LREC 2026

Speech Language Models in Low-Resource Settings: Performance, Evaluation, and Bias Analysis (SPEAKABLE) @ LREC 2026

Workshop Proceedings

Editors
Nina Hosseini-Kivanani, Alessio Brutti, Marco Matassoni, Sandipana Dowerah, Davide Liga

11 May 2026
Palma, Mallorca (Spain)

Proceedings of Speech Language Models in Low-Resource Settings: Performance, Evaluation, and Bias Analysis (SPEAKABLE) @ LREC 2026

ISBN 978-2-493814-83-8
EAN 9782493814838

Preface

Welcome to the SPEAKABLE 2026 Workshop on *Speech Language Models in Low-Resource Settings: Performance, Evaluation, and Bias Analysis*.

SPEAKABLE 2026 is a full-day workshop co-located with LREC 2026. It brings together researchers and practitioners working on speech-native language models, with a particular focus on low-resource languages, dialects, and speaker communities. The workshop was created in response to a growing need in the speech and language technology community: while speech foundation models and Speech LLMs have advanced rapidly, their benefits remain unevenly distributed across languages, domains, devices, and user groups.

Low-resource speech settings continue to face persistent constraints related to limited training data, uneven annotation quality, scarce evaluation benchmarks, and restricted computational budgets. These difficulties are often intensified by real-world deployment conditions, including accent and dialectal variation, channel and microphone mismatch, spontaneous speech phenomena, code-switching, noisy environments, and limited availability of lexicons or grapheme-to-phoneme resources. As a result, systems that appear strong under standard benchmark conditions may still fail to provide reliable, fair, and useful performance for many underrepresented communities.

The goal of SPEAKABLE 2026 is to provide a focused forum for addressing these challenges through three closely connected strands. The first strand is **efficient adaptation**, including parameter-efficient fine-tuning, multilingual transfer, knowledge distillation, and edge- or streaming-constrained inference for low-resource speech tasks. The second strand is **meaningful evaluation**, with an emphasis on moving beyond aggregate scores such as WER toward task-appropriate metrics, calibration, robustness analysis, abstention, and slice-aware reporting by accent, dialect, channel, speaker group, and speaking style. The third strand is **responsible practice**, treating bias analysis, data documentation, synthetic-data disclosure, privacy, and safety considerations as routine parts of scientific reporting rather than optional additions.

The call for papers welcomed work on efficient adaptation of Speech LLMs for low-resource languages, evaluation methods for ASR, spoken language understanding, and speech generation, robustness under domain shift, cascaded versus end-to-end error propagation, low-resource corpus creation, lexicon and G2P development, and ethics-by-default reporting. We especially encouraged submissions that combine methodological innovation with strong empirical evidence, transparent documentation, calibrated uncertainty, and, where possible, openly released resources or code.

This first edition of SPEAKABLE reflects the increasing importance of speech technologies for the long tail of languages and communities. The accepted contributions address a broad range of topics, including multilingual and cross-lingual modelling, low-resource adaptation, evaluation design, robustness analysis, data-centric approaches, and practical deployment challenges. Together, they show that progress in speech technology cannot be measured only by performance on high-resource benchmarks. It must also be assessed by how reliably systems work under realistic constraints, how transparently they are evaluated, and how equitably they serve diverse speakers.

The workshop program includes oral and poster presentations, as well as an invited talk by Jordi Luque from Telefónica Research. We are grateful to our Program Committee, consisting of confirmed domain experts from academia and industry, for their careful and constructive reviews. Their work was essential in shaping a balanced and high-quality program. We also thank all authors for submitting their work, revising their papers, and contributing to the scientific scope of this first edition.

We hope that SPEAKABLE 2026 will serve not only as a venue for presenting current research, but also as a catalyst for collaboration, shared evaluation practices, and community-building around inclusive speech technologies. Our broader aim is captured by the workshop's guiding message: build strong models, measure what matters, and make bias analysis routine for speech in the long tail.

Finally, we thank the LREC 2026 workshop chairs and organizers for hosting SPEAKABLE as part of the LREC 2026 workshop program. We also thank our invited speaker, reviewers, authors, participants, and supporting institutions for helping make this first edition possible.

Further information about the workshop, including the program and updates, is available on the SPEAKABLE 2026 website: `https://speakable-2026.github.io/`.

The SPEAKABLE 2026 Organizing Committee

Organizing Committee

- Nina Hosseini-Kivanani, Radio Télévision Luxembourg & University of Luxembourg, Luxembourg
- Alessio Brutti, Fondazione Bruno Kessler, Italy
- Marco Matassoni, Fondazione Bruno Kessler, Italy
- Sandipana Dowerah, Tallinn University of Technology, Estonia
- Davide Liga, University of Luxembourg, Luxembourg
- Christoph Schommer, University of Luxembourg, Luxembourg

Program Committee

- Badr M. Abdullah, Saarland University, Germany
- Şeymanur Aktı, Karlsruhe Institute of Technology (KIT), Germany
- Tanel Alumäe, Tallinn University of Technology, Estonia
- Dimitra Anastasiou, Luxembourg Institute of Science and Technology, Luxembourg
- Leonardo Badino, Almawave, Italy
- Stefano Bannò, University of Cambridge, UK
- Carlos Carvalho, INESC-ID, Portugal
- Shammur Absar Chowdhury, Qatar Computing Research Institute (QCRI), Qatar
- Matt Coler, University of Groningen, Netherlands
- Lorenzo Concina, Fondazione Bruno Kessler (FBK), Italy
- Miguel Couceiro, Universidade de Lisboa, Portugal
- Rohan Kumar Das, Fortemedia, Singapore
- Ioannis Douros, Stavros Niarchos Foundation (SNF), Greece
- Yassine El Kheir, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany
- Daniele Falavigna, Fondazione Bruno Kessler (FBK), Italy
- Saeed Farzi, Fondazione Bruno Kessler (FBK), Italy
- Maxime Fily, Inalco (Institut national des langues et civilisations orientales), France
- Peter Gilles, University of Luxembourg, Luxembourg
- Nabarun Goswami, University of Tokyo, Japan
- Felix Herron, Université Paris Dauphine – PSL, France
- Aditya Joshi, UNSW Sydney, Australia
- Joonas Kalda, Pyannote.ai, France
- Ajinkya Kulkarni, Idiap Research Institute, Switzerland
- Spyretta Leivaditi, University of Groningen, Netherlands
- Damien Lolive, IUT of Vannes (University of South Brittany), France
- Keerthana Murugaraj, University of Luxembourg, Luxembourg
- Maria Onoeva, Charles University, Czech Republic
- Ludovica Pannitto, University of Bologna, Italy
- Fernando Perez Tellez, Technological University Dublin, Ireland
- Ben Peters, INESC-ID & Instituto Superior Técnico, Portugal

- Fred Philippy, SnT, University of Luxembourg, Luxembourg
- Bornali Phukon, University of Illinois Urbana-Champaign, USA
- Tina Raissi, RWTH Aachen University, Germany
- Thomas Rolland, Orange, France
- Beatrice Savoldi, Fondazione Bruno Kessler (FBK), Italy
- Imran Sheikh, Vivoka, France
- Ravi Shekhar, University of Essex, UK
- Golshid Shekoufandeh, University of Amsterdam, Netherlands
- Francisco Teixeira, INESC-ID, Portugal
- Hawau Olamide Toyin, Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), UAE
- Preben Vangberg, Bangor University, UK
- Jelena Vasić, Technological University Dublin, Ireland
- Martijn Wieling, University of Groningen, Netherlands
- Enrico Zovato, Almawave, Italy
- Juan Pablo Zuluaga-Gomez, AGIGO, Switzerland

Invited speaker

- Jordi Luque, Telefónica, Spain

Table of Contents

Workshop Program

May 11, 2026

Room: W11

09:00–09:20 **Introduction and general remarks**

09:20–10:20 **Invited speaker: Jordi Luque**

10:20–10:30 **Remote posters**

Adapting Foundational ASR Models to Efik: An Empirical Study of an Extremely Low-Resource Tonal Language
Offiong Bassey Edet, Stephen Orok Duke, Enoima Essien Umoh, Benjamin Okon Nyong and Andrew Asuquo Nkpanam

PAREDA: A Multi-Accent Speech Dataset of Natural Language Processing Research Discussions
Sicheng Jin, Dipankar Srirag and Aditya Joshi

10:30–12:00 **Coffee Break and Poster session**

Say Again? The Limits of Whisper with Conversation. A Case Study on the KIParla Corpus.
Martina Simonotti, Ludovica Pannitto, Caterina Mauri, Adriano Ferraresi and Gabriele Carioli

Not All Polar Questions Are the Same: ASR, Humans, and Russian
Maria Onoeva

Quantizing Whisper: How Design Choices Affect ASR Performance
Arthur Söhler, Julian Irigoyen and Andreas Søeborg Kirkedal

"OK Aura, Be Fair with Me": Demographics-Agnostic Training for Bias Mitigation in Wake-up Word Detection
Fernando López, Paula Delgado-Santos, Pablo Gómez, David Solans and Jordi Luque

Scalable Expansion of Multilingual Speech LLMs for ASR: A Continual Learning Approach
Lorenzo Concina, Marco Matassoni and Alessio Brutti

May 11, 2026 (continued)

Responsible Benchmarking of Fairness for Automatic Speech Recognition
Felix E. Herron, Ange Richard, François Portet, Alexandre Allauzen and Solange Rossato

Addressing Accent Disparities in Automatic Speech Recognition: A Comparative Study of Single and Two-Step Adaptation
Mykhailo Danilevskyi, Fernando Perez-Tellez and Jelena Vasic

Investigating Speaker Pronunciation Variability in Speech Embeddings: Speaker and L1 Effects on French as a Second Language
Maxime Fily, Martine Adda-Decker and Guillaume Wisniewski

What LID Systems Say About Dialectal Variation. The Case of Yiddish, Quechua and Mande
Johanna Cordova, Eric Jordan and Valentina Fedchenko

