicity and interpretability, it comes with several limitations and controverses [1, 2, 3, 4]. Human annotations are inherently subjective, prone to biases, and highly variable depending on the context, listener demographics, and experimental setup. Moreover, the process of collecting MOS ratings is time-consuming and expensive, which makes it difficult to scale, especially when dealing with multiple languages, domains, or synthesis methods. To address these issues, automatic quality prediction models have been proposed in the literature [5, 6, 7]. The need to evaluate synthetic speech has sparked interest in developing automatic quality prediction models. One recent interest is the VoiceMOS Challenge [8, 9], which aims to promote the development of models capable of generalising quality prediction across different languages, speakers and synthesis systems. Nevertheless, building effective predictors remains a difficult task. The models often require large-scale annotated datasets to reach competitive performance levels. This dependency on costly human-labeled data poses significant challenges, especially when scaling to multiple languages or new application domains. These systems aim to estimate perceived synthetic speech quality with minimal human intervention. As a result, the generalization and practical deployment of these models are still constrained by the limited availability of reliable annotations. In this context, active learning emerges as a promising solution [10] in other fields. By intelligently selecting the most informative samples for annotation, active learning aims to reduce the amount of labeled data needed while maintaining or even improving model performance [11]. Applied to the domain of synthetic speech quality assessment, active learning could help alleviate the dependence on costly human evaluations, facilitate the creation of diverse and effective datasets, and ultimately enhance the generalization capabilities of quality prediction models across languages and domains. This thesis addresses these challenges through the following research questions:

RQ1 : how can human intervention be minimized to improve the automatic prediction of synthetic speech quality across diverse contexts (languages, domains), by using an active learning approach ?

RQ2 : how can we design an automated loop for synthetic speech evaluation and dataset enrichment that enhances predictor generalization?

RQ3 : how can we model individual differences in perceived quality of synthetic speech?

RQ4 : how can a quality predictor be used to systematically identify the textual features that induce failure modes in Text-to-Speech systems (TTS)?

2. Experiments Conducted

To address **RQ1**, we conducted a first study aiming to improve the generalization of a synthetic speech quality predictor using active learning. In this context, active learning refers to the iterative training paradigm where a model selects the most informative samples to be annotated, based on selection criteria based on **uncertainty** and **diversity**. This process is designed to reduce labeling costs while maximizing model performance. In our incremental active learning framework, we start with a pretrained baseline model and, at each iteration, select new samples from the unlabeled pool using uncertainty-based and diversity-based strategies. The selected samples are then annotated and added to the training data, progressively refining the model over multiple iterations. In this first study, active learning is used to generalize the automatic predictor across two distinct scenarios: (i) cross-domain adaptation, generalizing the model to a different domain than the one it was trained on, using the SOMOS corpus [12]; and (ii) cross-lingual adaptation, transferring the model to a different language (French), using the Blizzard Challenge 2023 (BC 2023) corpus[13]. Here, cross-lingual refers to the challenge of generalizing model performance from one language (e.g., English) to another (e.g., French). For this purpose, we employed one of the well-established baselines from the literature: SSL-MOS [14], which combines Wav2Vec 2.0 (W2V2) [15] with an additional hidden layer to predict the MOS. We evaluated using utterance-level Spearman's Rank Correlation

Coefficient (SRCC), widely used for MOS prediction [8]. Our

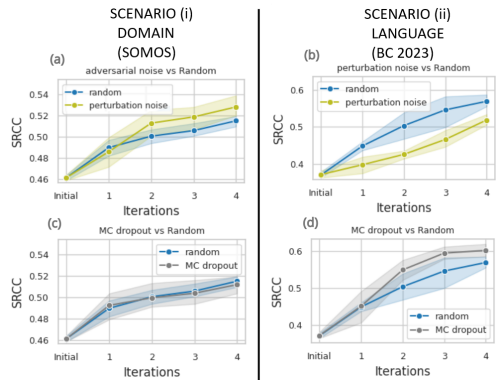


Figure 1: *Evaluation of uncertainty strategies in active learning with SRCC metric in utterance-level (y-axis) for each iteration (x-axis).*

observations shown in Figure 1 revealed contrasting behaviors depending on the task and the selected strategy. Here we present the strategies that yielded interesting results: perturbation noise, which measures uncertainty by adding white noise to the audio to see if it confuses the model, and MC dropout, which estimates uncertainty by performing multiple stochastic forward passes. For cross-domain adaptation (i), the perturbation noise strategy, an uncertainty-based strategy, proved to be the most effective. From the second iteration onward, this approach outperformed random selection in terms of SRCC at the utterance level. In contrast, the MC Dropout strategy, an uncertainty-based approach, did not yield significant improvements in this scenario. Conversely, in the cross-lingual adaptation (ii), the MC Dropout strategy was more successful, achieving better SRCC performance than random selection. In contrast, the perturbation noise approach failed to capture meaningful uncertainty in this setting, resulting in performance degradation. In order to refine the active learning strategies, we analyze the distribution of the selected data from active learning process for a good performance of the predictor with two criteria : MOS and duration. It was observed that perturbation noise selects longer, high-MOS samples, boosting cross-domain adaptation. MC Dropout keeps a balanced selection, helping cross-lingual transfer. The analysis of the distribution of the selected data from active learning process could also give the information to enrich the dataset.

3. Current Insights and Challenges

While our work on **RQ1** focuses on minimizing human intervention through active learning strategies to select the most informative samples for annotation, it still reveals several challenges. Notably, even with active learning, human annotation remains a bottleneck, motivating further automation. This becomes particularly challenging in scenarios where a new synthesis system is developed for a low-resource language, and no prior annotations are available. As a result, it motivates the development of a more autonomous quality evaluation system **RQ2**. Figure 2 presents an overview of this proposed autonomous learning loop for synthetic speech quality prediction. This next research question explores how to fully automate the loop of synthetic speech evaluation and dataset enrichment, enabling continuous improvement of the model without relying on manual labeling. In this direction, we examined whether alternative metrics (Figure 2 (1.2)), such as the bona fide probability from a DeepFake detection model [16], could serve as a proxy for perceptual quality, offering an alternative to MOS. Prior work in fake audio detection [17] shows that MOS scores can help differentiate real from fake speech. Consistently, we observed a correlation between MOS and bona fide probability, particularly in the French corpus (BC 2023). This insight enables automatic sample selection and evaluation, fostering continuous model improvement.

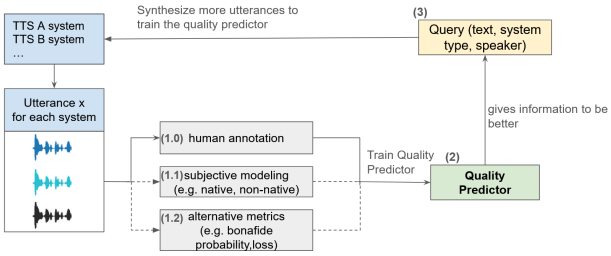


Figure 2: *Overview of the automated improvement process for the speech quality predictor in **RQ2**. A fixed pool of TTS systems generates utterances from selected (text, system, speaker) triplets chosen via active learning (3) to maximize predictor training efficiency (2) using alternative metrics (1.2).*

4. Future Work Plan

The next phase of our work focuses on automating the continuous improvement of synthetic speech quality prediction **RQ2**. This objective builds upon our earlier findings and aims to move from human-supervised (Figure 2 (1.0)) active learning toward a fully self-sustaining evaluation loop (Figure 2 (1.2)). To achieve this, we will first continue developing and refining active learning strategies capable of identifying which types of samples are most beneficial for enriching the training dataset. These strategies will generate structured queries that specify target characteristics such as *text content*, *TTS system type*, and *speaker identity* to guide TTS systems (Figure 2 (3)) in synthesizing informative samples that are added to the training dataset. Evaluation of these strategies will consider not only their impact on model performance but also the nature of the selected data (e.g., duration, linguistic diversity). In parallel, enabling automation requires a way to annotate new data without human intervention. For this, we will explore the use of alternative metrics (Figure 2 (1.2)) such as prediction loss as an indicator of perceptual quality, which could replace or reduce the need for manual MOS labels. Furthermore, we could study phonetic measures that might be potentially related to speech naturalness.

We aim to extend the quality prediction system to model individual listener perception (Figure 2 (1.1)). By capturing how specific raters evaluate synthetic speech, we hope to develop more personalized and perceptually aligned quality predictors (e.g. native vs non-native). This directly relates to **RQ3**, which investigates how individual differences in quality perception can be modeled using information derived from human annotations. Such personalized models may help us better understand subjectivity in speech evaluation and improve prediction alignment with listener expectations.

In addition, **RQ4** focuses on using our generalized quality predictor as a tool to diagnose weaknesses in TTS systems. By analyzing which texts consistently receive low predicted quality scores, we aim to identify the characteristics of text that are most challenging for synthesis. This information could then be used to guide targeted training or adaptation of TTS systems.

5. Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101007666. This work was granted access to the HPC resources of IDRIS under the allocation 2025-AD011014876R2 made by GENCI.

6. References

[1] E. Cooper, S. L. Maguer, E. Klabbers, and J. Yamagishi, "Good practices for evaluation of synthesized speech," *arXiv preprint arXiv:2503.03250*, 2025.

[2] A. Kirkland, S. Mehta, H. Lameris, G. E. Henter, E. Szekely, and J. Gustafson, "Stuck in the mos pit: A critical analysis of mos test methodology in tts evaluation," in *12th ISCA Speech Synthesis Workshop (SSW2023)*, 2023, pp. 41–47.

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

[4] J. Camp, T. Kenter, L. Finkelstein, and R. Clark, "Mos vs. ab: Evaluating text-to-speech systems reliably using clustered standard errors," 08 2023, pp. 1090–1094.

[5] Z. Qi, X. Hu, W. Zhou, S. Li, H. Wu, J. Lu, and X. Xu, "LE-SSL-MOS: Self-Supervised Learning MOS Prediction with Listener Enhancement," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–6.

[6] H. Wang, S. Zhao, X. Zheng, and Y. Qin, "RAMP: Retrieval-Augmented MOS Prediction via Confidence-based Dynamic Weighting," in *Interspeech*, 2023, p. 1095–1099.

[7] W.-C. Huang, E. Cooper, and T. Toda, "MOS-Bench: Benchmarking Generalization Abilities of Subjective Speech Quality Assessment Models," *arXiv preprint arXiv:2411.03715*, 2024.

[8] E. Cooper, W.-C. Huang, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "The Voicemos Challenge 2023 : Zero-Shot Subjective Speech Quality Prediction for Multiple Domains," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–7.

[9] W.-C. Huang, E. Cooper, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "The VoiceMOS Challenge 2022," in *Interspeech*, 2022, pp. 4536–4540.

[10] E. Surzhikova and J. Proppe, "regAL: Python Package for Active Learning of Regression Problems," 2024.

[11] D. Li, Z. Wang, Y. Chen, R. Jiang, W. Ding, and M. Okumura, "A Survey on Deep Active Learning: Recent Advances and New Frontiers," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–21, 2024.

[12] G. Maniati, A. Vioni, N. Ellinas, K. Nikitaras, K. Klapsas, J. S. Sung, G. Jho, A. Chalamandaris, and P. Tsiakoulis, "SOMOS: The Samsung Open MOS Dataset for the Evaluation of Neural Text-to-Speech Synthesis," in *Interspeech*, 2022, pp. 2388–239.

[13] O. Perrotin, B. Stephenson, S. Gerber, and G. Bailly, "The Blizzard Challenge 2023," in *18th Blizzard Challenge Workshop*, 2023, pp. 1–27.

[14] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization Ability of MOS Prediction Networks," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8442–8446, 2021.

[15] A. Baevski, Y. 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*, vol. 33, 2020, pp. 12 449–12 460.

[16] J. weon Jung, H.-S. Heo, H. Tak, H. jin Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," 2021.