HARNESS: Lightweight Distilled Arabic Speech Foundation Models
Vrunda Nileshkumar Sukhadia and Shammur Absar Chowdhury

When Does OmniASR Fail? A Fine-Grained Human Evaluation on Saudi Arabic Dialects
Hend Al-Khalifa

12:00–13:00 **Spotlight papers**

SpeechLM for Automatic Speech Recognition in Low-resource Languages
Md Abdur Razzaq Riyadh, Eneko Agirre, Eva Navas and Claudia Borg

Improving Low-resource ASR Using Bilingual Fine-tuning with Language Identification: A Cross-linguistic Evaluation
Reihaneh Amooie, Yun Hao, Wietse de Vries, Jelske Dijkstra, Matt Coler and Martijn Wieling

May 11, 2026 (continued)

13:00–14:00 **Lunch break**

14:00–16:00 **Architectures and learning methods**

May 11, 2026 (continued)

16:00–16:30	**Coffee break**
16:30–17:00	**Best paper and closing remarks**

Adapting Foundational ASR Models to Efik: An Empirical Study of an Extremely Low-Resource Tonal Language

Offiong Bassey Edet[1,4], Stephen Orok Duke[1], Enoima Essien Umoh[1], Benjamin Okon Nyong[2], Andrew Asuquo Nkpanam[3]
[1]University of Cross River State, [2]Arthur Jarvis University, [3]University of Calabar, [4]ML Collective
offiongbassey99@gmail.com, orokduke2003@unicross.edu.ng,
enoimaumoh@unicross.edu.ng, benokon26@gmail.com, drewsuqi6@gmail.com

Abstract

Automatic Speech Recognition (ASR) has significantly transformed human-computer-interaction and natural language processing. However, many African spoken languages, including Efik, remain severely underrepresented in ASR research. This paper investigates the adoption of state-of-the-art foundational ASR models such as XLS-R and Whisper through fine-tuning for Efik, a low-resource tonal language and empirically evaluates their performance. We curate a 3-hour Efik speech dataset and conduct a comparative evaluation using standard ASR metrics. We further augmented the XLS-R CTC model with a 3-gram KenLM language model trained on an Efik text corpus. Experimental results show that XLS-R-300M + KenLM achieves a word error rate (WER) of 10.86% and a character error rate (CER) of 3.16%, substantially outperforming both the baseline XLS-R (WER: 29.2%, CER: 6.4%) and Whisper across noisy and multi-speaker conditions. These findings suggest that lightweight CTC models augmented with language model integration offer a more robust and practical approach for extremely low-resource tonal languages than larger sequence-to-sequence models.

Keywords: Automatic Speech Recognition, Low-Resource Language, Efik, Tonal Language, Transfer Learning

1. Introduction

Automatic Speech Recognition (ASR) has significantly advanced Natural Language Processing (NLP), particularly with the emergence of state-of-the-art (SOTA) models. Recent deep learning-based ASR systems, including self-supervised encoder-only models such as XLS-R (Babu et al., 2021) and encoder-decoder architectures such as Whisper (Radford et al., 2022) are trained on hundreds of thousands of speech data that have demonstrated strong performance in transcribing spoken language into text. Beyond transcription, these models have enabled the development of robust virtual assistants, real-time speech-to-speech translation systems and language learning tools.

Despite these advances, the benefit of modern ASR systems have been largely concentrated on high-resource languages such as English and Mandarin (Imam et al., 2025). In contrast, many African languages, spoken by over one billion people (Adelani et al., 2022), remain severely underrepresented. Africa is home to over 2,000 languages (Abbott and Martinus, 2019), with Nigeria alone accounting for more than 1,000 distinct languages, among which Efik is included.

Efik, spoken by 1.5 million native speakers and about 3 million second-language speakers (Mensah and Mensah, 2014), has received little attention in ASR. As a tonal language with morphological complexity and limited publicly available speech resources and linguistic tools, Efik presents unique challenges that have contributed to its underrepresentation in both ASR and NLP research more broadly.

This study investigates the adaptation of powerful SOTA ASR models to Efik and empirically evaluates their performance in a low-resource setting.

2. A Brief Discussion on Efik Language

Efik is a Benue-Congo language belonging to the Niger-Congo family. It is primarily spoken in Cross River State and parts of Akwa Ibom State in south-eastern Nigeria, and is also spoken in the South Western part of Cameroon (Offiong and Ansa, 2013). Efik language is amongst the earliest Nigerian languages to be written and studied in Nigeria. Its orthography, lexicon and grammatical structure were largely developed through early missionary linguistic efforts (Offiong and Ansa, 2013).

As a tonal language, Efik employs pitch variations to distinguish lexical and grammatical meaning, which introduces additional challenges for automatic speech recognition systems. Combined with its morphological complexity and limited availability of publicly accessible speech resources, these characteristics contribute to the underrepresentation of Efik in ASR research.

Proceedings of SPEAKABLE @ LREC 2026, pages 1–7

3. Related Work

Recent studies have explored ASR development for African and other low-resource languages. Rufai et al. (2020) developed an end-to-end ASR system for Nigeria Pidgin English, demonstrating improvement using SOTA models such as Nemo Quartznet, Wav2Vec2.0 Base-100H, and Wav2Vec XLS-R-Large-53, achieving a word error rate (WER) of 28.6% from Wav2Vec2.0 XLS-R-Large-53.

Similarly, Chanie et al. (2023) developed ASR systems for the three East African languages - Kinyarwanda, Swahili and Luganda trained on 3,900 hours of code switched speech data, achieving competitive WERs across all languages.

Benchmarking efforts by Nahabwe et al. (2025) empirically evaluates four SOTA ASR models - Whisper, XLS-R, MMS, and wav2bert - across 13 African languages using a transcribed dataset of approximately 400 hours, highlighting the strength and limitation of each model in low-resource setting.

Other notable ASR research efforts in Africa have focused on major Nigerian languages such as Yoruba, Hausa, and Igbo, supported by relatively larger speech datasets (Tolúlọ́pẹ̀ Ògúnrèmí, 2024; DeRenzi et al., 2025). However, Efik remains largely absent from these studies.

To the best of our knowledge, no prior work has addressed automatic speech recognition for Efik, making this study the first empirical investigation of Efik ASR.

4. Efik Language Dataset Curation

The dataset used for this study consists of self-recorded speech data collected from a single native Efik speaker. It comprises 2,632 clips with a total duration of approximately 3 hours. The recording materials were drawn primarily from educational texts, folktales and storybooks to ensure linguistic diversity and natural sentence structure.

All recordings were conducted in a quiet environment using a wireless microphone to minimize background noise and ensure consistent audio quality.

4.1. Labeling and Validation of Audio Recordings

Due to the unavailability of existing Efik speech technologies, a hybrid method was adopted for dataset labeling and validation. First, complete audio recordings with their corresponding transcripts were collected. A custom python script was developed to segment the recordings into smaller utterances. The segmentation process relied on short, natural pauses in speech to avoid generating overly long clips, with a maximum clip duration capped at 16 seconds. Additionally, a padding of 120 milliseconds was applied at the beginning of each clip to prevent truncation of initial phonemes and to reduce overly aggressive splitting by the segmentation algorithm.

Second, dataset validation was performed entirely through manual inspection, ensuring that each audio clip aligned with its corresponding transcription. Although forced alignment tools were considered, existing SOTA ASR models including Whisper, wav2vec2.0 and XLS-R - performed poorly on Efik speech and were therefore unsuitable for reliable alignment.

4.2. Statistical Analysis of Dataset

The dataset contains 2,632 utterances with a total duration of 3.08 hours. Clips lengths range from 0.48s to 15.94s with an average duration of 4.21s and a median of 3.49s. In Figure 1, the y-axis represents the number of clips and the x-axis represents the duration in seconds.

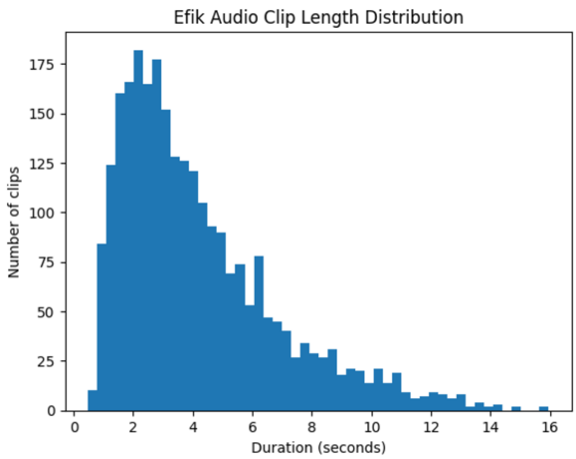


Figure 1: Distribution of audio clips duration.

The corresponding text corpus consists of 23,986 word tokens with a vocabulary size of 3,276 unique words. The average utterance length is 9.11 words, with transcripts ranging from 1 to 33 words. In Figure 2, the y-axis represents the number of clips and the x-axis represents the number of words per utterance.

4.3. Dataset Preprocessing

All transcripts were normalized using Unicode NFC normalization to ensure consistent character representations. Punctuation marks not corresponding to acoustic events were removed. Following standard CTC-based ASR practice, whitespace was replaced with a dedicated word bounding symbols (|) to facilitate alignment and decoding. The dataset

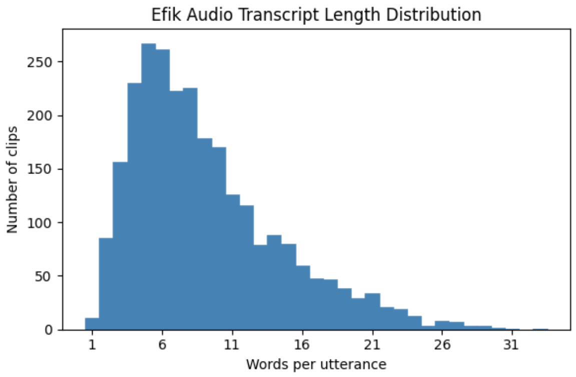


Figure 2: Distribution of transcript length per audio clip.

was partitioned into training and test sets as summarized in Table 1.