[17] W. Zhou, Z. Yang, C. Chu, S. Li, R. Dabre, Y. Zhao, and K. Tatsuya, "Mos-fad: Improving fake audio detection via automatic mean opinion score prediction," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 876–880.

Enhancing Ultrasound Tongue Imaging Based Articulation-to-Speech Synthesis

Ibrahim Ibrahimov

Budapest University of Technology and Economics, Hungary

Enhancing Ultrasound Tongue Imaging Based Articulation-to-Speech Synthesis

Ibrahim Ibrahimov[1]

[1]Department of Telecommunications and Artificial Intelligence, Budapest University of Technology and Economics, Hungary

ibrahim@tmit.bme.hu

Abstract

Within the research scope of Silent Speech Interfaces, a process known as articulation-to-speech synthesis targets the conversion of movements of articulatory parts (e.g., tongue) that are actively utilized in speech production into audible speech. To capture tongue motion during speech, utilizing ultrasound tongue imaging technique became popular in research due to its non-invasive and cost-efficient structure. However, this method has its own limitations (e.g., data scarcity, noisy image results, speaker and session dependency) which hinder the development of the synthesis process. In my doctorate research, I aim to propose innovative solutions towards more robust, session-independent, direct ultrasound-to-speech synthesis system. To make this goal possible, I base my works on three research directions: eliminating session dependency, creating a universal (speaker and language independent) process, improving utilized neural network techniques. In this paper, I present the recent works and their results followed by observed limitations and future directions.

Index Terms: articulation-to-speech synthesis, ultrasound tongue imaging

1. Introduction and Motivation

The tongue plays a central role in speech production, contributing significantly to the articulation of both consonants and vowels [1]. Ultrasound tongue imaging (UTI) has become a valuable tool in speech research, offering real-time, high-resolution, and ultra-fast imaging of tongue motion in two dimensions, with ongoing developments towards three-dimensional capabilities [2, 3]. As a data acquisition technique, UTI produces sequences of 2D images, commonly referred to as ultrasound tongue image frames (UTIF), which capture the dynamic articulatory movements during speech.

Articulation-to-speech synthesis (ATS) using UTIF has been researched by employing mainly two different approaches. The first involves converting UTIF data into text, followed by applying a text-to-speech (TTS) system to generate the final acoustic output [4, 5]. While conceptually straightforward, this pipeline introduces significant latency and computational demands, making it less suitable for real-time applications. On the other hand, so-called ”direct synthesis” approach, relies on the fundamental principle that the movement of tongue during speech production is intricately linked to the acoustic speech signal [6]. In this approach, UTIF features are directly mapped to an intermediate representation of acoustic features (e.g., mel-spectrogram), which are then used by a vocoder to synthesize speech [7, 8]. In the literature, the term articulatory-to-acoustic mapping (AAM) is sometimes used to describe the entire ATS process. However, in the context of this research, AAM specifically refers to the component of the direct synthesis approach that performs the mapping from articulatory input (UTIF) to acoustic features, prior to the vocoding stage.

Despite the progress made in utilizing UTIF for ATS using ”direct synthesis” approach, several limitations persist that hinder its broader application and performance [9]. UTI is a speaker-dependent technique because of the diversity in tongue shapes, sizes, and anatomical variations among individuals and this poses a challenge for creating a generalized ultrasound-to-speech (UTS) system. Additionally, session dependency is a significant barrier. Even for the same speaker, variations in transducer placement, head posture, and probe pressure causes trouble to keep the consistency in limited open-access dataset. Current AAM strategies predominantly rely on convolutional layers (e.g. 2D-CNN [8, 10], 3D-CNN [11, 12]). While effective to a degree, these architectures have limited capacity to capture the complex spatiotemporal dependencies in tongue motion.

In response to these limitations, my doctoral research aims to develop a more robust and practical direct ultrasound-to-speech synthesis system that can function reliably across different recording sessions, speakers, and languages. The ultimate goal is to enable a real-time, speaker-independent, and language-flexible speech synthesis system that could, in the long term, support assistive communication technologies for individuals with speech impairments. To achieve this, my work is structured around three key research directions: (1) eliminating session dependency to ensure consistent performance across different data acquisition scenarios; (2) designing universal models that generalize across speakers and languages; and (3) improving neural network architectures and training strategies for more accurate and efficient AAM.

2. Contributions and Results

Towards the goal of my research, we adopt the ”direct synthesis” approach, wherein the original UTIF data - initially at a resolution of 64x842 and subsequently downsampled via bicubic interpolation to 64x128 - is mapped to an 80-dimensional mel-spectrogram representation of the corresponding audio signal. As of now, speaker-specific models are trained to generate mel-spectrograms, which then are fed to neural vocoder to synthesize speech as the output of UTS systems. In this section, brief report is provided about our research until now followed by the obtained results.

2.1. Data augmentation methods on UTIF for UTS

Knowing the fact that in openly accessible UTIF datasets (UltraSuite-TaL80 [13]), each speaker has very small amount of recordings (150-240 utterances), data scarcity becomes a

drawback for training a CNN-based neural network for AAM. To address this issue, data augmentation techniques are often utilized to increase the quantity of training data available. We have developed six distinct augmentation methods tailed for UTIF. After training speaker-specific 2D-CNN models for generating mel-spectrograms, WaveGlow neural vocoder was utilized for synthesizing speech [14]. The efficacy of each approach for UTS systems was evaluated by measuring Mean-Squared Error (MSE) between generated and original mel-spectrograms as well as Mel-Cepstral Distortion (MCD) between synthesized and original speech. According to obtained results of the objective metrics, employing a specific data augmentation technique can effectively improve the generalization of a 2D-CNN-based AAM model.

2.2. Dynamic Time Warping towards cross-speaker UTS

We explored the integration of Dynamic Time Warping (DTW) alignment in the context of cross-speaker UTS towards eliminating speaker dependency. The DTW is performed on the speech signals, which is in synchrony with UTIF. The alignment of UTIF is done based on the calculated DTW distance. We tested cross-speaker UTS systems for four subjects from the UltraSuite-TaL dataset. Using aligned ultrasound data, we did ATS experiments consisting of 2D-CNN based AAM and WaveGlow vocoder with each speaker pair, and compared the MSE and MCD metrics. We found that for two speakers, the DTW-aligned input ultrasound images resulted in lower errors than using the original ultrasound data of the same speaker, indicating that anatomical differences or speckle noise in ultrasound images are an important factor. Overall, the results underlined the potential of DTW as a valuable tool in enhancing UTS towards cross-speaker application.

2.3. Investigation of ”known” limitations of UTS system

The limitations of UTI technique towards UTS has been shown in context of previous works in the field, however to clearly state problems, we have tried to answer below research questions in our previous works.

We know that UTI provides noisy UTIF but what is the best possible representation of tongue shape that can be captured using ultrasound imaging? - For answering this, the cartesian coordinate results of DeepLabCut pose estimation software [15, 16] on original UTIF were used to regenerate tongue images. These regenerated images were also utilized for developing UTS systems. The objective metrics (MSE, MCD) were utilized to compare regenerated and original UTIF performance in UTS system using 2D-CNN based AAM. According to the obtained results, making less noisy UTIF by regenerating it reduced the captured differences among the frames which caused drawback in the AAM stage of UTS which ended to generate less-detailed mel-spectrograms.

Addressing the open question of language dependency in UTS systems, we investigated how the performance of UTS system gets affected by utilizing different language than the language on which the models are trained. A dedicated bilingual (Azerbaijani and English) ultrasound tongue image and audio dataset was recorded to facilitate this analysis. A 2D-CNN AAM and a HiFi-GAN vocoder were employed for developing speaker-specific UTS systems [17]. Based on the objective evaluation using MSE and MCD metrics, we found that our quite standard UTS system is language-dependent both on AAM and vocoding stages. This underscores the necessity of considering linguistic context and individual pronunciation

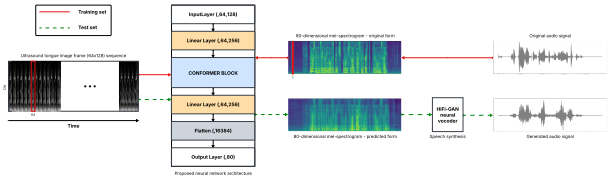


Figure 1: *Schematic diagram of the ultrasound-to-speech system pipeline. Training set with development set flows through the solid arrow, while test set flows through the dashed arrow after the training phase was done.*

patterns in the development of UTS systems

To find an answer to the assumption in the field of UTS about silent articulation is the same as normal articulation, in our previous study, we investigated how silent and whispered speech deteriorates the quality of the synthesized speech samples. For this aim, we trained speaker-dependent 2D-CNN based AAM on UTIF, which were recorded during normal (audible) speech, then we synthesized speech samples from recordings of normal, hyper-articulated and normal articulated silent speaking styles. We found that synthesized speech using HiFi-GAN vocoder for these silent ”speaking” styles is significantly less intelligible than for audible speech, suggesting a difference in the articulatory movements, which should be considered when training silent speech restoration models.

2.4. Conformer-based UTS system

In our research, for the first time, we implemented convolution-augmented transformer (i.e. conformer [18]) block for the UTIF-to-mel mapping step as a part of the ultrasound-to-speech conversion task (please see Figure 1 for full pipeline). From a performance perspective the conformer model has been shown to be parameter-efficient, while effectively learning both local and global dependencies in sequential data like UTIF. In this work, we also explored a more complex architecture by adding bidirectional LSTMs (bi-LSTM) into our ‘Conformer base’ structure. For both UTI-to-mel mapping approaches, we generated the synthesized speech samples using the HiFi-GAN neural vocoder and compared the results to 2D-CNN based AAM. MUSHRA listening test revealed that Conformer with bi-LSTM provided better perceptual quality, while Conformer Base matched the performance of the baseline along with a three times faster training time due to its simpler architecture. These findings suggest that Conformer-based models, especially the Conformer with bi-LSTM, offer a promising alternative to CNNs for ultrasound-to-speech conversion.

3. Future work

Future work will focus on several key directions to further improve the robustness, generalizability, and practicality of UTS systems. First, we plan to investigate the use of diffusion models for the AAM component. Addressing session dependency remains a critical challenge; to this end, we aim to develop strategies that enhance system stability across recording sessions without the need for extensive retraining. In parallel, we will evaluate the effectiveness of various neural vocoders to identify those best suited for UTS applications. Furthermore, we aim to remove intermediate levels and predefined acoustic targets from the UTS system.

4. References

[1] M. Stone, "A guide to analysing tongue motion from ultrasound images," *Clinical linguistics & phonetics*, vol. 19, no. 6-7, pp. 455–501, 2005.

[2] S. M. Lulich and W. G. Pearson Jr, "Three-/four-dimensional ultrasound technology in speech research," American Speech-Language-Hearing Association, Tech. Rep., 2019.

[3] E. Naga Karthik, E. Karimi, S. M. Lulich, and C. Laporte, "Automatic tongue surface extraction from three-dimensional ultrasound vocal tract images," *The Journal of the Acoustical Society of America*, vol. 147, no. 3, pp. 1623–1633, 2020.

[4] L. Tóth, G. Gosztolya, T. Grósz, A. Markó, and T. G. Csapó, "Multi-task learning of speech recognition and speech synthesis parameters for ultrasound-based silent speech interfaces," in *Interspeech 2018*, 2018, pp. 3172–3176.

[5] T. Hueber, E.-L. Benaroya, G. Chollet, B. Denby, G. Dreyfus, and M. Stone, "Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips," *Speech Communication*, vol. 52, no. 4, pp. 288–300, 2010.

[6] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

[7] T. G. Csapó, T. Grósz, G. Gosztolya, L. Tóth, and A. Markó, "Dnn-based ultrasound-to-speech conversion for a silent speech interface," in *Interspeech 2017*, 2017, pp. 3672–3676.