Split	Utterances
Train	2,368
Validation	264
Test	393

Table 1: Dataset split statistics.

4.4. Data Availability

We have open-sourced the dataset used in this study. It is publicly available at `https://huggingface.co/datasets/offiongbassey/efik_audio_dataset` (offiongbassey/efik_audio_dataset).

5. Methodology

5.1. Model Selection

Transfer learning from high resource languages has been widely shown to be an effective approach for improving end-to-end ASR performance in low-resource settings (Xu et al., 2016; Imam et al., 2025; Olatunji et al., 2023). In this study, we select two state-of-the-art (SOTA) ASR models - XLS-R and Whisper for fine-tuning and evaluation of Efik speech.

Both models provide multiple pretrained checkpoints ranging from small to large. Due to limited computational resources and relatively small size of available Efik speech data, we adopt smaller checkpoints for both architectures. Leveraging cross-lingual transfer learning is both theoretically and practically preferable to training from scratch in such low-resource scenarios.

XLS-R (Babu et al., 2021) is a large-scale multilingual speech representation extending wav2vec2.0 (Baevski et al., 2020). It was pretrained on approximately 436,000 hours of unlabeled speech across 128 languages, including several African languages. The pretraining data includes Swahili (91 hours), Yoruba (75 hours), Zulu (56 hours), Lingala (72 hours), Kinyarwand (1,199 hours) and Afrikaans (87 hours), among others, making XLS-R well suited for cross-lingual adaption for African languages.

Whisper (Radford et al., 2022) is an encoder-decoder ASR model trained on 680,000 hours of multilingual and multitask speech data covering approximately 96 languages. Its sequence-to-sequence and large-scale weakly supervised training enable strong robustness to noise, speaker variation and domain mismatch, often requiring minimal fine-tuning for reasonable performance.

5.2. Language Model Integration

CTC-based models such as XLS-R decode speech by independently predicting the most likely token at each time step, without explicit modelling of linguistic context. To address this limitation, we augment the fine-tuned XLS-R model with a 3-gram KenLM language model (Heafield, 2011) trained on an Efik text corpus of 166,977 tokens and 15,662 unique word types. The language model was integrated into the decoding pipeline using pyctcdecode with beam search (beam width=100), a language model weight of $\alpha = 0.5$ and a word insertion bonus of $\beta = 1.0$. This approach allows the decoder to favour linguistically plausible word sequences during inference, without requiring additional annotated speech data.

5.3. Evaluation Metrics

To empirically evaluate the performance of the fine-tuned SOTA ASR models, we used Word Error Rate (WER), a widely adopted metrics for measuring accuracy in ASR systems. WER means the number of substitutions, deletions and insertions required to transform a hypothesis transcription into a reference transcription (Park et al., 2008).

$$\text{WER} = \frac{S+D+I}{N} = \frac{S+D+I}{S+D+C} \quad (1)$$

Where S is the number of substitutions; D is the number of deletions; I is the number of insertions; C is the number of correct words; and N is the number of words in the reference (N = S + D + C).

In addition to WER, we report Character Error Rate (CER) which measures the percentage of incorrect characters between the hypothesis and reference text. Unlike WER, which struggles with morphological complex languages, CER has been shown to correlate well with human judgements in multilingual evaluations and even handles unclear word boundaries (K et al., 2024).

$$\text{CER} = \frac{S + D + I}{N} \quad (2)$$

Where S is the number of substitutions; D is the number of deletions; I is the number of insertions and N is the number of words in the reference.

To assess semantic robustness beyond WER and CER, we adopt a Machine Translation (MT) based evaluation pipeline (Figure 3), where Efik ASR outputs are translated into English and evaluated using BLEU, ChrF and human judgement. This approach better reflects real-world usage, particularly under domain and speaker mismatch.

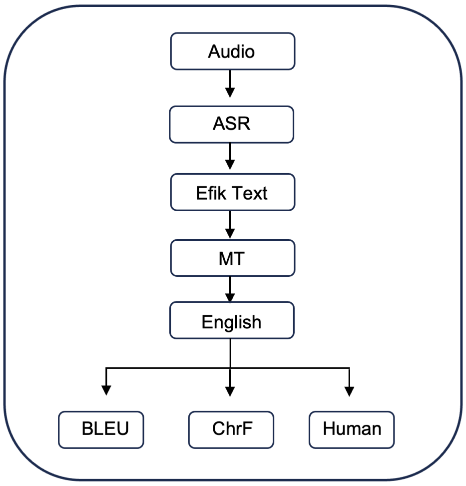


Figure 3: MT - Based Semantic Evaluation Pipeline for Efik ASR

5.4. Experiments

We fine-tuned XLS-R and Whisper using 3 hours of self-recorded Efik speech data, leveraging cross-lingual transfer learning. Each model was trained for 12 epochs.

For XLS-R, we used a learning rate of 3×10^{-4} with a batch size of 4, while Whisper was fine-tuned using a learning rate of 1×10^{-5} with a batch size of 8. All experiments were conducted on NVIDIA A100 with high-ram using the Google Colab Pro+ platform.

The AdamW optimizer was employed with a linear learning rate warm-up of 500 steps. Furthermore, we used the default dropout configuration of the pretrained models during fine-tuning. Mixed precision (FP16) training and gradient checkpoint were enabled to reduce memory usage.

During training, both models exhibited a steady decrease in training and validation loss, indicating progressive learning and effective generalization despite the limited size of the dataset.

6. Results

The results of the study using the two SOTA models - XLS-R and Whisper are presented below:

6.1. Quantitative Analysis

Table 2 presents the performance of both models evaluated on speech from the same speaker in a quiet environment. Both models demonstrate measurable improvements in automatic speech recognition (ASR) for Efik, a low-resource tonal language, despite the limited amount of available training data. The fine-tuned XLS-R-300M model achieves lower WER and substantially lower CER compared to the Whisper baseline. Adding a 3-gram KenLM language model further improves performance, yielding the lowest WER and CER overall.

Model	**WER**	**CER**
Whisper	37.5%	28.9%
XLS-R-300M	29.2%	6.4%
XLS-R + KenLM 3-gram	**10.86%**	**3.16%**

Table 2: ASR performance comparison on the Efik test set.

6.2. Qualitative and Semantic Evaluation beyond WER and CER

Quantitative metrics such as WER and CER provide a limited view of transcription quality, particularly concerning semantic preservation. To complement these measures, we conduct a qualitative evaluation across diverse acoustic and speaker scenarios, including single-speaker clean speech and multi-speaker noisy environments. This analysis reveals the interplay between phonetic precision and contextual reasoning, elucidating the trade-offs observed in our automated semantic metrics (BLEU/ChrF).

Table 3 presents representative examples comparing reference transcripts with outputs from Whisper, XLS-R, and XLS-R augmented with a 3-gram KenLM language model, highlighting differences in semantic fidelity and robustness under domain shift. The KenLM-enhanced model demonstrates improved transcription quality in both clean and noisy conditions, producing outputs that more closely align with the reference transcripts, particularly in preserving morphological boundaries and reducing fragmentation.

We further adopt a translation-based semantic evaluation, in which Efik ASR outputs are translated into English and assessed using BLEU and

Scenario	Reference (Source)	XLS-R	XLS-R + KenLM 3-gram	Whisper
Single speaker, quiet environment.	idem eyen esie isọñke, enye odu ke ufọk ibọk.	idem eyen esisọñke enye odu ke ufọk ibọk	idem eyen esie isọñke enye odu ke ufọk ibọk	idem eyen esi isọñke, enye eduk ufọk ibọk.
Multi-speaker, noisy environment.	ndien akpa ñke nyenede edi, oyom ikpọ ñkpọ.	nyenakpa ke eyen ededi ikaa oyom ikpọñ kpọ	ndian akpañ ke enyenede edi yak oyom ikpọ ñkpọ	ndien akpañke enyenede edi, oyom ikpọñkpọ.
Multi-speaker, noisy environment.	ọkọñọ mbañ enañ. Utu ke mbọk osun udi. Yak edim, o, edep.	añ ambae namikpokke mbọk osim di ia edim aaedem	ọkọñọ mbañ enam koko ke mbọk esin udi yak edim edem	Okoñ amaña enañ. Otuk embọk ọsunu di. Ya edim, o, edep.

Table 3: Qualitative semantic comparison of XLS-R, XLS-R + KenLM, and Whisper on Efik ASR under domain and speaker mismatch

ChrF. This approach approximates downstream usability, reflecting scenarios where ASR outputs are consumed by non-Efik speakers or integrated into multilingual pipelines.

As shown in Table 4, XLS-R + KenLM-3-gram achieves the highest BLEU and ChrF scores across both single-speaker and multi-speaker conditions, consistently outperforming both the base XLS-R model and Whisper. Notably, under noisy multi-speaker conditions where Whisper previously showed an advantage in ChrF over the base XLS-R model, the KenLM-enhanced XLS-R surpasses Whisper across all metrics. These results demonstrate that incorporating a language model not only improves phonetic accuracy (as reflected in WER/CER) but also yields substantial gains in semantic preservation, even under challenging acoustic conditions. This underscores the value of leveraging language models for low-resource tonal languages like Efik, where both phonetic and semantic fidelity are critical for downstream usability.

6.3. Discussion

Despite being trained on limited audio data from a single speaker in a quiet environment, both models demonstrate notable gains in Efik automatic speech recognition, highlighting the effectiveness of modern self-supervised and encoder–decoder architectures in extremely low-resource tonal settings.

On a held-out single-speaker evaluation set recorded independently from the training data, the fine-tuned XLS-R-300M model achieved a WER of 29.2% and a CER of 6.4%, outperforming a Whisper baseline trained on the same Efik corpus. The addition of a 3-gram KenLM language model further improved performance significantly, achieving a WER of 10.86% and a CER of 3.16%, a relative reduction of 62.8% in WER and 50.6% in CER over the base XLS-R model. This substantial gain underscores the value of incorporating language model priors, particularly for low-resource tonal languages where phonetic ambiguity is high.