[8] T. G. Csapó, C. Zainkó, L. Tóth, G. Gosztolya, and A. Markó, "Ultrasound-based articulatory-to-acoustic mapping with waveglow speech synthesis," in *Interspeech 2020*, 2020, pp. 2727–2731.

[9] Z. Xia, R. Yuan, Y. Cao, T. Sun, Y. Xiong, and K. Xu, "A systematic review of the application of machine learning techniques to ultrasound tongue imaging analysis," *The Journal of the Acoustical Society of America*, vol. 156, no. 3, pp. 1796–1819, 09 2024. [Online]. Available: https://doi.org/10.1121/10.0028610

[10] L. Tóth, A. Honarmandi Shandiz, G. Gosztolya, and T. G. Csapó, "Adaptation of tongue ultrasound-based silent speech interfaces using spatial transformer networks," in *Interspeech 2023*, 2023, pp. 1169–1173.

[11] L. Tóth and A. H. Shandiz, "3d convolutional neural networks for ultrasound-based silent speech interfaces," in *Artificial Intelligence and Soft Computing: 19th International Conference, ICAISC 2020, Zakopane, Poland, October 12-14, 2020, Proceedings, Part I 19*. Springer, 2020, pp. 159–169.

[12] C. Zainkó, L. Tóth, A. H. Shandiz, G. Gosztolya, A. Markó, G. Németh, and T. G. Csapó, "Adaptation of tacotron2-based text-to-speech for articulatory-to-acoustic mapping using ultrasound tongue imaging," in *11th ISCA Speech Synthesis Workshop (SSW 11)*, 2021, pp. 54–59.

[13] M. S. Ribeiro, J. Sanger, J.-X. Zhang, A. Eshky, A. Wrench, K. Richmond, and S. Renals, "Tal: A synchronised multi-speaker corpus of ultrasound tongue imaging, audio, and lip videos," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 1109–1116.

[14] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3617–3621.

[15] A. Mathis, P. Mamidanna, K. M. Cury, T. Abe, V. N. Murthy, M. W. Mathis, and M. Bethge, "Deeplabcut: markerless pose estimation of user-defined body parts with deep learning," *Nature neuroscience*, vol. 21, no. 9, pp. 1281–1289, 2018.

[16] A. Wrench and J. Balch-Tomes, "Beyond the edge: Markerless pose estimation of speech articulators from ultrasound and camera images using deeplabcut," *Sensors*, vol. 22, no. 3, p. 1133, 2022.

[17] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in neural information processing systems*, vol. 33, pp. 17 022–17 033, 2020.

[18] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Interspeech 2020*, 2020, pp. 5036–5040.

Attacks by Human-Like Synthetic Speech on Anti-Spoofing Systems: Challenges, Insights, and Solutions

Aurosweta Mahapatra

Johns Hopkins University, USA

Attacks by Human-Like Synthetic Speech on Anti-Spoofing Systems: Challenges, Insights, and Solutions

Aurosweta Mahapatra[1]

[1]Center for Language and Speech Processing, Johns Hopkins University, USA

amahapa2@jhu.edu

1. Motivation

This work investigates the growing vulnerability of anti-spoofing systems to human-like spoofed speech generated by modern text-to-speech (TTS) and voice conversion (VC) models. While prior research has explored various spoofing techniques and generation models, it often overlooks the nuanced and diverse nature of real human speech, particularly emotional variation. Our experiments show that recent synthesis models can produce speech that, while not always convincing to human listeners, is natural enough to deceive state-of-the-art anti-spoofing systems. We demonstrate that attackers can exploit this by launching emotion-targeted attacks, significantly increasing the likelihood of bypassing existing defenses. This reveals a critical gap in current approaches. The objective of this research is to analyze these vulnerabilities and to develop more robust models capable of detecting and withstanding emotionally expressive, human-like spoofing attacks.

2. Introduction

Speech anti-spoofing focuses on detecting spoofed audio generated through replay attacks, speech synthesis, and voice conversion techniques [1]. Among these, attacks based on TTS and VC systems are of particular concern. This is due to recent advancements in speech synthesis that enable the generation of human-like, emotionally expressive speech using TTS and VC methods. Such capabilities raise the risk of misuse, including the deception of biometric verification systems and the spread of misinformation through non-consensual voice impersonation [1, 2].

State-of-the-art speech synthesis systems are increasingly capable of generating expressive speech, capturing a wide range of emotions with high fidelity [3–8]. While such synthetic speech may not consistently deceive human listeners, our findings indicate that it is often sufficient to bypass current anti-spoofing systems. Moreover, zero-shot TTS models [9–11] can generate emotionally rich speech from only a few seconds of reference audio, making it easier and more accessible for adversaries to craft convincing spoofing attacks. Although state-of-the-art anti-spoofing models [12, 13] achieve strong performance on neutral synthetic speech, their detection accuracy degrades significantly when evaluated on emotionally expressive spoofed speech. These findings underscore the need for anti-spoofing systems that are robust to emotion-targeted attacks and more effective at detecting human-like synthetic speech.

Challenges like ASVspoof [14] and the Audio Deep Synthesis Detection (ADD) Challenge [15, 16] have fostered progress in this domain by supporting the creation of benchmark datasets including ASVspoof 2019, 2021, and 2024 [17–19]. However, these corpora do not account for emotional variability, leaving a critical gap in addressing emotion-driven spoofing.

Current anti-spoofing approaches include end-to-end models like RawNet2 [12] and AASIST [13], as well as self-supervised learning (SSL) models [20–23], which operate directly on raw waveforms and eliminate the need for handcrafted features. Alternatively, some methods combine handcrafted features with supervised classifiers, such as LCNN [24], ASSERT [25], and ResNet34 [26]. However, these models have predominantly been trained and evaluated on ASVspoof datasets [17–19], which largely comprise emotionally neutral speech. As a result, their robustness to emotionally expressive spoofed speech remains underexplored. This highlights the need for a dedicated dataset containing emotionally expressive synthetic speech, which can serve both as a benchmark for evaluating model performance under emotion-targeted attacks and as a foundation for developing more robust anti-spoofing methods.

3. Emotion Vs Anti-spoofing

3.1. Dataset

Emotion-targeted spoofing attacks represent a potentially serious yet underexplored threat to the reliability of current anti-spoofing systems. Due to the lack of datasets that capture emotional variability in synthetic speech and to show the impact empirically, we introduced EmoSpoof-TTS. It is a dataset specifically developed to facilitate the study of emotion-targeted attacks. It comprises spoofed speech synthesized using three state-of-the-art zero-shot TTS models: StyleTTS2 [9], F5-TTS [10], and CosyVoice [11], all capable of generating high-quality, emotionally expressive speech from minimal reference input. The dataset spans four emotional categories: *Happiness*, *Anger*, *Sadness*, and *Neutral*, and includes 36,000 spoofed utterances from 10 speakers (5 male and 5 female) across the three TTS models. These are paired with 12,000 bona fide utterances drawn from the ESD dataset. EmoSpoof-TTS serves as an essential resource for empirically demonstrating and analyzing the vulnerability of state-of-the-art anti-spoofing systems to emotion-targeted synthetic speech.

3.2. Analysis

We use the pre-trained anti-spoofing model RawNet2 [12], trained on the ASVspoof 2019 training and development partitions, for our analysis. To examine the impact of emotion-targeted attacks on anti-spoofing systems, we first evaluate RawNet2 on ASVspoof datasets that lack emotional variability. Table 1 reports the Equal Error Rates (EERs) for the Logical Access (LA) track of ASVspoof 2019 and 2021, as well as

Track 1 of ASVspoof 2024. As shown, RawNet2 performs well on ASVspoof 2019 and 2021, with relatively low EERs. However, its performance deteriorates considerably on ASVspoof 2024, a more recent dataset, highlighting the model's limited generalization to newer and more sophisticated spoofing techniques.

Table 1: *Performance of Pre-Trained RawNet2 model on test-set of Traditional Datasets (EER%)*

	Test Dataset		
Model	**ASVspoof 2019**	**ASVspoof 2021**	**ASVspoof 2024**
RawNet2	4.60	14.73	40.67

We then evaluated the pre-trained RawNet2 model on EmoSpoof-TTS, using spoof samples from four speakers while maintaining speaker consistency across all three TTS models.

Table 2: *Performance of Pre-Trained RawNet2 model on EmoSpoof-TTS (EER%)*

EmoSpoof-TTS						
TTS Model	**HAS**	**Neutral**	**Happy**	**Angry**	**Sad**	**Overall**
StyleTTS2	**44.03**	39.00	45.00	35.17	46.75	43.06
F5-TTS	**42.97**	38.08	38.42	39.42	47.25	41.92
CosyVoice	**42.44**	42.17	39.58	39.75	47.50	42.65

The evaluation presented in Table 2 provides insights into the vulnerabilities of the pre-trained RawNet2 model when exposed to high-quality, emotionally expressive synthetic speech. The model struggles to detect fake samples from EmoSpoof-TTS, with EER values higher than those for the ASVspoof 2024 dataset across all TTS models. The EER for HAS across all three models is consistently higher than for Neutral. For instance, the EER for HAS in the case of StyleTTS2 is 44.03%, whereas for Neutral, it is 39%. This indicates lower confidence of the anti-spoofing model in distinguishing between spoofed and bona-fide speech when it is exposed to fake emotional speech. These results reinforce our claim that SOTA anti-spoofing models, despite performing well on traditional datasets like ASVspoof 2019 and ASVspoof 2021, struggle with speech samples from recent emotionally expressive TTS models, especially when handling emotional speech, making them unreliable in such scenarios. Additionally, our analysis reveals that EER values vary across emotion categories; for instance StyleTTS2, the EER for Happiness is 45.00%, for Sadness is 46.75%, and for Anger is 35.17%. This highlights the unreliable nature of spoof detection across emotion and generation models and it is vulnerable to emotion-targeted attacks.

3.3. Proposed Model

To improve performance on emotionally expressive synthetic speech, we propose the Gated Ensemble Method (GEM). GEM integrates multiple emotion-specialized anti-spoofing models with a Speech Emotion Recognition (SER) model that functions as a gating mechanism. This design allows the system to selectively prioritize decision of the most appropriate model based on the detected emotional content of the input. By doing so, GEM enhances detection performance for high-quality, emotionally expressive speech and helps reduce performance disparities across different emotions. Our method is tailored specifically to address the challenges posed by emotion-targeted spoofing attacks. Experimental results show that GEM outperforms the baseline model when evaluated on the EmoSpoof-TTS dataset.

4. Future Work

As discussed in the previous section, current anti-spoofing models struggle to detect emotionally expressive synthetic speech. These limitations are evident even with recent datasets such as ASVspoof 2024, which feature advanced TTS models. Interestingly, humans often outperform machines in identifying real speech, while machines tend to perform slightly better at detecting fake speech [27]. This suggests that if machines could better understand or model real speech in the way humans do, their performance in detecting fake speech might improve significantly.

Humans are known to rely on subtle cues that are often underutilized or overlooked by current anti-spoofing models. I aim to explore which specific aspects enable humans to generalize better across both emotional and non-emotional real speech, and investigate how such perceptual strategies can be integrated into anti-spoofing models.

As part of this research direction, I plan to conduct a human perception study focused on emotionally expressive speech to compare human and machine performance more systematically. This study will provide further insights into how humans identify authenticity and how models could benefit from learning similar representations.

My broader objective is to design models that not only improve performance on datasets like EmoSpoof-TTS but also generalize well to diverse spoofing scenarios. I am particularly interested in three key research directions:

- Human-inspired model design: Incorporating human perception of audio into machine to improve model robustness.
- Multilingual and accented speech: Investigating vulnerabilities and robustness across different languages and accents in anti-spoofing.
- Audio-visual integration: Exploring the fusion of audio and visual modalities to enhance detection performance in multimodal settings.