Under real-world noisy and multi-speaker conditions, Whisper exhibits greater robustness and semantic fidelity compared to the base XLS-R model, likely due to its integrated autoregressive language modeling and encoder-decoder architecture, despite Efik's morphological and tonal complexity. However, the KenLM-enhanced XLS-R model bridges this gap considerably, demonstrating improved semantic preservation in translation-based evaluations while maintaining its phonetic accuracy advantages. These findings suggest that effective language model integration, whether through an external n-gram LM or an architecture-internal component, is critical for achieving robust ASR performance in low-resource tonal languages under domain shift.

7. Future Work

Our results indicate that SOTA ASR models can be significantly improved for Efik with more high-quality, multi-speaker, and well-annotated speech data. We plan to curate additional audio spanning multiple domains, speakers ranging from children to adults, and diverse noisy environments. Expanding the dataset beyond the current 3 hours of single-speaker recordings will help the models better capture tonal variations, semantic nuances, and morphological complexity inherent to Efik. Additionally, we will explore data augmentation techniques such as speed and pitch perturbation and noise injection to further enhance robustness and generalization.

8. Conclusion

We evaluated the adaptation of two pre-trained SOTA ASR models, XLS-R and Whisper, for Efik, a tonal language with morphological complexity.

Scenario	Model	BLEU	ChrF	Human Score	Notes
Single speaker, quiet environment.	XLS-R	43.01	74.27	8/10	Very close to reference.
Single speaker, quiet environment.	Whisper	14.21	35.06	7/10	Slight token error, meaning still intact.
Single speaker, quiet environment.	**XLS-R + KenLM**	**59.54**	**80.40**	**9.2/10**	**Very close to reference with minimal token errors**
Multi-speaker, noisy environment.	XLS-R	4.07	13.59	2.8/10	Errors due to unseen speakers and noisy background. Severe degradation in semantic coherence is observed.
Multi-speaker, noisy environment.	Whisper	4.52	28.82	3.6/10	Slight errors in unseen speakers and noisy background, slight semantic preservation and meaning.
Multi-speaker, noisy environment.	**XLS-R + KenLM**	**6.88**	**29.95**	**3.9/10**	**Moderate improvement over base XLS-R; better character-level preservation than Whisper, though semantic coherence remains challenged by domain shift.**

Table 4: Qualitative semantic comparison of XLS-R and Whisper on Efik ASR under domain and speaker mismatch.

Both models were fine-tuned on 3 hours of single-speaker audio (2,632 clips) over 12 epochs each, showing significant improvements in low-resource settings.

The integration of a 3-gram KenLM language model with XLS-R yielded substantial gains across all evaluation conditions. KenLM-enhanced XLS-R achieved the lowest overall WER (10.86%) and CER (3.16%) on the single-speaker test set, outperforming both the base XLS-R model and Whisper. Under noisy multi-speaker conditions, the KenLM-enhanced model also demonstrated improved robustness compared to the base XLS-R model, achieving competitive performance with Whisper while maintaining its phonetic accuracy advantages.

Whisper's ability to model punctuation contributed to better context preservation in certain cases, highlighting the importance of architectural choices in low-resource ASR. Our results demonstrate that combining self-supervised acoustic models with language model integration offers a promising pathway for building robust ASR systems for under-resourced tonal languages.

9. Ethics Statement

We obtained informed consent from all volunteers who participated in the data recording process. The dataset does not contain sensitive, personal, or otherwise violatory content. All recordings were collected and used in accordance with ethical research practices.

10. Acknowledgements

We would like to sincerely thank all the volunteers, native speakers, and linguists who contributed to the recording process and assisted in the evaluation of the results.

11. Bibliographical References

Jade Abbott and Laura Martinus. 2019. Benchmarking neural machine translation for southern african languages. In *Proceedings of the Workshop on Widening NLP*, pages 98–101, Florence, Italy. Association for Computational Linguistics.

David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba O. Alabi, Shamsuddeen H. Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Elvis Mboning, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo L. Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Adeyemi, Gilles Q. Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu Ngoli, and Dietrich Klakow. 2022. Masakhaner 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint*.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint*.

Yonas Chanie, Moayad Elamin, Paul Ewuzie, and Samuel Rutunda. 2023. Multilingual automatic speech recognition for kinyarwanda, swahili, and luganda. In *Conference Proceedings*.

Brian DeRenzi, Anna Dixon, Mohamed Aymane Farhi, and Christian Resch. 2025. Synthetic voice data for automatic speech recognition in african languages. *arXiv preprint*.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Association for Computational Linguistics.

Sukairaj Hafiz Imam, Babangida Sani, Dawit Ketema Gete, Bedru Yimam Ahamed, Ibrahim Said Ahmad, Idris Abdulmumin, Seid Muhie Yimam, Muhammad Yahuza Bello, and Shamsuddeen Hassan Muhammad. 2025. Automatic speech recognition for african low-resource languages: Challenges and future directions. In *Proceedings of the Workshop on African NLP*.

Thennal D K, Jesin James, Deepa P Gopinath, and Muhammed Ashraf K. 2024. Advocating character error rate for multilingual asr evaluation. *arXiv preprint*.

Eyo Mensah and Eyamba Mensah. 2014. The adaptation of english consonants by efik learners of english. *English Language Teaching*, 7(3).

Alvin Nahabwe, Sulaiman Kagumire, Denis Musinguzi, Bruno Beijuka, Jonah Mubuuke Kyagaba, Peter Nabende, Andrew Katumba, and Joyce Nakatumba-Nabende. 2025. Benchmarking automatic speech recognition models for african languages. *arXiv preprint*.

Offiong Ani Offiong and Stella Ansa. 2013. The efik language: A historical profile. *Research in Humanities and Social Sciences*, 3(6).

Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, et al. 2023. Afrispeech-200: Pan-african accented speech dataset for clinical and general domain asr. *Transactions of the Association for Computational Linguistics*, 11:1669–1685.

Youngja Park, Siddharth Patwardhan, Karthik Visweswariah, and Stephen C. Gates. 2008. An empirical analysis of word error rate and keyword error rate. In *INTERSPEECH 2008*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint*.

Amina Mardiyyah Rufai, Afolabi Abeeb, Esther Oduntan, Tayo Arulogun, Oluwabukola Adegboro, and Daniel Ajisafe. 2020. Towards end-to-end training of automatic speech recognition for nigerian pidgin. *arXiv preprint*.

Anuoluwapo Aremu Tolúlópè Ógúnrèmí, Kòlá Túbòsún. 2024. Ìròyìnspeech: A multi-purpose yorùbá speech corpus. In *Proceedings of LREC 2024*.

Haihua Xu, Hang Su, Chongjia Ni, Xiong Xiao, Hao Huang, Eng-Siong Chng, and Haizhou Li. 2016. Semi-supervised and cross-lingual knowledge transfer learnings for dnn hybrid acoustic models under low-resource conditions. In *INTERSPEECH 2016*.

PAREDA: A Multi-Accent Speech Dataset of Natural Language Processing Research Discussions

Sicheng Jin, Dipankar Srirag, Aditya Joshi
University of New South Wales
{stefan_zalkoszin.jin, d.srirag, aditya.joshi}@unsw.edu.au

Abstract

While modern Automatic Speech Recognition (ASR) systems achieve high accuracy on benchmark corpora, their performance often degrades when there is real-world variability. This work focuses on variability arising due to accented, spontaneous, and domain-specific speech. In particular, we introduce PAper REading DAtaset (PAREDA), a first-of-its-kind multi-accent speech dataset consisting of discussions on academic Natural Language Processing (NLP) papers between speakers with Australian, Indian-English, and Chinese English accents. Each session elicits a spontaneous monologue (a summary of a paper's abstract) and a non-monologue (a question-and-answer session between participants), resulting in a corpus rich with technical jargon and conversational phenomena. We evaluate the performance of SOTA ASR models on PAREDA, analysing the impact of accent mixing and increased speech rate. Our results show that, in the zero-shot setting, models perform worse, confirming the dataset's challenging nature. However, fine-tuning on PAREDA significantly reduces the Word Error Rate (WER), demonstrating that our dataset captures linguistic characteristics often missing from existing corpora. PAREDA serves as a valuable new resource for building and evaluating more robust and inclusive ASR systems for specialised, real-world applications.

Keywords: Accents, World Englishes, Dialogue, Research Discussions

1. Introduction

Automatic speech recognition (ASR) models are increasingly deployed in academic settings such as lecture transcription, workplace meetings, and conference presentations. As these settings become more diverse, ASR systems are required to handle a wide range of linguistic phenomena, including accented speech, dialectal variation, and domain-specific terminology (Mehrish et al., 2023). While recent ASR models achieve strong performance on general-domain monologue speech for mainstream accents, studies have shown that state-of-the-art ASR models such as Whisper (Radford et al., 2023a) exhibit degraded transcription performance for non-mainstream English accents, including Nigerian English (en-NG) and Indian English (en-IN), when compared to mainstream American English (en-US) (Eisenstein et al., 2023). This evaluation is for general-domain dialogue.

Academic conversations are inherently domain-specific and frequently involve speakers with diverse linguistic backgrounds, as in the seminal AMI meeting corpus (Kraaij et al., 2005). To our knowledge, there is currently no publicly available resource designed to evaluate ASR performance on domain-specific academic speech involving non-mainstream English accents. In this work, we describe the creation of PAREDA (which stands for PAper REading DAtaset), a small-scale multi-accent speech dataset comprising 3.9 hours of recorded audio. The dataset consists of conversations between pairs of speakers with non-mainstream English accents, specifically Australian, Chinese, and Indian English. Speakers exhibit varying levels of expertise in natural language processing, ranging from students to researchers. The conversations center on discussions of NLP research papers sourced from the ACL Anthology [1]. This design reflects realistic academic interactions, where the use of technical terminology varies according to a speaker's familiarity with the subject matter, resulting in differing densities of domain-specific vocabulary across utterances. Our evaluation on PAREDA highlights that ASR models like Whisper (Radford et al., 2023a), Phi-4 (Abouelenin et al., 2025), and CrisperWhisper (Zusag et al., 2024) are able to achieve SOTA accuracy, but certain conditions of the input audio have to be met. The domain choice (academic NLP discussions) and its intersection with accent variability make PAREDA a useful dataset for future research.