5. Conclusion

This work presented our investigation into the vulnerability of anti-spoofing models to emotionally expressive synthetic speech. We introduced EmoSpoof-TTS to study this problem and showed that models like RawNet2 fail under emotion-targeted attacks. To address this, we proposed the Gated Ensemble Method (GEM), which improves robustness by leveraging emotion-based gating. Moving forward, I aim to explore human-inspired modeling, multilingual and accented speech, and audio-visual integration to build more generalizable and perceptually aligned anti-spoofing systems. In addition, I plan to conduct human perception studies on emotionally expressive fake speech to better understand the cues humans rely on when identifying deepfakes. These insights will guide the development of models that "listen" more like humans. Ultimately, our goal is to bridge the gap between machine and human perception in spoof detection.

6. References

[1] A. Khan, K. M. Malik, J. Ryan, and M. Saravanan, "Battling voice spoofing: a review, comparative analysis, and generalizability evaluation of state-of-the-art voice spoofing counter measures," *Artificial Intelligence Review*, vol. 56, no. Suppl 1, pp. 513–566, 2023.

[2] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward," *Applied intelligence*, vol. 53, no. 4, pp. 3974–4026, 2023.

[3] Y. Guo, C. Du, X. Chen, and K. Yu, "Emodiff: Intensity controllable emotional text-to-speech with soft-label guidance," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[4] H. Tang, X. Zhang, N. Cheng, J. Xiao, and J. Wang, "Ed-tts: Multi-scale emotion modeling using cross-domain emotion diarization for emotional speech synthesis," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12 146–12 150.

[5] R. Shimizu, R. Yamamoto, M. Kawamura, Y. Shirahata, H. Doi, T. Komatsu, and K. Tachibana, "Prompttts++: Controlling speaker identity in prompt-based text-to-speech using natural language descriptions," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12 672–12 676.

[6] S.-H. Lee, S.-B. Kim, J.-H. Lee, E. Song, M.-J. Hwang, and S.-W. Lee, "Hierspeech: Bridging the gap between text and speech by hierarchical variational inference using self-supervised representations for speech synthesis," in *Advances in Neural Information Processing Systems*, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds., 2022. [Online]. Available: https://openreview.net/forum?id=awdyRVnfQKX

[7] M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar, and W.-N. Hsu, "Voicebox: Text-guided multilingual universal speech generation at scale," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [Online]. Available: https://openreview.net/forum?id=gzCS252hCO

[8] S. Chen, C. Wang, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Neural codec language models are zero-shot text to speech synthesizers," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 705–718, 2025.

[9] Y. A. Li, C. Han, V. Raghavan, G. Mischler, and N. Mesgarani, "Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 19 594–19 621. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/3eaad2a0b62b5ed7a2e66c2188bb1449-Paper-Conference.pdf

[10] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. Zhao, K. Yu, and X. Chen, "F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching," *arXiv preprint arXiv:2410.06885*, 2024.

[11] Z. Du, Q. Chen, S. Zhang, K. Hu, H. Lu, Y. Yang, H. Hu, S. Zheng, Y. Gu, Z. Ma *et al.*, "Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens," *arXiv preprint arXiv:2407.05407*, 2024.

[12] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with rawnet2," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6369–6373.

[13] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6367–6371.

[14] Z. Wu, T. Kinnunen, N. Evans, and J. Yamagishi, "Asvspoof 2015: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," *Training*, vol. 10, no. 15, p. 3750, 2014.

[15] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan, S. Liang, S. Wang, S. Zhang, X. Yan, L. Xu, Z. Wen, and H. Li, "Add 2022: the first audio deep synthesis detection challenge," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 9216–9220.

[16] J. Yi, J. Tao, R. Fu, X. Yan, C. Wang, T. Wang, C. Y. Zhang, X. Zhang, Y. Zhao, Y. Ren *et al.*, "Add 2023: the second audio deepfake detection challenge," *arXiv preprint arXiv:2305.13774*, 2023.

[17] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee, L. Juvela, P. Alku, Y.-H. Peng, H.-T. Hwang, Y. Tsao, H.-M. Wang, S. L. Maguer, M. Becker, F. Henderson, R. Clark, Y. Zhang, Q. Wang, Y. Jia, K. Onuma, K. Mushika, T. Kaneda, Y. Jiang, L.-J. Liu, Y.-C. Wu, W.-C. Huang, T. Toda, K. Tanaka, H. Kameoka, I. Steiner, D. Matrouf, J.-F. Bonastre, A. Govender, S. Ronanki, J.-X. Zhang, and Z.-H. Ling, "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech Language*, vol. 64, p. 101114, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230820300474

[18] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, and K. A. Lee, "Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2507–2522, 2023.

[19] X. Wang, H. Delgado, H. Tak, J.-w. Jung, H.-j. Shim, M. Todisco, I. Kukanov, X. Liu, M. Sahidullah, T. Kinnunen *et al.*, "Asvspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale," *arXiv preprint arXiv:2408.08739*, 2024.

[20] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[21] 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.

[22] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "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.

[23] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[24] Z. Wu, R. K. Das, J. Yang, and H. Li, "Light convolutional neural network with feature genuinization for detection of synthetic speech attacks," *arXiv preprint arXiv:2009.09637*, 2020.

[25] C.-I. Lai, N. Chen, J. Villalba, and N. Dehak, "Assert: Anti-spoofing with squeeze-excitation and residual networks," *arXiv preprint arXiv:1904.01120*, 2019.

[26] K. Z. Mon, K. Galajit, C. O. Mawalim, J. Karnjana, T. Isshiki, and P. Aimmanee, "Spoof detection using voice contribution on lfcc features and resnet-34," in *2023 18th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*. IEEE, 2023, pp. 1–6.

[27] K. Warren, T. Tucker, A. Crowder, D. Olszewski, A. Lu, C. Fedele, M. Pasternak, S. Layton, K. Butler, C. Gates *et al.*, "" better be computer or i'm dumb": A large-scale evaluation of humans as audio deepfake detectors," in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 2024, pp. 2696–2710.

Developing a Dynamical Model of the Coordination Between Melody and Syllabic Duration in Brazilian Portuguese

Gustavo Silveira

University of Campinas, Brazil

Developing a dynamical model of the coordination between melody and syllabic duration in Brazilian Portuguese

Gustavo Silveira[1]

[1]IEL, University of Campinas, Brazil

silveira@tuta.io

Abstract

Since the 1960s, speech researchers have proposed various quantitative models for the generation of fundamental frequency (F0) contours in an effort to describe the mechanisms underlying vocal fold tension control during speech. Although it is widely acknowledged that F0 timing interacts with the temporal organization of syllables within an utterance, most of these models focus exclusively on F0, abstracting away its relationship with syllables and typically treating syllable duration as fixed. The aim of this study is to advance such models by capturing the interplay between syllable duration and F0 during the realization of informational and contrastive narrow focus in Brazilian Portuguese (BP). Specifically, we propose to extend Barbosa's coupled-oscillator model of speech rhythm by integrating an F0 generation mechanism that interacts with the existing syllabic production oscillatory system, making the model capable of emulating the coordination between F0 and syllable duration under narrow focus in this language.

Index Terms: prosody, intonation, dynamical systems, Brazilian Portuguese, narrow focus

1. Previous works

Since Öhman's pioneering work [1], speech researchers have sought to model the control processes underlying fundamental frequency (F_0) production [2, 3, 4, 5, 6]. These models describe time-varying F_0 during speech as the output of a control system responsive to communicative functions.

The most well-known model is Fujisaki's [6], which posits two types of input commands: phrase commands (impulses) and accent commands (step functions). These correspond to the two main functions of prosody: phrasing (grouping words into higher-level units) and prominence (enhancing the salience of certain words). The F_0 contour results from superimposing the system's responses — critically damped second-order linear systems — to these commands and producing a proportional change in a logarithmic curve, representing F_0 declination.

Recently, this approach has begun to attract the attention of linguists seeking to address two key related problems: (i) how phonological units — such as the autosegmental-metrical (AM) tones H and L [7, 8]— are "phonetically implemented" in time by the vocal tract; (ii) while linguists increasingly recognize the role of continuous prosodic variation in conveying linguistic information, the computational nature of phonological theories limits their capacity to account for such gradient phenomena. In response, new models have been proposed as dynamical counterparts to intonational theories [2, 3]. These models tend to be local in scope, aiming to capture the intonation of only the portion associated with an underlying tonal target. An exception is PENTA [9, 5], which adopts a local approach to modeling global utterance melody by positing that every syllable has an underlying pitch target. Each syllable's F_0 curve is a solution of a critically damped second-order linear system. The system's final state at the end of one syllable serves as the initial condition for the next one, and the concatenation of the solution curves yields the utterance's overall melody.

Although these models have successfully emulated F_0 under the conditions for which they were designed, none accounts for the interaction between F_0 and the temporal dynamics of syllable production. While most assume some kind of temporal relationship between F_0 and syllables, they do not explicitly model this relationship. Instead, they focus solely on F_0 and, for the sake of simplicity, treat syllable duration as fixed. As a result, temporal aspects of syllable production, such as syllable duration and speech rate, are excluded — even as input parameters.

In this doctoral research, the goal is to advance previous quantitative models of F_0 by developing a model, applied to Brazilian Portuguese, that accounts for F_0 production *coupled* with the dynamics of syllable production. Specifically, within a dynamical systems framework, the model should address the following three questions:

- How is F_0 coordinated with syllables?
- How does variation in syllable duration influence F_0?
- How does F_0 movement affect syllable duration?

2. The coupled-oscillators model

This research builds upon an existing model of syllable production dynamics, Barbosa's coupled-oscillator model of speech rhythm [10, 11, 12], with the goal of extending it by incorporating a melody-generation component.

Barbosa's model posits that speech is temporally structured in terms of vowel onsets and proposes a planning mechanism that anticipates the timing of upcoming vowel onsets, with the interval between two adjacent vowel onsets defining a syllable-sized unit — referred to as the VV unit.

The mechanism comprises two coupled oscillators. The first, the syllabic oscillator, marks the temporal distribution of vowel onsets. Each oscillatory cycle corresponds to one VV unit, and its frequency typically hovers around 5 Hz. Hierarchically, it imposes the syllabic structure on segmental gestures [13]. The second oscillator, the phrase-stress oscillator, marks the temporal distribution of prosodic phrases, with a frequency around 1 Hz. The two oscillators represent competing organizing forces: the syllabic oscillator promotes regularity and cohesion at the syllabic level (around 5 Hz), whereas the phrase-stress oscillator imposes regularity and cohesion at the phrasal level (around 1 Hz). The dynamical interaction of these opposing forces give rise to the complex patterns of syllable duration

along the utterance. Specifically, the phrase-stress oscillator exerts an exponential decelerating force on the syllabic oscillator, causing the period of the syllabic oscillations to increase exponentially over the course of the phrase-stress cycle. At the end of this cycle, a reset occurs, and the exponential deceleration begins again. The magnitude of the decelerating force is proportional to the prominence strength intended by the speaker. Consequently, the coupling between the phrase-stress and syllabic oscillators serves as the mechanism underlying the prosodic realization of phrasing and prominence.

Barbosa's coupled-oscillators model is currently implemented as a discrete-time dynamical system. The discrete time dimension corresponds to the sequence of syllables in the utterance, and a difference equation governs the change ΔT in the period T of the syllabic oscillator, as influenced by the phrase-stress oscillator, from one syllable to the next.

3. What I have done so far

To inform my modeling decisions, I collected experimental data on the production of narrow focus [14] in Brazilian Portuguese — both informational and contrastive — across varying speech rates. I chose to focus on narrow focus realization because it is a prosodically marked function with stable and salient acoustic correlates, typically involving modifications to both F_0 and syllable duration. Eight participants took part in the experiment, which used a standard question–answer paradigm for focus elicitation. The recordings were transcribed and segmented into phones, syllables, and words. I am currently in the analysis phase. While I have outlined preliminary ideas for the modeling stage, the formal development of the model has not yet begun.