2. Data Collection

We collect data of speech recordings, using the methodology illustrated in Figure 1. We use NLP research papers collected from ACL Anthology to create our dataset. We cover *three* locales, with *one* participant from each locale. Due to the nature of this study, we limit our participant group to *one* speaker per accent. Each speaker is given 21 papers for the elicitation exercise.

[1] https://www.aclanthology.org/

Proceedings of SPEAKABLE @ LREC 2026, pages 8–15
11 May 2026.

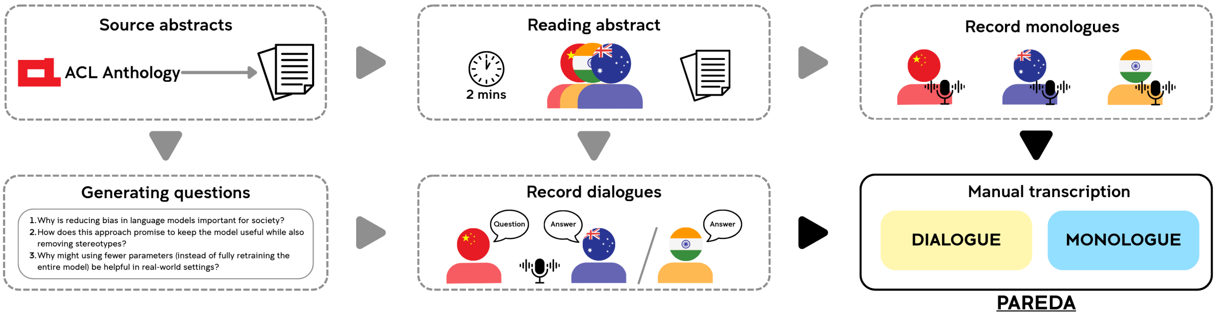


Figure 1: Methodology for dataset collection

2.1. Speakers and Prompts

We conduct elicitation with three participants, one for each locale. The three locales covered in this dataset are: Australian (en-AU), Indian (en-IN), Northern Chinese (en-ZH). We did not collect American (en-US) samples as there is already an excessive amount of en-US speech samples available in other datasets, and the models we use have already exposed to such audio extensively during training. All speakers have experience in NLP research and were willing to engage in recording for this research. The Indian speaker is an NLP lecturer with extensive research expertise in the academic field, and the other two speakers are NLP research students. All participants are also above 18 years of age, with native or superior proficiency in English[2]. We prepared a corpus of 21 NLP research papers on the ACL Anthology website which, according to their focus, can be broadly categorised as: (A) NLP in applied linguistics, (B) NLP in linguistic research, (C) Mitigating NLP bias, (D) NLP in historical and cultural linguistics.

2.2. Recording

The speech samples are either collected online or in person in an indoor environment in a closed meeting room. During each recording session, the participant is given two minutes to read the abstract. If the abstract was found to be non-informative, the participant is allowed to read the full body and summarise the paper. Following this, we also record discussion sessions involving a host asking questions. These questions are specific to each paper, and were prepared to elicit utterances and guide conversation between the host and the participant. This discussion session lasts for upto five minutes. Figure 1 demonstrates an example of this procedure. We manually segment the audio files to less than 30 seconds to ensure the audio is not split abruptly mid-sentence. This is done as some models experimented in this paper only support audio samples of less than 30 seconds. The final audio files are saved in the .wav format. We denote the summarisation session of the speech sample as *monologue*, and the discussion sessions as the *dialogue*. The *dialogue* subset does not involve full-duplex speech; rather, it consists of structured question–answer exchanges.

[2]We determine the proficiency of non-native speakers using English language tests like IELTS.

	en-AU	en-IN	en-ZH
Monologue	16:19	36:22	64:24
Non-Monologue	75:21	41:58	–
Total	91:40	78:20	64:24

Table 1: Speech duration (minutes:seconds) by accent and interaction type. Non-monologue duration reflects respondent speech only; the en-ZH speaker serves as the host during dialogues and is therefore not included as a dialogue respondent.

2.3. Transcription

We first generate raw transcriptions using the ASR model CrisperWhisper (Zusag et al., 2024), which are then manually post-edited. As the models evaluated in this study default to American English spelling conventions, the transcriptions are further standardised to follow the same convention. To assess inter-annotator reliability, an independent annotator post-edited 10 randomly selected transcriptions from the *dialogue* subset. Agreement between annotators was quantified using WER, computed after standard text normalisation. The resulting mean symmetric WER was 2.77%, indicating high agreement between annotators.

2.4. Dataset Statistics

Table 1 summarises the distribution of speech duration across accents and interaction types. Monologue recordings were collected from all three accents. Non-Monologue recordings consist of structured question–answer exchanges in which speakers with en-AU and en-IN accents act as respondents, while the en-ZH speaker serves exclusively as the host and questioner. Consequently, Non-Monologue duration is reported only for en-AU and en-IN accents. The dataset contains 20 en-AU, 23 en-IN, and 39 en-ZH monologue samples, as well as 50 en-AU and 28 en-IN dialogue samples. We split the dataset into training, validation and test with 80:10:10 ratio.

3. Experiment Setup

We fine-tune four model sizes of Whisper in a two-stage training procedure. In the first stage, the models are fine-tuned on GLOBE, a large-scale multi-accent English speech corpus, to improve robustness to accented speech. All GLOBE fine-tuned models are trained for 5,000 steps with an initial learning rate of 1 e-5 and a warmup ratio of 10%. Model checkpoints are evaluated every 500 steps, and the resulting models are subsequently evaluated on our dataset to establish baseline performance prior to in-domain adaptation. In the second stage, the GLOBE-tuned models are further fine-tuned on our collected dataset to adapt the models to domain-specific academic speech. As the in-domain dataset is substantially smaller than GLOBE, we increase the evaluation frequency and employ early stopping to mitigate overfitting. OpenAI's Whisper API, Microsoft's Phi-4 and CrisperWhisper are also used for evaluation, as these are leading models on Hugging Face's Open ASR Leaderboard at the time of experiment. CrisperWhisper was selected as it is specifically tailored for producing verbatim transcripts for casual speech. Whisper API used Whisper Large as the model, while Phi-4 and CrisperWhisper were obtained from HuggingFace. Experiments requiring no fine-tuning were carried out using one Nvidia A100 GPU and all fine-tuning experiments were performed on one Nvidia Tesla Volta GPUs. The total training time for both stages was approximately 70 hours.

4. Results

We address the following questions in our evaluation using PAREDA and other datasets: (a) Would mixed accents worsen the performance of ASR models? (b) Would speech speed influence the performance of ASR models? (c) Can models fine-tuned on non-casual speech perform equally well on casual speech? Table 2 details the performance of these three models under different speech accent and scenario combinations using word error rate (WER) as the metric. For our dataset, there are two major accents: Australian and Indian, designated by en-AU and en-IN; while en-ZH represents Northern Chinese, an accent added into some speech samples of the two major accents. Since en-ZH is used as a condition accent to help assessing other accents, no separate en-ZH samples are evaluated.

4.1. Mixed Accents

Table 2 shows that, model-wise, CrisperWhisper achieves the highest WER across all speech accents and speech types, with the overall lowest WER being 4.08% on Indian speech samples, while Phi-4 performs slightly lower. Whisper API exhibits the lowest scores, producing the highest WER of 18.21% on Australian samples. However, there is no clear evidence of how introducing en-ZH influences WER. We observe that for Phi-4, the WER for both accents are higher when adding en-ZH: Australian WER increases from 8.62% to 8.69% and Indian WER increases 8.96% to 9.15%; however, for the other two models, Australian WER drops considerably with en-ZH. The en-ZH monologue samples score relatively similar to Australian ones, with the whisper WER being slightly lower. Indian samples typically result in lower WER, although this trend does not apply to all models, as observed in the 8.96% result from Phi-4, which is higher than both Australian samples 8.62% and 8.69%. For the other two models, it is clear that Indian monologue samples performs better than Australian monologue ones, while Australian non-monologue samples might show lower WER than the Indian counterparts, as in Whisper API results, or higher WER, as in CrisperWhisper results. Similarly, adding another accent such as en-ZH does not result in a noticeable shift in the WER. Whether the mixed speech would score a higher or lower WER relates to the model used. In our example, the Indian non-monologue samples had a higher WER than the monologue ones, however this pattern reverses of Australian samples. What did not alter is the overall pattern of WER of the models, where CrisperWhisper performs significantly better than the other two.

4.2. Noise Robustness

We see the opportunity to evaluate for robustness to noise by synthetically adding variations. We consider two additional configurations: increasing the speed of the audio and adding background noise. When we accelerate the speech to 1.5 times faster or add a -10dB background white noise, recognition accuracy drops sharply for every model, as shown in Table 2. For accelerated tests, Whisper API's error rate rises from about 18% to 26% on the Australian samples and from 10% to 15% on the Indian samples; Phi-4 shows a similar pattern, jumping to roughly 21% and 16% respectively. CrisperWhisper experienced the worst degradation where its Australian error climbs significantly to 26%, and its

Condition	Model	en-AU	en-AU/ZH	en-IN	en-IN/ZH	en-ZH	en-US
Normal	Whisper API	18.21	15.04	9.56	10.62	15.04	3.91
	Phi4	8.62	8.69	8.96	9.15	8.61	3.82
	CrisperWhisper	5.10	4.29	4.08	4.66	4.38	3.97
1.5x Speed	Whisper API	25.98	23.56	14.76	16.49	20.76	-
	Phi4	20.77	22.37	16.16	21.23	22.98	-
	CrisperWhisper	25.57	25.74	17.05	19.05	22.24	-
-10dB Noise	Whisper API	22.51	19.11	14.65	15.40	21.10	-
	Phi4	14.12	12.80	10.95	13.18	14.98	-
	CrisperWhisper	10.87	12.41	9.51	17.94	27.67	-

Table 2: WER (%) Benchmark Across ASR Architectures Under Varied Linguistic and Environmental Conditions. Note that the last column, en-US, is taken from the Hugging Face Open ASR Leaderboard on the Librispeech-other dataset.