4. Key challenges

4.1. How to analyze the data?

The first challenge is to determine which types of measurements and techniques are most appropriate for investigating how F_0 aligns with syllables and how F_0 movements interact with syllable duration. Regarding alignment, my hypothesis is that pitch targets (see next challenge) are consistently aligned with vowel onsets. To test this hypothesis, I plan to identify critical points — such as local maxima, minima, and inflection points — in the F_0 contour as well as in its first and second derivatives. I will then measure the distance between these points and several syllable-related landmarks: vowel onset, midpoint, and offset; and phonological syllable onset, midpoint, and offset. I will try not only to determine which of these distances tends to be shortest, but more importantly, to identify which exhibits the least variance, as lower variance would indicate greater temporal stability. As for the interaction between F_0 and syllabic duration, I intend to examine correlations between the magnitude of the F_0 critical points and syllable duration. I am also doing qualitative examinations of the F_0 curves to investigate what happens to the shape of F_0 movements when the syllable is significantly shortened or lengthened. I will also look at potential correlations between F_0 and syllable duration curves [15].

4.2. What is the target?

The second challenge in analyzing and modeling F_0 is to determine what constitutes the unit of melodic action (a pitch gesture) [16, 17] — that is, the unit corresponding to an intended communicative articulatory action by the speaker. This challenge translates into how linguistic information is mathematically represented in the model. For instance, AM theory posits that there are two basic melodic actions: reaching a relatively high pitch level (H tone) or a low pitch level (L tone) [8]. Pitch rises and falls are considered incidental by-products of sequencing L and H tones (or vice versa). AM-based models tend to represent H and L tones as point-attractors [2, 3]. The PENTA model, by contrast, assumes that pitch targets are best represented as linear functions with varying slope and intercept [9]. The speaker's task is then to asymptotically approach these target lines through the control of F_0.

4.3. Integration of discrete-time and continuous-time dynamical systems

Perhaps the most significant challenge in this research is the temporal mismatch between F_0 and syllable duration: they do not operate on the same time scale. Syllables are discrete units, and a sequence of syllables yields a discrete sequence of durations. For this reason, discrete-time systems — defined by difference equations where time varies over $\mathbb{Z}$ —, are well-suited to capturing the temporal evolution of syllable duration across an utterance. In contrast, vocal fold tension varies continuously over time. For this reason, most models of F_0 production dynamics employ differential equations, where the independent variable t varies over $\mathbb{R}$.

One possible strategy to address this temporal incompatibility is to abandon the discrete characterization of syllables by redefining them in terms of continuous acoustic (e.g., the periodic energy envelope) or articulatory (e.g., jaw opening) variables. Another strategy — one which I do not intend to pursue — is to downsample the F_0 contour to match the syllabic time scale. Instead, I plan to adopt a third approach: to represent syllable timing as a continuous function, such as an impulse train or a periodic oscillatory function. The specific mathematical implementation of this strategy remains an open question and will be the focus of my ongoing work.

5. Plans for the future

I am currently analyzing the experimental data to investigate how speakers simultaneously control and coordinate F_0 and syllable duration to encode narrow focus. These results, together with existing literature on prosodic production, will form the basis for defining the model's assumptions. I envision a dynamical model defined by a set of ordinary differential equations, in which the two dependent variables are $F_0(t)$, representing the F_0, and $T(t)$, representing the period of the syllabic oscillator cycle at time t. These two variables will be coupled, meaning that each functions as a time-varying parameter in the differential equations governing the other.

The model will be tested and refined using natural speech data. The experimental dataset will be split into two non-overlapping subsets: training and testing. The training subset will be used to estimate model parameters for simulating narrow focus in Brazilian Portuguese. These parameters will then be applied to generate F_0 contours for the utterances in the testing subset, which are distinct from those used for training to avoid overfitting. Model accuracy will be evaluated through at least three methods: (i) visual comparison of generated and original F_0 contours; (ii) quantitative measures of fit, such as root mean square error and Pearson correlation; and (iii) perceptual judgments of the naturalness of the synthesized contours by human listeners.

6. References

[1] S. Öhman, "Word and sentence intonation: A quantitative model," *Speech Transmission Laboratory - Quarterly Progress and Status Report*, vol. 8, no. 2-3, pp. 020–054, 1967.

[2] K. Iskarous, J. Cole, and J. Steffman, "A minimal dynamical model of Intonation: Tone contrast, alignment, and scaling of American English pitch accents as emergent properties," *Journal of Phonetics*, vol. 104, p. 101309, May 2024.

[3] S. Roessig, D. Mücke, and M. Grice, "The dynamics of intonation: Categorical and continuous variation in an attractor-based model," *PloS one*, vol. 14, no. 5, p. e0216859, 2019.

[4] G. Bailly and B. Holm, "SFC: A trainable prosodic model," *Speech Communication*, vol. 46, no. 3, pp. 348–364, 2005.

[5] S. Prom-on, Y. Xu, and B. Thipakorn, "Modeling tone and intonation in Mandarin and English as a process of target approximation," *The Journal of the Acoustical Society of America*, vol. 125, no. 1, pp. 405–424, Jan. 2009.

[6] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *Journal of the Acoustical Society of Japan (E)*, vol. 5, no. 4, pp. 233–242, 1984.

[7] M. E. Beckman and J. Pierrehumbert, "Japanese prosodic phrasing and intonation synthesis," in *24th Annual Meeting of the Association for Computational Linguistics*, 1986, pp. 173–180.

[8] J. B. Pierrehumbert, "The phonology and phonetics of English intonation," Ph.D. dissertation, Massachusetts Institute of Technology, 1980.

[9] Y. Xu, "Speech melody as articulatorily implemented communicative functions," *Speech Communication*, vol. 46, no. 3, pp. 220–251, 2005.

[10] P. A. Barbosa, "From syntax to acoustic duration: A dynamical model of speech rhythm production," *Speech Communication*, vol. 49, no. 9, pp. 725–742, 2007.

[11] ——, "Análise e modelamento dinâmicos da prosódia do português brasileiro," *Revista de Estudos da Linguagem*, vol. 15, no. 2, pp. 75–96, 2007.

[12] ——, *Incursões em torno do ritmo da fala*. Campinas, SP: Pontes, 2006.

[13] C. P. Browman and L. Goldstein, "Articulatory Phonology: An Overview," *Phonetica*, vol. 49, no. 3-4, pp. 155–180, 1992.

[14] J. Aissen, "Documenting topic and focus," in *Key Topics in Language Documentation and Description*, P. Jenks and L. Michael, Eds. Honolulu: University of Hawai'i Press, 2023.

[15] P. A. Barbosa, "The interplay between syllabic duration and melody to signal prosodic functions in reading and story retelling in Brazilian Portuguese," 2024.

[16] C. A. Fowler, P. Rubin, R. E. Remez, and M. T. Turvey, "Implications for speech production of a general theory of action," *Language production*, vol. 1, pp. 373–420, 1980.

[17] J. Krivokapić, "Prosody in Articulatory Phonology," in *Prosodic Theory and Practice*. The MIT Press, Feb. 2022.

Controllability and Disentanglement in Expressive Text-to-Speech Systems

David Lindevelt

Leiden University, The Netherlands

Controllability and Disentanglement in Expressive Text-to-Speech Systems

David Lindevelt[1,2]

[1]LIACS, Leiden University, Netherlands
[2]Daisys.ai, Netherlands

1. Introduction

Speech is a multifaceted form of human communication. We communicate not only through words, but also use variations in tone, rhythm, pause, emphasis, and emotion. Human-machine interaction is becoming common, as we can see in recent audio chat models and virtual assistant services. Natural interaction in such settings requires the replication of complex expressions.

Text-to-speech (TTS) model architectures have advanced over time. Most neural TTS models have an Encoder-Decoder architecture, often with an additional vocoder [1]. Conventional TTS models primarily make use of Mel spectrograms as representations. With recent developments in data compression [2, 3], TTS architectures that use discrete audio codes as representation have become the go-to choice for producing natural sounding speech.

Although modern TTS systems exhibit impressive natural sounding speech, these systems struggle with providing detailed *control* over dimensions that impact speech synthesis, such as the identity of the speaker and the emotion expressed. Prior work reveals only moderate success due to the challenge of expressive feature disentanglement and the necessity of feature extraction in order to do so [4, 5, 6, 7, 8, 9]. Recent studies integrate audio samples to achieve more natural style generation [10]. Models like VALL-E [11, 12] demonstrate impressive voice cloning capabilities with just 3-seconds of audio prompt. Prompting techniques also enable the model to transfer not only the identity of the speaker but also style, pace, and acoustic environment conditions. Even though current models present promising results in controllability, they still lack fine-grained control over speech segments, and rely on very large datasets compared to earlier model architectures.

2. Motivation & Research Questions

While recent TTS systems have made impressive improvements in naturalness and voice cloning with short audio samples, they typically lack fine-grained control over various dimensions of speech synthesis. Prompt-based conditioning methods transfer speaker identity along with other aspects of speech such as acoustic background, rhythm, and emotion [13]. These methods are limited in providing users with selective control over individual dimensions of speech characteristics.

Our recent research on emotion alignment between text and speech has revealed that even emotion values extracted by state-of-the-art (SOTA) models show limited alignment between text emotion and speech emotion beyond the Valence dimension [14]. We found that the genre of content is also an important factor that affects alignment. This misalignment highlights that models cannot effectively capture and disentangle complex features, emphasizing the need for more nuanced disentanglement approaches, hence more sophisticated conditioning of TTS models for controllable emotional speech generation.

By advancing our theoretical understanding of how different speech dimensions can be effectively disentangled and independently controlled, we aim to develop TTS systems that offer both natural sounding results and precise control over expressive characteristics. Such advances would benefit applications ranging from accessibility tools to creative content production, while also making sophisticated speech synthesis more accessible to languages and domains with limited resources.

We address the following research questions.

- How can we develop text-to-speech systems that effectively disentangle and independently control multiple expressive dimensions of speech?
- What architectural and conditioning mechanisms enable fine-grained control over the full spectrum of prosodic and stylistic elements in neural text-to-speech systems?
- How can contextual information and multimodal signals be leveraged to create more naturally expressive and controllable speech synthesis?

3. Key Challenges

We have identified four key challenges that currently limit progress in controllable and expressive speech synthesis.

First, many cutting-edge models don't share their code or detailed training procedures, which hinders reproducibility and slows down collective progress in the field. This challenge is compounded by the fact that these models are often trained on a massive scale, requiring computational resources that are out of reach for many researchers. This disparity creates a divide in the research community, potentially limiting the diversity of voices and ideas that contribute to TTS technology. We observe a trend in recent TTS models with similar architectures (operating on discrete audio codes) using increasingly larger datasets. Massive black-box models may achieve impressive results through scaling, but often at the cost of interpretability and targeted control (and training cost). We believe that more efficient approaches focusing on the principled disentanglement of speech attributes could not only reduce resource requirements but also enhance controllable systems that understand the components of expressive speech rather than merely mimicking it.

Second, the field continues to face a trade-off between naturalness and controllability [13]. Many systems that offer detailed control parameters produce speech that sounds mechanical or artificial, while the most natural-sounding systems often lack interpretable control mechanisms. This suggests a need for novel architectures and training approaches that can maintain naturalness while providing interpretable, fine-grained control over speech synthesis.

Third, the lack of high-quality, fine-grained annotations hinders the development of disentangled speech synthesis. Many TTS systems use measurements such as pitch, energy, or keyword-based annotations such as personality/style traits. These are solid choices, but are not yet sufficiently defined. For example, context-dependent features are often used to represent speaker identity. While the features tied strongly to speaker identity such as mean pitch values generally used as style of speech [15].

Finally, the evaluation of model training procedure and generated speech remains one of the major bottlenecks for the development of TTS systems. The manual evaluation of synthesized speech requires carefully set-up experiments with human annotators. When it comes to the annotation of subtle differences, even human annotators struggle to follow instructions during evaluation. This creates additional difficulty in assessment of the methods and models.

4. Discussion of results

We faced significant challenges in our initial research trajectory due to the limited availability of code resources for recent TTS architectures. This constraint hindered our ability to reproduce recent advancements and build on top of these developments. Therefore we spent considerable effort to investigate and implement discrete code-based architectures. This foundational work provided several important benefits.

1. Comprehensive understanding of transformer-based TTS architectures.
2. Building a robust training and experimental framework.
3. Creation of adaptable baseline models for testing novel approaches. As an example of this capability, we successfully implemented the architectural design of Battenberg et al. [16] within our discrete code-based TTS model.

Building on this technical foundation, we have formulated experiments investigating the use of surrounding contextual information to enhance emotionally coherent speech synthesis. Our hypothesis was that controlling expressivity of speech using surrounding textual content to guide the generation process could produce more natural and contextually coherent synthesized speech. However, prior to full-implementation, we conducted a systematic cross-modal analysis of emotional alignment between text and speech in podcasts, audiobooks, and TED talks. This study revealed crucial insights: while LLMs are good at recognizing emotions from situational text [17], the alignment with speech emotions is limited to the Valence dimension. Additionally, the genre of the text is an important moderator for the alignment. Furthermore, increasing context did not improve the alignment (See Table 1).

These findings highlight a critical insight for TTS research: the development of emotionally expressive speech synthesis systems cannot rely solely on the to be spoken textual input for comprehensive emotional control.

5. Future Plans & Expected Contributions

Our planned investigations focus on two complementary research directions: disentangling speaker identity and style, and improving emotional coherence through latent state modeling.

First, we aim to disentangle speaker identity and speaking style by introducing two distinct natural language descriptions as conditioning signals. To support this distinction we propose a novel depth-dependent gating mechanism, with different fusion ratios applied at different layers of the model architecture. This approach is inspired by human perception processes, where style recognition typically precedes speaker identification [18]. Consequently, our model design will emphasize style conditioning representations in earlier layers while progressively increasing the influence of speaker identity in deeper layers. Through this research, we aim to contribute to a novel framework for disentangling and independently controlling speaker identity and speech style through separate natural language prompts. Additionally, we will address the prevalent misconception in current literature where speaker identity and speech style are often mixed. We clarify that speaker identity encompasses stable, context-independent features (such as pitch range, gender, age), while speech style relates to dynamic, context-dependent attributes (like emphasis, emotional expression). Distinction is important for accurate modeling and annotation.

Table 1: *Correlation between* contextual *text-based and speech-based emotion prediction on Libriheavy dataset*

	Arousal	Dominance	Valence
llama3.3:70b	.416 (.016)	.175 (.054)	**.700** (.006)
llama3.1:70b	.427 (.021)	.015 (.025)	.697 (.016)
phi3:14b	.392 (.080)	.108 (.064)	.591 (.066)
gemma2:9b	.450 (.015)	**.297** (.026)	.641 (.023)
llama3.1:8b	**.473** (.064)	.119 (.117)	.642 (.040)
qwen2:7b	.267 (.067)	.064 (.044)	.679 (.016)
openchat:7b	.320 (.099)	.006 (.037)	.591 (.089)
gemma:7b	.230 (.052)	.144 (.100)	.549 (.053)

In our second line of investigation, we propose a model for contextually coherent emotional speech synthesis based on latent state tracking. Rather than relying on extensive contextual history or reference speech samples as input, we propose representing the speaker's emotional state as a continuously evolving multi-dimensional latent vector that encapsulates affective qualities while minimizing computational requirements.

By encoding emotional context as a compact latent representation that evolves with each utterance, the system can maintain psychological continuity while significantly reducing computational demands. The latent state vector, updated through attention mechanisms similar to those in state tracking research [19], provides a more cognitively plausible model of emotional evolution in human speech [20]. This framework enables more nuanced transitions than discrete category-based approaches, potentially improving both the naturalness and computational efficiency of expressive speech synthesis systems. Through this research, we expect to contribute both theoretical insights into the representation of emotional states in speech and practical advances in creating more coherent and natural-sounding expressive TTS systems.

6. Conclusion

We presented challenges and limitations in the field of controllable text-to-speech systems. We described the preliminary results from our initial study and future directions to address aforementioned challenges under the scope of Controllability and Disentanglement in Expressive TTS. These directions aim to advance TTS systems offering both naturalness and fine-grained expressive control while reducing computational demands through principled feature disentanglement.

7. References

[1] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A Survey on Neural Speech Synthesis," Jul. 2021, arXiv:2106.15561 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2106.15561

[2] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "SoundStream: An End-to-End Neural Audio Codec," Jul. 2021, arXiv:2107.03312 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2107.03312

[3] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High Fidelity Neural Audio Compression," Oct. 2022, arXiv:2210.13438 [cs, eess, stat]. [Online]. Available: http://arxiv.org/abs/2210.13438

[4] Z. Ju, Y. Wang, K. Shen, X. Tan, D. Xin, D. Yang, Y. Liu, Y. Leng, K. Song, S. Tang, Z. Wu, T. Qin, X.-Y. Li, W. Ye, S. Zhang, J. Bian, L. He, J. Li, and S. Zhao, "NaturalSpeech 3: Zero-Shot Speech Synthesis with Factorized Codec and Diffusion Models," Apr. 2024, arXiv:2403.03100 [eess]. [Online]. Available: http://arxiv.org/abs/2403.03100

[5] X. An, F. K. Soong, and L. Xie, "Disentangling Style and Speaker Attributes for TTS Style Transfer," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 646–658, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9693198/

[6] Y.-J. Zhang, S. Pan, L. He, and Z.-H. Ling, "Learning Latent Representations for Style Control and Transfer in End-to-end Speech Synthesis," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 6945–6949, iSSN: 2379-190X. [Online]. Available: https://ieeexplore.ieee.org/document/8683623/

[7] Z. Du, B. Sisman, K. Zhou, and H. Li, "Disentanglement of emotional style and speaker identity for expressive voice conversion," in *Proc. Interspeech 2022*, 2022, pp. 2603–2607.

[8] G. Zhang, Y. Qin, W. Zhang, J. Wu, M. Li, Y. Gai, F. Jiang, and T. Lee, "iemotts: Toward robust cross-speaker emotion transfer and control for speech synthesis based on disentanglement between prosody and timbre," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1693–1705, 2023.

[9] S. Dutta and S. Ganapathy, "Zero shot audio to audio emotion transfer with speaker disentanglement," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 371–10 375.

[10] Y. Li, C. Han, V. S. Raghavan, G. Mischler, and N. Mesgarani, "StyleTTS 2: Towards Human-Level Text-to-Speech through Style Diffusion and Adversarial Training with Large Speech Language Models," *arXiv.org*, 2023.

[11] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers," Jan. 2023, arXiv:2301.02111 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2301.02111

[12] M. Łajszczak, G. Cámbara, Y. Li, F. Beyhan, A. van Korlaar, F. Yang, A. Joly, Á. Martín-Cortinas, A. Abbas, A. Michalski, A. Moinet, S. Karlapati, E. Muszyńska, H. Guo, B. Putrycz, S. L. Gambino, K. Yoo, E. Sokolova, and T. Drugman, "BASE TTS: Lessons from building a billion-parameter Text-to-Speech model on 100K hours of data," Feb. 2024, arXiv:2402.08093 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2402.08093

[13] T. Xie, Y. Rong, P. Zhang, W. Wang, and L. Liu, "Towards Controllable Speech Synthesis in the Era of Large Language Models: A Survey," Mar. 2025, arXiv:2412.06602 [cs]. [Online]. Available: http://arxiv.org/abs/2412.06602

[14] Anonymized. et al., "Emotion alignment between text and speech is limited: A cross-modal study," Submitted, manuscript under review.

[15] M. Kawamura, R. Yamamoto, Y. Shirahata, T. Hasumi, and K. Tachibana, "LibriTTS-P: A Corpus with Speaking Style and Speaker Identity Prompts for Text-to-Speech and Style Captioning," Jun. 2024, arXiv:2406.07969 [eess]. [Online]. Available: http://arxiv.org/abs/2406.07969

[16] E. Battenberg, R. J. Skerry-Ryan, D. Stanton, S. Mariooryad, M. Shannon, J. Salazar, and D. Kao, "Very Attentive Tacotron: Robust and Unbounded Length Generalization in Autoregressive Transformer-Based Text-to-Speech," Oct. 2024, arXiv:2410.22179 version: 1. [Online]. Available: http://arxiv.org/abs/2410.22179

[17] J. Broekens, B. Hilpert, S. Verberne, K. Baraka, P. Gebhard, and A. Plaat, "Fine-grained Affective Processing Capabilities Emerging from Large Language Models," in *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. Cambridge, MA, USA: IEEE, Sep. 2023, pp. 1–8. [Online]. Available: https://ieeexplore.ieee.org/document/10388177/

[18] K. N. Spreckelmeyer, M. Kutas, T. Urbach, E. Altenmüller, and T. F. Münte, "Neural processing of vocal emotion and identity," *Brain and cognition*, vol. 69, no. 1, pp. 121–126, Feb. 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2642974/

[19] X. Feng, X. Wu, and H. Meng, "Injecting linguistic knowledge into BERT for Dialogue State Tracking," *IEEE Access*, vol. 12, pp. 93 761–93 770, 2024, arXiv:2311.15623 [cs]. [Online]. Available: http://arxiv.org/abs/2311.15623

[20] F.-J. Ren, Y.-Y. Zhou, J.-W. Deng, K. Matsumoto, D. Feng, T.-H. She, Z.-Y. Jiao, Z. Liu, T.-H. Li, S. Nakagawa, and X. Kang, "Tracking Emotions Using an Evolutionary Model of Mental State Transitions: Introducing a New Paradigm," *Intelligent Computing*, vol. 3, p. 0075, Jan. 2024. [Online]. Available: https://spj.science.org/doi/10.34133/icomputing.0075

Development of End-to-End Speech Translation Models for Indian Languages

Jamaluddin

Aligarh Muslim University, India

Development of End-to-End Speech Translation Models for Indian Languages

Jamaluddin [1]

[1]Aligarh Muslim University, India,

GI3860@myamu.ac.in

Abstract

For my doctoral work, I am trying to develop a novel sequence-to-sequence-based direct speech translation framework capable of translating speech from one Indian language to another without relying on any intermediate text representation. Through bypassing the traditional ASR-MT-TTS pipeline, the proposed work tries to streamline the translation process and reduce cumulative errors. Recent advances in deep learning have demonstrated that direct speech translation can outperform conventional cascaded systems in both efficiency and translation quality.

Index Terms: Speech-to-speech translation, Indian languages, deep learning, low-resource languages, multilingual corpora

1. Introduction

Speech-to-Speech Translation (S2ST) is a crucial research domain as it bridges linguistic gaps to facilitate clear and effective communication across diverse language speakers globally. In a world rich with voice, tone, and emotion, converting speech from one language into speech of another without detouring through text isn't just innovation, it's a revolution. The current state of S2ST is marked by significant advancements due to deep learning technologies [1] [2]. Presently, cascade(or indirect) speech-to-speech translation as shown in figure 1, which involves converting speech to text, translating the text, and then synthesizing it back into speech, is a common approach [3]. This approach produces good performance because it benefits from the maturity of text-based translation technologies and has extensive training datasets available. However, cascade S2ST approach faces several limitations. Among many, the multi-step process faces difficulty in handling speech nuances such as tone, emotion, and cultural context may also be lost when translating through intermediate textual representations [4]. Therefore, recent developments focus on end-to-end models that translate speech from one language directly into another without intermediate text representation [5]. Direct translation methods also have many challenges, primarily due to unavailability of datasets in the form of language pairs. They also struggle with maintaining the emotional and prosodic features of the original speech [4]. However, ongoing research and improvements in neural network architectures and training datasets continue to push the boundaries of what S2ST systems can achieve.