Indian error almost quadruples to 17%. For noise tests, the error rate did not increase to the extent of the accelerated samples, however, still visibly worse than the original normal test results. Whisper's WER rises to 22.5% for Australian samples and from 10% to 14.6% on Indian samples. The other models follow the same pattern. In general, faster delivery or excessive background noise hinders the advantages of the strongest model and narrows the gap between systems. This suggests that none of the three architectures has learned a truly tempo-invariant acoustic representation, and that further speed-augmentation or specialised front-end processing may be required to restore their baseline performance. Similar results are observed for the introduction of background noise. For all three models, combining fast speech with accent mixing still produces error rates in the low-to-mid 20% range, roughly double the figures observed when only the extra speaker is present. Despite the overall degradation, we can observe in Figure 2 that Whisper API remains the least robust, CrisperWhisper the most accurate, and Phi-4 sits in between. The fact that accent diversity has so little impact once speed increases implies that speaking-rate variation is a dominant constraint. Future work should therefore prioritise systematic tempo-augmentation and dynamic time-warping techniques, as they appear more promising than simply expanding accent coverage for enhancing real-world resilience.

4.3. Cross-variety Evaluation

We also conduct a cross-variety evaluation to examine whether models fine-tuned on non-casual datasets can generalize to PAREDA's casual, technical speech, and whether domain-specific fine-tuning provides significant gains. Table 3 compares performance across three stages: zero-shot baseline, GLOBE-tuned, and PAREDA-tuned. In the baseline stage, Whisper Medium achieved the lowest WER (13.46%), notably outperforming the Large model, while the Tiny version performed worst at 22.20%. Surprisingly, Stage 1 fine-tuning with GLOBE resulted in "peculiar" results, as WER actually increased across all model sizes—ranging from a 1.02% rise for Large to 2.98% for Small. Following this, we further fine-tuned the models using PAREDA until WER and validation loss stagnated, as detailed in Section 3. Evaluating these Stage 2 models against the PAREDA test set revealed substantial improvements, with WER for every version dropping by at least 10% compared to previous stages. These findings demonstrate that while broad multi-accent datasets like GLOBE provide breadth, they may not contribute to improved recognition in specialized professional fields. This highlights the necessity of incorporating casual, domain-specific speech during training to help ASR models comprehend the distinct phonological nuances found in expert-level discourse.

4.4. Per-Accent Evaluation

We then perform a per-accent evaluation, fine-tuning GLOBE-tuned models on individual PAREDA accent subsets to observe if targeted tuning reduces the Word Error Rate (WER) for specific accents. We maintained the Stage 1 configuration, using an early stopping patience of 3 and a 1e-5 learning rate. Evaluation steps were set to 250 for Tiny/Small models and 50 for Medium/Large versions. Figures 2 and 3 present the raw WER and relative performance against the all-accent baseline, with rows representing the tuning accent and columns representing the test accent.

For Tiny and Small models, per-accent tuning generally worsened performance compared to the all-accent baseline. For instance, the Indian-tuned Tiny model's WER rose from 11.9 to 12.6 on Indian

Fine-Tuning Stage	Whisper Model Size			
	Tiny	Small	Medium	Large
Baseline (Not Fine-tuned)	22.20	15.03	13.46	15.39
Stage 1 (GLOBE-tuned)	23.95	18.01	15.84	16.41
Stage 2 (PAREDA-tuned)	12.85	6.68	4.53	4.87

Table 3: WER Comparison when fine-tuning Whisper with/without PAREDA.

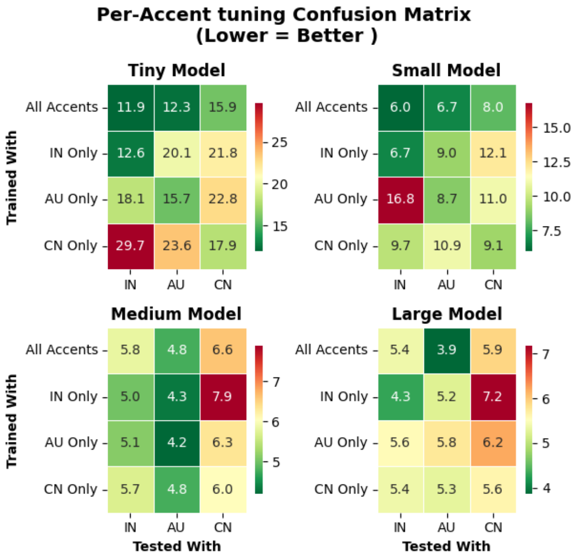


Figure 2: Per-Accent Tuning Results

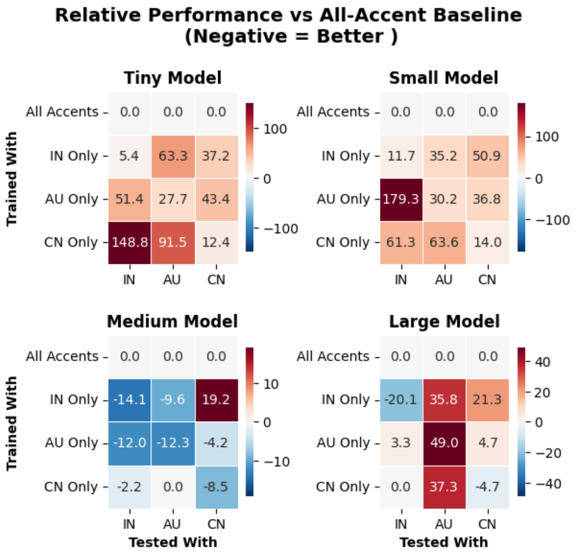


Figure 3: Per-Accent Relative Results

samples, while cross-accent testing (e.g., tuning on Indian/Australian and testing on Chinese) resulted in significantly higher errors. Conversely, per-accent tuning benefited larger models, particularly Medium, where WER decreased across all accents. The Large model showed a contradictory pattern for Australian samples, where tuning on its own accent caused the highest WER increase, though other accents saw modest improvements. These results suggest that while phonology is a factor, model complexity and contextual speech characteristics—such as technical vocabulary and filler words—are equally decisive in ASR performance.

4.5. Qualitative Analysis

Beyond aggregate metrics, we performed a linguistic error analysis on the Whisper-tiny model's PAREDA test performance. Using NLTK and WordNet, we categorised errors at the word level to understand how specific accent conditions influence ASR failure. In order to conduct the analysis, we utilised a sequence alignment approach using Python's `difflib.SequenceMatcher`[3] to identify three types of errors: (1) substitutions (one-to-one word replacements), (2) deletions (words present in reference but missing in prediction), and (3) insertions (words present in prediction but absent in reference). To avoid double-counting, substitutions were identified first through sequence alignment, and only words not involved in substitutions were considered for deletion/insertion analysis.

Each error word was automatically categorised using NLTK's part-of-speech tagging, stopword identification, and WordNet's semantic taxonomy. This approach eliminated manual categorisation bias and provided consistent, reproducible linguistic classifications across all 488 unique error instances identified in the 77-sentence test set. The analysis revealed 84 unique substitutions, 196 deletions, and 208 insertions, totalling 734 errors and 488 distinct occurrences.

The top three semantic classes remain the same, but their relative weight has shifted: (A) **Function words**, which mostly serve grammatical purposes such as *'a', 'do', 'can'* now account for 241 tokens (29.3%), still the single largest source of error. Deletions of articles (*a*, *the*) and prepositions dominate, confirming the model's difficulty with reduced, unstressed items; (B) **Short**

[3] `https://docs.python.org/3/library/difflib.html`

filler tokens (*uh*, *um*, single letters) rose sharply to 131 errors (16%), with a striking 11.9 errors per unique filler, by far the densest error class; **Communication-related terms**, including many NLP words such as *'prompting', 'slang', 'dialogue'* contribute 54 errors (6.6%). We present additional examples fo this class in Table4. Smaller but noteworthy movements are observed in the following cases. **Unclassified/OOV items** climbed to 49 tokens, many are the result of hallucinated insertions rather than deletions, including non-existent words such as *'afimouth'* and *'dissist',* hinting at domain-dependent lexical over-generation, and that accented pronunciations still confuse the ASR. Similarly, **morphological-technical terms** remain rare: only one token ("tokenisation") was mis-recognised, which was consistent with the previous finding that complex morphology is handled reasonably well. Function-word errors break down into deletions 40%, insertions 44%, substitutions 16% (percentages of the 241 tokens). The prevalence of function-to-function substitutions including *a*→*the* and *will*→*for* signals acoustic confusion among weak syllables, though their absolute counts have increased with the larger error pool.

Our targeted word-list comparison covers 54 NLP terms and 46 functional words, as seen in Table 5. Some words are not observed in the test set, and those are not counted towards the results. These include: (A) **Domain words**: 45.8% average error rate (35 errors in 117 appearances); (B) **Grammatical words**: 7.6% (62 errors in 966 appearances). The ratio 6× confirms that technical vocabulary is still markedly harder for the model to capture.

These patterns reveal two primary pathways for ASR failure in multi-accent, technical discourse. First, a "prosodic deletion" pathway affects function words and fillers, where accent-conditioned vowel reduction leads to the omission of unstressed items like articles and prepositions. Second, an "acoustic mis-modeling" pathway targets low-frequency specialist vocabulary. For the 54 NLP terms analysed, the average error rate was 45.8% which is six times higher than the 7.6% rate for grammatical words. This 6× disparity suggests that uncommon phoneme sequences combined with accent shifts overwhelm the recognizer, regardless of its underlying lexical knowledge. Addressing these challenges requires a dual-track approach: prosody-aware modeling to capture variant pronunciations of function words and lexical specialisation to boost the recognition of domain-specific terminology.

5. Related Work

SOTA ASR systems often achieve high accuracy on standard benchmarks but exhibit significant performance degradation across diverse speaker groups, a phenomenon known as ASR bias (Mehrish et al., 2023; Feng et al., 2024). Variations in pronunciation, accent, and intonation remain persistent challenges, often rooted in the lack of diversity within training corpora (Basak et al., 2023; Feng et al., 2024). Recent efforts to address this include the GLOBE corpus, which provides read speech across 164 accents, though it lacks spontaneous conversational nuances (Wang et al., 2024). Conversely, the Multi-Dialect Dataset of Dialogues (MD3) offers task-oriented conversational data but focuses on general-domain tasks like guessing games (Eisenstein et al., 2023). Datasets of meeting corpora either rely on general topics (McCowan et al., 2005) or use TTS to generate audio (Lee et al., 2023). PAREDA addresses a critical gap these corpora overlook: the intersection of dialectal variation and domain-specific, technical vocabulary in spontaneous speech. Unlike read-speech datasets or general-knowledge dialogues, PAREDA targets the challenge of accented speech and technical jargon, which is known to cause robustness issues even in SOTA models like Whisper (Jain et al., 2024; Radford et al., 2023b).

6. Conclusion

We introduced PAper REading DAtaset (PAREDA), a novel multi-accent speech dataset comprising spontaneous monologues and non-monologues from Australian, Indian, and Chinese English speakers discussing technical NLP research papers. The design of PAREDA elicits natural, casual speech laden with domain-specific jargon, creating a challenging and realistic testbed for modern ASR systems. Our experiments demonstrate that a SOTA model, Whisper, yields unsatisfactory Word Error Rates (WER) on the PAREDA corpus in a zero-shot setting, confirming that dialectal and domain-specific speech remains a significant hurdle (Mehrish et al., 2023). The analysis revealed particular difficulties with domain-specific vocabulary, accent mixing, and variations in speaking rate. However, we also showed that finetuning the Whisper model, even with the limited data in PAREDA, leads to considerable performance improvements. This finding aligns with research in other specialised, low-resource domains, such as child speech recognition, where finetuning is a highly effective strategy (Jain et al., 2024). Future work should focus on expanding the PAREDA dataset to include more speakers and a wider variety of global Englishes. PAREDA can be used to investigate the propagation of ASR errors into downstream NLP tasks and serve as a crucial benchmark for developing novel bias mitigation techniques aimed at creating more inclusive speech recognition technologies (Feng et al., 2024).

Reference	Prediction
...just a parser may not be sufficient and there will be other tools such as stemmers ...	...just a **person** may not be sufficient and there will be other tools such as **standard** ...
...it is a low resource language ...	...it is a **only source** language ...
...how many n grams they get right ...	...how many **enzymes** they get **ripe** ...

Table 4: Examples of ASR transcription errors. Mismatches between the Reference (Ground Truth) and Model Prediction are highlighted.

Category	Total Words	Words in Test Set	Total Appearances	Total Errors	Words w/ Errors	Avg. Error Rate
Content (NLP)	54	41	117	35	26	45.80%
Grammatical	46	45	966	62	25	7.57%

Table 5: Linguistic Error Analysis: Technical vs. Functional Vocabulary

7. Bibliographical References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Sneha Basak, Himanshi Agrawal, Shreya Jena, Shilpa Gite, Mrinal Bachute, Biswajeet Pradhan, and Mazen Assiri. 2023. Challenges and limitations in speech recognition technology: A critical review of speech signal processing algorithms, tools and systems. *CMES-Computer Modeling in Engineering and Sciences*.

Siyuan Feng, Bence Mark Halpern, Olya Kudina, and Odette Scharenborg. 2024. Towards inclusive automatic speech recognition. *Computer Speech & Language*, 84:101567.

Rishabh Jain, Andrei Barcovschi, Mariam Yahayah Yiwere, Peter Corcoran, and Horia Cucu. 2024. Exploring native and non-native english child speech recognition with whisper. *IEEE Access*, 12:41601–41610.

Ambuj Mehrish, Navonil Majumder, Rishabh Bharadwaj, Rada Mihalcea, and Soujanya Poria. 2023. A review of deep learning techniques for speech processing. *Information Fusion*, 99:101869.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023a. Robust speech recognition via large-scale weak supervision. 202:28492–28518.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023b. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Mario Zusag, Laurin Wagner, and Bernhad Thallinger. 2024. Crisperwhisper: Accurate timestamps on verbatim speech transcriptions. In *Interspeech 2024*, page 1265–1269. ISCA.

8. Language Resource References

Jacob Eisenstein, Vinodkumar Prabhakaran, Clara Rivera, Dorottya Demszky, and Devyani Sharma. 2023. Md3: The multi-dialect dataset of dialogues. In *INTERSPEECH 2023*.

Wessel Kraaij, Thomas Hain, Mike Lincoln, and Wilfried Post. 2005. The ami meeting corpus. In *Proc. International Conference on Methods and Techniques in Behavioral Research*, pages 1–4.

Keon Lee, Kyumin Park, and Daeyoung Kim. 2023. Dailytalk: Spoken dialogue dataset for conversational text-to-speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

I McCowan, J Carletta, W Kraaij, S Ashby, S Bourban, M Flynn, M Guillemot, T Hain, J Kadlec, V Karaiskos, et al. 2005. The ami meeting corpus. In *Proceedings of Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research*, pages 137–140. Noldus Information Technology.

Wenbin Wang, Yang Song, and Sanjay Jha. 2024. Globe: A high-quality english corpus with global accents for zero-shot speaker adaptive text-to-speech. In *INTERSPEECH 2024*.

Say Again? The Limits of Whisper with Conversation. A Case Study on the KIParla Corpus.

Martina Simonotti, Ludovica Pannitto, Caterina Mauri
Adriano Ferraresi, Gabriele Carioli
University of Bologna, Bologna - Italy
martina.simonotti@studio.unibo.it,
{ludovica.pannitto, caterina.mauri, adriano.ferraresi, gabriele.carioli}@unibo.it

Abstract

This study investigates how Whisper handles interactional phenomena in spontaneous Italian conversation, focusing on backchannels, repairs, and filled pauses. We compare standard Word Error Rate (WER) optimization with a decoding strategy that explicitly rewards the preservation of interactional events. Results show that decoding choices have limited impact on overall accuracy, while recognition remains strongly phenomenon-dependent, suggesting structural limitations in the handling of interactional phenomena, with systematic linearization of repairs and frequent suppression of short conversational items.

Keywords: ASR, Interaction, Italian, Backchannels, Repairs, Filled Pauses

1. Introduction

End-to-end Automatic Speech Recognition (ASR) systems based on large neural models have reached high levels of performance. Models such as OpenAI's Whisper (Radford et al., 2023) are widely used thanks to their robustness across languages, recording conditions and speaker variability. However, most ASR systems are trained primarily on monologic, non-spontaneous speech and tend to normalize conversational input, treating interactional phenomena as noise (Lopez et al., 2022; Yamasaki et al., 2023). Previous studies report systematic limitations in the production of short conversational items, interjections, disfluencies, overlap and paralinguistic cues (Liesenfeld et al., 2023; Lopez et al., 2022; Umair et al., 2022; Zayats et al., 2019). All of these elements play a crucial role in the organization of spoken discourse and, from a linguistic perspective, constitute structurally relevant resources. This paper focuses on three specific phenomena: backchannels, conversational repairs and filled pauses. Backchannels signal attention or speaker alignment (Dideriksen et al., 2019; Mereu et al., 2024; Blomsma et al., 2024); repairs manage problems of speaking and understanding (Drew, 1997; Schegloff et al., 1977; Dingemanse and Enfield, 2015, 2024); and filled pauses reflect speech planning and turn management (Christenfeld et al., 1991; Spreafico, 2012; Cossavella and Cevasco, 2021). Because they are brief, prosodically subtle and often produced in overlap, they are especially vulnerable to omission or normalization in ASR output (Lopez et al., 2022). Studies show that limited representation in output transcripts does not necessarily imply their absence from the acoustic signal or from the model's internal representations. Interactional elements may be suppressed during decoding, when the system selects which hypotheses to render as text, implying that ASR output should therefore be treated as configuration-dependent (Vitale et al., 2024; Dinkar et al., 2023). This study adopts a linguistically oriented perspective to examine how different Whisper decoding configurations affect the representation of interactional phenomena in selected conversations from the KIParla corpus (Mauri et al., 2019)[1].

2. Backchannels, Repairs and Filled Pauses

Backchannels (Example (1)) are listener responses that display attention and support the speaker's ongoing turn (Blomsma et al., 2024; Dideriksen et al., 2019; Mereu et al., 2024). They may consist of short tokens or multi-unit sequences and contribute to maintaining interactional organization (Pernas and Borreguero Zuloaga, 2010). Following the definition proposed by Ward and Tsukahara (2000), a backchannel (i) responds directly to the other's utterance, (ii) is optional, and (iii) does not require acknowledgement.

(1) BOI115: ci sono esami sempre e
there are always exams and
si studia anche durante il
you have to study also during your
tirocinio anzi per forza
internship well actually you must
BOR031: madonna ok wow
oh my god ok wow

[1] Data and code are provided at `https://github.com/KIParla/say-again`.

Proceedings of SPEAKABLE @ LREC 2026, pages 16–30
11 May 2026.

BOI115: quindi non c'è un periodo di stop
so there isn't a break
che dici studio e basta
where you can just focus on studying
BOR031: **mhmhmh**
uh-huh, uh-huh
PBA030, ParlaBO (Mauri et al., 2024b).