2. Related Works

Recently, there has been significant advances in the direct S2ST models [1,2,5-8]. Jia et al. [6] introduced Translatotron 2, a neural direct speech-to-speech translation model that can be trained end-to-end and exhibited significant improvement upon its predecessor, Translatotron. The model integrated a speech encoder, a linguistic decoder, an acoustic synthesizer, and a single attention module to facilitate high translation and speech generation quality. Translatotron 2 achieved performance comparable to cascade systems while improving speech naturalness. Lee et. al. [7] presented a novel approach to direct speech-to-speech translation (S2ST) by utilizing discrete speech units. The authors employed a self-supervised discrete speech encoder and a sequence-to-sequence speech-to-unit translation model to enhance performance without reliance on text transcripts. There are many other recent studies on direct S2ST technologies that produced promising results [8]. Nguyen el at. [9] presents an effective method to generate a large amount of synthetic S2ST data from unlabeled text corpora, it offers a practical method of making use of the enormous amount of unlabeled text that is currently available in a variety of languages.

3. Key Challenges

In existing speech translation systems, speech is first converted into text in the source language through automatic speech recognition (ASR), then translated into the target language via machine translation (MT), and finally synthesized back into speech using text-to-speech (TTS) models. While this pipeline has been useful, it introduces multiple sources of error at each stage, such as misrecognitions in ASR, incorrect translations in MT, and unnatural speech synthesis in TTS. Furthermore, these systems fail to capture critical aspects of communication, such as emotional tone, rhythm, and speaker intent, resulting in robotic and contextually inadequate translations. Speech-to-speech translation currently relies on a multi-step process (ASR $\rightarrow$ MT $\rightarrow$ TTS) that introduces errors at each stage, reducing translation accuracy. Intermediate text-based systems fail to capture important aspects of speech. Eliminating the text-based step can improve accuracy and preserve prosodic nuances, but this approach requires advanced models and high quality parallel speech datasets. The main challenges include creating diverse datasets and developing pre-trained models capable of handling direct speech translation, particularly for low resource Indian languages. Languages with limited labeled data, like many Indic languages, face difficulties in developing accurate speech translation models due to the scarcity of parallel speech datasets which hinders the development of effective speech translation models. Addressing these challenges is critical for developing robust, inclusive, and scalable S2ST models. Initially, we would focus on Indian languages that are native to the authors (i.e. Hindi and Urdu), as it would make the data collection and its analysis easier.

4. Motivation

Many end to end speech translation models have been developed in recent years on foreign languages typically Spanish - English. No End-to-End model has yet been designed for any Indian language in India. India is linguistically a rich country. There are 22 languages listed in the 8th schedule of the Indian constitution, belonging to 4 different language families. These languages have a collective speaker base of 1.2B, spread across 742 districts in India [10]. This motivates me to pursue research on direct speech translation in Indian languages to bridge the communication gap across diverse linguistic communities. India's rich multilingualism demands efficient translation tools to preserve cultural identity, enhance access to knowledge, and promote inclusivity. This research could revolutionize education, governance, and social interaction, fostering national unity and global engagement.

5. Research Questions

Among the many challenges to build a direct S2ST model, one is the compilation of vast and diverse datasets [4] to train models effectively across various languages, accents, and dialects. The scope of the problem is even bigger in a country like India, which is home to a vast array of languages and dialects, with over 20 officially recognized languages and hundreds of dialects[10]. The diversity makes it difficult to gather sufficient training data for each language, which is essential for developing robust direct S2ST models. The major research questions are as follows:

- How can the existing S2ST models be extended to other language pairs?
- Can the existing standards and best practices of collecting speech dataset be adopted for the diverse Indian context?
- Can we use pre-trained models which are trained on other language pairs (English-Spanish). In the case of Indian languages, would pre-trained S2ST models perform better or newly developed model from scratch?
- Can we train on synthetic data and test on real data. In that case we can create small dataset for testing purpose only.his is the original sentence.

6. Proposed Plan

Aligarh Muslim University (AMU), one of India's largest and most prestigious residential universities, accommodates nearly 25,000 students in its on-campus hostels. This academic environment is enriched by the linguistic and cultural diversity of its student body. While English is widely used for academic communication, the majority of students come from Hindi and Urdu-speaking regions, particularly from states such as Uttar Pradesh, Bihar, Jharkhand, Madhya Pradesh, and Uttarakhand. In addition, AMU hosts a significant number of students from linguistically diverse states like Jammu & Kashmir, West Bengal, Assam, and Kerala. This confluence of regional languages, cultures, and dialects makes AMU a microcosm of India's multilingual society—often referred to as a "mini-India." Recognizing this unique setting, our research team was inspired to curate a rich and diverse speech dataset aimed at building and evaluating direct speech-to-speech translation (S2ST) models, with a particular focus on addressing the challenges of Indian language translation. We are currently reviewing existing literature related to S2ST. We have written a review paper on direct speech translation that critically evaluates various approaches, highlights their trade-offs, and discusses future directions for enhancing real-time multilingual communication. The paper is available on arXiv:https://arxiv.org/abs/2503.04799 . Moreover, we have developed a website https://dr-recorder.onrender.com/ to collect speech samples for Hindi-English language pair from bilingual speakers of Hindi-speaking regions. We have already identified around 50 speakers and have collected more than 1,500 samples (sample rate is 44.1 kHz and file type is .wav), which includes ≈ 4 hours of real speech data. The duration of each sample ranges from as short as ≈ 1 second to as long as ≈ 69 seconds, with an average of ≈ 9 seconds. The Hindi-English text pairs which the speakers record are taken from Bharat Parallel Corpus Collection (BPCC)[10]. Following previous works [11], our target is to create a dataset of at least 120,000 speech samples for the first release within a time frame of 6-12 months. After collecting a decent amount of dataset, we would first try to implement existing pre-trained direct S2ST models and compare their performance with cascade models (may take around 8-12 months). It would be interesting to observe how pre-trained models, such as Spanish-English direct S2ST model, will perform for Hindi-English scenario as Hindi and Spanish sound very different. We have also planned to develop our own model from scratch, if the performance of pre-trained models is not satisfactory. We plan to extend this work to the Urdu language as well, since most students/teachers at our institute come from Hindi/Urdu heartland and are fluent in Hindi, English, and Urdu. We also plan to expand the dataset to include other Indian languages.

7. Thesis Contribution

Our primary purpose is to create a corpus of Indian languages for direct S2ST system. Moreover, the dataset can add value to existing Speech-to-Text and Text-to-Speech translation systems as well and can serve as a test benchmark for indirect or direct S2ST models. In the long run, our goal is to create a speech data hub for direct S2ST systems for the Indian languages pairs the speakers of which are easily available within University. Overall, we are focusing on enhancing multilingual speech translation research for Indian languages which can enable better communication across diverse linguistic communities in India and abroad. Our work will help the development of local voice assistants, language learning, and accessibility tools and will enrich the research arena of speech translation for Indian languages. The major contributions of the thesis can be summarized as follows:

- Addressing challenges associated with developing direct S2ST systems for Indian languages.
- Creating a corpus of Indian languages for direct S2ST systems.
- Developing and training end-to-end models on the parallel speech corpora for direct speech translation.
- Enhancing multilingual speech translation research for Indian languages.
- Enabling development of local voice assistants, accessibility tools etc.

8. References

[1] T. Kano, S. Sakti, and S. Nakamura, "Transformer-based direct speech-to-speech translation with transcoder," in *2021 IEEE spoken language technology workshop (SLT)*. IEEE, 2021, pp. 958–965.

[2] Q. Fang, Y. Zhou, and Y. Feng, "Daspeech: Directed acyclic transformer for fast and high-quality speech-to-speech translation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 72 604–72 623, 2023.

[3] L. Bentivogli, M. Cettolo, M. Gaido, A. Karakanta, A. Martinelli, M. Negri, and M. Turchi, "Cascade versus direct speech translation: Do the differences still make a difference?" *arXiv preprint arXiv:2106.01045*, 2021.

[4] M. Sperber and M. Paulik, "Speech translation and the end-to-end promise: Taking stock of where we are," *arXiv preprint arXiv:2004.06358*, 2020.

[5] Y. Zhu, Z. Gao, X. Zhou, Z. Ye, and L. Xu, "Diffs2ut: A semantic preserving diffusion model for textless direct speech-to-speech translation," *arXiv preprint arXiv:2310.17570*, 2023.

[6] Y. Jia, M. T. Ramanovich, T. Remez, and R. Pomerantz, "Translatotron 2: High-quality direct speech-to-speech translation with voice preservation," in *International Conference on Machine Learning*. PMLR, 2022, pp. 10 120–10 134.

[7] A. Lee, P.-J. Chen, C. Wang, J. Gu, S. Popuri, X. Ma, A. Polyak, Y. Adi, Q. He, Y. Tang *et al.*, "Direct speech-to-speech translation with discrete units," *arXiv preprint arXiv:2107.05604*, 2021.

[8] E. Nachmani, A. Levkovitch, Y. Ding, C. Asawaroengchai, H. Zen, and M. T. Ramanovich, "Translatotron 3: Speech to speech translation with monolingual data," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 686–10 690.

[9] X.-P. Nguyen, S. Popuri, C. Wang, Y. Tang, I. Kulikov, and H. Gong, "Improving speech-to-speech translation through unlabeled text," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[10] T. Javed, J. A. Nawale, E. I. George, S. Joshi, K. S. Bhogale, D. Mehendale, I. V. Sethi, A. Ananthanarayanan, H. Faquih, P. Palit *et al.*, "Indicvoices: Towards building an inclusive multilingual speech dataset for indian languages," *arXiv preprint arXiv:2403.01926*, 2024.

[11] M. Gupta, M. Dutta, and C. K. Maurya, "Benchmarking hindi-to-english direct speech-to-speech translation with synthetic data," *Language Resources and Evaluation*, pp. 1–39, 2025.

Speech Deepfake Detection and Beyond

Yassine El Kheir

DFKI, Germany

Speech Deepfake Detection and Beyond

Anonymous submission to Interspeech 2025

Abstract

With the rapid development of deep learning and generative AI, it has become increasingly easy to create or manipulate synthetic media. Even inexperienced users can now generate highly realistic content with minimal effort. While this progress brings exciting opportunities, it also introduces significant risks, particularly when these technologies are misused for malicious purposes such as spreading misinformation or compromising security systems. To mitigate these threats, we emphasize the need for:

- Robust deepfake detection models.
- Fairness-aware and explainability-driven training strategies to understand and mitigate performance disparities.

Index Terms: speech synthesis, robustness, interpretability, spoofing countermeasures

1. Introduction

Recent progress in generative artificial intelligence has significantly transformed the field of ultra-realistic speech synthesis. Advances in state-of-the-art text-to-speech (TTS) and voice conversion (VC) technologies have made it feasible to produce speech outputs that closely mimic natural human utterances [1]. Moreover, the emergence of large language models (LLMs) in audio and speech generation workflows has further elevated the potential of creating complex speech signals with high fidelity and nuance [2, 3, 4].

Despite these encouraging advancements, they have also introduced significant challenges. The increasing accessibility of tools capable of generating highly realistic speech has raised concerns about their misuse by malicious actors [5]. These technologies are being exploited to disseminate misinformation, incite hate speech, and even support acts of terrorism—activities that pose a direct threat to societal trust and the integrity of public discourse [6]. Synthetic speech has also become a tool for financial fraud; a recent survey reported voice-cloning attempts have affected approximately 7.7% of individuals, including high-profile cases of CEOs in the United Kingdom [7].

Current countermeasures predominantly rely on data-driven, black-box models fine-tuned on labeled examples of genuine versus fake audio. While these models often achieve high in-domain accuracy, they frequently fail to generalize to unseen attack algorithms or novel acoustic conditions. Moreover, they provide little insight into how decisions are made or why errors occur, especially across different demographic groups, limiting transparency and user trust [8, 9].

To address these challenges, this work proposes speech deepfake countermeasures that are both robust and distributionally fair. We structure our contributions along two complementary directions:

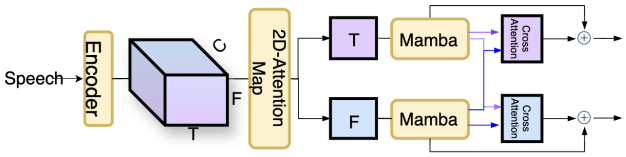


Figure 1: *Overview of the proposed BiCrossMamba-ST Framework.*

Robustness: Building deepfake detection models that fuse experts' observations with modeling architectures such as Transformers [10], Mamba [11].

Fairness and Explainability: Evaluating and mitigating disparities in model performance across demographic, linguistic, and environmental subgroups, while providing structured diagnostics to understand under which conditions the model is likely to succeed or fail.

2. Research Problem I: Robustness

2.1. Research Question

How can we design deepfake-detection architectures that integrate experts' observations to capture deepfake artifacts and generalize to diverse audio modalities?

2.2. Motivation

While recent deepfake detection models have made progress using large-scale self-supervised learning (SSL) and end-to-end training pipelines, many fail to explicitly incorporate well-established expert observations into their design. Not using such insights during model development can limit robustness and generalization. For instance, spoofing artifacts often appear as localized distortions in both spectral and temporal domains [12, 13], yet many models process the signal globally. Similarly, hand-crafted features like CQCCs [14], which are data-independent and sensitive to synthesis artifacts, are often ignored with SSL models. Designing models that integrate these expert cues can lead to more effective and generalizable deepfake detection systems.

2.3. Major contributions

2.3.1. BiCrossMamba-ST

To effectively capture localized deepfake artifacts, which often appear in specific spectral bands or temporal intervals, we propose **BiCrossMamba-ST** as shown in Figure 1, a spectro-temporal architecture designed with expert-inspired modular-

ity. The model begins by applying a **2D-attention** map over encoder features to highlight critical regions in the time–frequency space. These attended features are then passed through two parallel bidirectional Mamba blocks, each specialized to model either spectral or temporal dependencies. To unify the insights from both branches, we introduce a mutual **cross-attention** mechanism that enables effective information exchange between the spectral and temporal representations. Compared to traditional Graph Attention Networks (GATs), which rely on graph pooling to select the most informative nodes, Mamba's internal selection mechanism offers a more efficient alternative, it dynamically controls which input elements contribute to the hidden state at each time step, thereby filtering out irrelevant temporal and spectral features and emphasizing discriminative cues critical for deepfake detection. BiCrossMamba-ST achieves significant performance improvements, a 67.74% and 26.3% relative Equal Error Rate (ERR) gain over previous SOTA models AASIST [15] on LA21 [16] and DF21 [17] benchmarks, respectively, and a 6.80% improvement over SE-Rawformer [18] on DF21.

2.3.2. Multi-View Hybrid Fusion

Hybrid models that integrate both SSL representations and traditional handcrafted spectral features (SF) have demonstrated efficacy in various tasks, including ASR [19] and speaker verification [20], as well as recently in spoofing detection [21]. Handcrafted features such as MFCCs, which offer data-independence and task-robust features and are sensitive to synthetic artifacts, can enhance overall system robustness when combined with SSL features. We investigate the effectiveness of three widely used spectral features—MFCC, LFCC, and CFCC—when combined with SSL representations learned (Wav2Vec2.0 XLSR-53). Among different combination and fusion techniques, **Cross-Attention (SSL → SF)** with CQCC yields the best overall performance, achieving the lowest average EER of 6.80% and outperforming all fusion systems on DF21 and ITW [22] with EERs of 2.71% and 6.03%, respectively. All examined fusion approaches exploit additional spectral cues to outperform the baseline using only SSL features. The strong performance of **Cross-Attention (SSL → SF)** can be attributed to the ability of cross-attention to dynamically learn which components of the spectral features are most relevant when conditioned on the richer SSL representations. By attending selectively to complementary cues in the handcrafted features, the model effectively enhances discriminative capacity in challenging scenarios.

3. Research Problem II: Fairness and Explainability

3.1. Research Question

How can we train deepfake detectors so that their decisions are transparent, that is, we can pinpoint exactly which cues the model relied on, and at the same time remove unwanted dependencies on non-generalizable factors such as language or speaker gender?

3.2. Motivation

Prior research in speaker and speech verification has shown that model performance can improve when identity-related or language-dependent features are explicitly removed [23, 24] which can reduce domain mismatch and improve cross-lingual generalization, especially when transferring models between speakers of different linguistic backgrounds. In the context of deepfake detection, recent work such as FairSSD [25] has highlighted that detection models often rely unintentionally on speaker demographics such as age and gender—attributes not causally related to whether speech is real or fake. This dependency can introduce bias. These findings motivate the need for detection systems that are not only accurate but also fair and explainable.

3.3. Major contributions

Table 1: *EER (%) of standard SSL-based fine-tuning vs. our method on MLAAD multilingual test sets.*

Model	MLAAD-DE	MLAAD-AR	MLAAD-UK
Wav2Vec2.0	16.43	10.70	12.55
Our Method	**14.52**	**11.96**	**9.25**

We propose a two-stage strategy aimed at improving the interpretability and fairness of deepfake detection models by explicitly separating content from manipulation-related information.

The first stage focuses solely on modeling speech content. We train a **content encoder–decoder** pipeline using only real speech data, with inputs from a pretrained self-supervised model (e.g., Wav2Vec 2.0). The training objective is speech reconstruction, allowing the encoder to learn linguistic and speaker identity features without exposure to spoofed audio. This ensures that the learned content embeddings are disentangled from any deepfake artifacts. In the second stage, the **content encoder** is frozen. We introduce deepfake detector, trained jointly using both real and fake data. The goal is to identify content-agnostic spoofing-related features. An orthogonality constraint is enforced between the content and deepfake detector embeddings to ensure that artifact representations do not retain language, identity, or other irrelevant attributes captured in the first stage. This promotes fairness and reduces model bias toward content-specific cues.

We evaluate the proposed approach on a multilingual version of the MLAAD dataset [26]. Both baseline and our model are trained only on the English subset. As shown in Table 1, our method shows improved generalization on non-English test sets, confirming reduced reliance on language-specific features.

4. Potential Future Directions

Most deepfake detectors treat the problem as binary artifact classification, learning to spot data-dependent fake cues that differ by attack method. This makes them brittle: new spoofing techniques can evade detection. In contrast, genuine speech exhibits stable, content-driven patterns—consistent phonetic transitions, prosodic contours, and style–content couplings—that hold across speakers, languages, and recording conditions. By modeling what real speech is, we can flag any deviation leading to more generalizable and explainable detectors.

Can we detect deepfakes by measuring the mismatch between linguistic content (phonetic and semantic patterns) and vocal style (prosody, speaker identity, emotion), exploiting the fact that genuine speech maintains consistent content–style dependencies that are disrupted in synthetic audio?

5. References

[1] L. Pham, P. Lam, D. Tran, H. Tang, T. Nguyen, A. Schindler, F. Skopik, A. Polonsky, and H. C. Vu, "A comprehensive survey with critical analysis for deepfake speech detection," *Computer Science Review*, vol. 57, p. 100757, 2025.

[2] H. Hao, L. Zhou, S. Liu, J. Li, S. Hu, R. Wang, and F. Wei, "Boosting large language model for speech synthesis: An empirical study," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[3] D. Zhang, S. Li, X. Zhang, J. Zhan, P. Wang, Y. Zhou, and X. Qiu, "Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities," *arXiv preprint arXiv:2305.11000*, 2023.

[4] P. K. Rubenstein, C. Asawaroengchai, D. D. Nguyen, A. Bapna, Z. Borsos, F. d. C. Quitry, P. Chen, D. E. Badawy, W. Han, E. Kharitonov *et al.*, "Audiopalm: A large language model that can speak and listen," *arXiv preprint arXiv:2306.12925*, 2023.

[5] X. Li, K. Li, Y. Zheng, C. Yan, X. Ji, and W. Xu, "Safeear: Content privacy-preserving audio deepfake detection," in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 2024, pp. 3585–3599.

[6] M. Meaker, "Deepfake audio is a political nightmare," 2023.

[7] McAfee, "Beware the artificial impostor: A mcafee cybersecurity artificial intelligence report_2023," May 2023.

[8] J. Yi, C. Wang, J. Tao, X. Zhang, C. Y. Zhang, and Y. Zhao, "Audio deepfake detection: A survey," *arXiv preprint arXiv:2308.14970*, 2023.

[9] M. Li, Y. Ahmadiadli, and X.-P. Zhang, "Audio anti-spoofing detection: A survey," *arXiv preprint arXiv:2404.13914*, 2024.

[10] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[11] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[12] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with rawnet2," in *ICASSP 2021-2021 IEEE*. IEEE, 2021, pp. 6369–6373.

[13] H. Tak, J.-w. Jung, J. Patino, M. Kamble, M. Todisco, and N. Evans, "End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection," *arXiv preprint arXiv:2107.12710*, 2021.

[14] R. K. Das, "Known-unknown data augmentation strategies for detection of logical access, physical access and speech deepfake attacks: Asvspoof 2021," *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, pp. 29–36, 2021.

[15] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *ICASSP 2022-2022 IEEE*. IEEE, 2022, pp. 6367–6371.

[16] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee *et al.*, "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101114, 2020.

[17] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans *et al.*, "Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection," in *ASVspoof 2021 Workshop-Automatic Speaker Verification and Spoofing Coutermeasures Challenge*, 2021.

[18] X. Liu, M. Liu, L. Wang, K. A. Lee, H. Zhang, and J. Dang, "Leveraging positional-related local-global dependency for synthetic speech detection," in *ICASSP 2023-2023*. IEEE, 2023, pp. 1–5.

[19] D. Berrebbi, J. Shi, B. Yan, O. López-Francisco, J. D. Amith, and S. Watanabe, "Combining spectral and self-supervised features for low resource speech recognition and translation," *arXiv preprint arXiv:2204.02470*, 2022.

[20] S. Peng, W. Guo, H. Wu, Z. Li, and J. Zhang, "Fine-tune pretrained models with multi-level feature fusion for speaker verification," in *Proc. Interspeech*, 2024, pp. 2110–2114.

[21] Z. Jin, L. Lang, and B. Leng, "Wave-spectrogram cross-modal aggregation for audio deepfake detection," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[22] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, and K. Böttinger, "Does audio deepfake detection generalize?" *arXiv preprint arXiv:2203.16263*, 2022.

[23] K. Nam, Y. Kim, J. Huh, H. S. Heo, J.-w. Jung, and J. S. Chung, "Disentangled representation learning for multilingual speaker recognition," *arXiv preprint arXiv:2211.00437*, 2022.

[24] S. H. Mun, M. H. Han, M. Kim, D. Lee, and N. S. Kim, "Disentangled speaker representation learning via mutual information minimization," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2022, pp. 89–96.

[25] A. K. S. Yadav, K. Bhagtani, D. Salvi, P. Bestagini, and E. J. Delp, "Fairssd: Understanding bias in synthetic speech detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4418–4428.

[26] N. M. Müller, P. Kawa, W. H. Choong, E. Casanova, E. Gölge, T. Müller, P. Syga, P. Sperl, and K. Böttinger, "Mlaad: The multi-language audio anti-spoofing dataset," in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–7.