Conversational Repair refers to practices through which speakers address problems of speaking, hearing or understanding (Schegloff et al., 1977; Fele, 2007; Dingemanse and Enfield, 2015; Clark, 2020). Repair preserves mutual understanding and conversational progressivity. It may be *self-initiated* (Example (3)), when speakers correct their own talk, or *other-initiated* (Example (2)), when recipients signal trouble and prompt clarification from the main speaker (Schegloff et al., 1977).

(2) PKP040: quando c'ha l'esame marco?
when does marco have his exam?
PKP041: il dieci
on the tenth
PKP126: quindi mo sta chiuso a studiare
so now he's locked up at home studying
[...]
PKP041: ma' tu te lo sei visto sherlock?
mum, did you watch sherlock?
PKP040: **prego?**
pardon?
PKP041: te lo sei visto sherlock? della bbc?
did you watch sherlock? the bbc one?
PKP040: no cioè se è quello lì che facevano
no I mean if that's
la serie sì
the series then yes
KPS008, KIPasti (Mauri et al., 2024a).

(3) PSB050: ti trovi meglio a bologna o pavia?
where do you feel most comfortable in Bologna or in Pavia?
PSB049: pavia
Pavia
eh no no scusa bologna scusa
no no sorry I mean Bologna
PSB050: ah era la risposta sbagliata
ah that was the wrong answer
PSB049: eh sì sì bologna
yeah yeah Bologna
SBIB006, StraParlaBO (Zucchini et al., 2026).

Filled Pauses (Example (4)) are hesitation markers that interrupt fluency without contributing to propositional content. They reflect speech planning and information retrieval processes, while often functioning as floor-holding devices (Christenfeld et al., 1991; Spreafico, 2012; Schettino and Cataldo, 2019; Cossavella and Cevasco, 2021).

(4) PST211: **ehm** ti vengono in mente dei casi
***um** can you think of any situations*
in cui si mischiano le due
where the two
lingue in casa
languages get mixed at home
cioè proprio stereotipico
like something typical
che succede?
that happens?
PST036: **ehm** quando mia mamma è
***um** when my mom gets*
arrabbiata
angry
STIR012, StraParlaTO (Bernasconi and Goria, 2026).

3. Whisper in Interactional Contexts

Whisper (Radford et al., 2023) represents a shift in ASR research. Trained on $680,000$ hours of audio, it shows strong zero-shot generalization across domains, speakers, and languages. Architecturally, it relies on a standard encoder–decoder Transformer architecture that formulates speech processing tasks as token prediction within a unified sequence-to-sequence framework. However, despite these recent advances, conversational speech remains challenging for ASR systems, including Whisper (Yamasaki et al., 2023). Elements central to interactional organization are underrepresented in ASR output, being disproportionately prone to be omitted or misrecognized, especially when utterances are very brief (Cumbal et al., 2021; Lopez et al., 2022). This issue is further compounded by the way transcription accuracy is typically evaluated, that is, through the Word Error Rate (WER, Klakow and Peters 2002). This metric has been criticized for conversational data because it oversimplifies performance and gives equal weight to all word-level errors, regardless of their interactional relevance (Liesenfeld et al., 2023; Gorisch and Schmidt, 2024). As a result, ASR systems may achieve acceptable global WER scores while still failing to capture interactional features that are crucial for Conversation Analysis. Therefore, in this work, we explicitly target two research questions, namely:

RQ1: How does decoding optimization affect the transcription accuracy of spontaneous conversational speech, as measured by global WER and event-specific metrics?

RQ2: Are certain interactional phenomena structurally more vulnerable to omission or normalization, regardless of optimization strategy?

4. Methodology

The experimental workflow can be summarized as follows: (i) the three interactional phenomena described in Section 2 were manually annotated on selected conversations from the KIParla corpus (see Section 4.1); (ii) audio data were processed through a speaker diarization and segmentation pipeline to obtain speaker-attributed segments (see Section 4.2); (iii) Whisper decoding was performed by optimizing inference-time parameters while keeping the ASR model fixed, using standard WER as a baseline objective function and an Interaction-aware objective function designed to promote the retention of interactional phenomena (see Section 4.3); (iv) the resulting transcriptions were normalized and evaluated both quantitatively, through (a) global WER and (b) Mean Match Ratio for annotated phenomena (see Section 4.4), and qualitatively, through detailed inspection of match and mismatch patterns and representative examples. Results were compared across configurations to assess the impact of the two optimization pipelines on the representation of interactional phenomena.

4.1. Dataset

We selected 26 2-speaker conversations drawn from the KIParla Corpus of Spoken Italian, covering three macro types of interaction: semi-structured interviews, student-professor meetings (i.e., oral exams and office hours) and free conversations. The sample is heterogeneous in terms of speakers' metadata and is in line with the general composition of the KIParla corpus. Linguistic variation is also present, including one L2 speaker of Italian and several conversations featuring dialectal traits from different regions. The dataset in its entirety amounts to 15h48m of audio.

The KIParla Corpus is a resource of spoken Italian and is entirely transcribed following a manual pipeline, in both ortographical and Conversation Analysis format, following Jefferson notation (Jefferson, 2004). The resource is available in vertical, pseudo-tokenized format where each token bears information about its type (e.g., linguistic, metalinguistic), ID, corresponding speaker code, Jefferson notation, etc.[2]

The first 20 minutes of each recording were manually annotated in INCEpTION (Klie et al., 2018) using a simple multi-layer scheme, aimed at indentifying backchannels (BC), self- and other-initated repairs (SR and OR) and filled pauses (FP), while also maintaining Conversation Analysis information such as Jefferson notation, overlaps and intonation patterns. A screenshot of the INCEpTION annotation interface is provided in Appendix B, Figure 7.

The annotated portion of the dataset resulted in 8h40m of conversation. Annotation was carried out by an expert linguist while listening to the corresponding audio in ELAN (Max Planck Institute for Psycholinguistics, 2025) to accurately capture interactional dynamics. The remaining 6h48m were split into two groups, and employed for Whisper's optimization (Subset A, 3h25m) and subsequent control analysis (Subset B, 3h22). Annotation criteria is described in Appendix A.

4.2. Diarization

Segmentation, diarization, transcription and optimization were performed using DIT.DaT, a modular pipeline developed for the Department of Translation and Interpreting of the University of Bologna[3], combining PyAnnote (Bredin et al., 2019) for speaker diarization and Whisper for transcription. The pipeline produces speaker-aligned outputs that can be manually reviewed step-by-step. Compared to alternative solutions such as WhisperX (Bain et al., 2023), this approach offers greater flexibility, a more controlled workflow and more reliable speaker segmentation for the interactional phenomena under investigation.

Four types of segmentation were explored: while sharing the same processing pipeline, they only differ with respect to a limited set of parameters that manage diarization and segmentation sensitivity. Table 1 shows the four segmentation configurations, denominated A, B, C and D. Configuration A is the most restrictive: it enables exclusive mode (which assigns overlapping speech to a single, dominant speaker) and uses conservative segmentation thresholds (2s minimum pause, 0.25s minimum duration), favoring cleaner but less overlap-sensitive turns. Configuration B is identical to A but disables exclusive mode, allowing overlapping speech to be attributed to multiple speakers. Configuration C further increases segmentation sensitivity by reducing the minimum pause threshold to 1 second, enabling finer-grained turn segmentation. Configuration D is the most permissive setup: it maintains the 1-second pause threshold and further lowers the minimum segment duration to 0.20 seconds, maximizing retention of very short turns

[2]These files are available on Github for each module. For more information: `https://github.com/KIParla/KIP?tab=readme-ov-file#verticalized-content`.

[3]`https://github.com/bilo1967/DIT.DaT`

(e.g., minimal responses and hesitation markers) at the cost of greater fragmentation. All conversations were finally segmented according to each configuration.

ID	exclusive mode	min. pause (s)	min. duration (s)	sensitivity
A	yes	2.0	0.25	low
B	no	2.0	0.25	medium
C	no	1.0	0.25	high
D	no	1.0	0.20	very high

Table 1: Overview of the four processing configurations used in the optimization process.

4.3. Optimization

Before optimization, a pre-processing pipeline was implemented to ensure comparability between the output of each Optuna trial with the gold standard transcription used for reference (i.e., the manually transcribed version available for consultation). As far as the optimization is concerned, Whisper decoding parameters were automatically optimized on Subset A with Optuna (Akiba et al., 2019). Two objective functions were explored. The first relied on standard WER, computed through word-level alignment between ASR output and the gold standard. WER is defined as:

$$\mathrm{WER} = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}$$

where S is the number of substitutions, D the number of deletions, I the number of insertions, C the number of correct tokens and N the total number tokens in the reference transcript. This standard WER-based optimization is used as a baseline condition against which the Interaction-aware objective described below is evaluated. Despite being widely criticized in the literature for disproportionately weighting the representation of interactional phenomena, WER was adopted as the primary optimization metric: at present, it remains the standard benchmark for evaluating ASR transcription accuracy and guiding parameter tuning. Also, the lack of widely established alternative metrics makes WER a suitable baseline objective function for this study.

A second, Interaction-aware objective function was introduced to better account for interactional items that are often suppressed or normalized in ASR output. Specifically, we introduce a new Loss function that weighs the WER according to the phenomena under investigation. Given the number of suppressed target items S_t and the number of correctly produced target items P_t, the objective function was defined as:

$$\mathcal{L} = \mathrm{WER} + \lambda \cdot S_t - \mu \cdot P_t$$

Suppressed target tokens are calculated as the proportion of backchannel, repair and filled pause tokens that are either deleted or substituted, and produced target tokens as the proportion of correctly transcribed tokens belonging to the same group. Target tokens were identified using a predefined form-based list of short conversational tokens provided in Appendix C. The coefficients were manually set to $\lambda = 0.3$ and $\mu = 0.1$: both weights were intentionally kept low to avoid distorting overall transcription quality or encouraging hallucinated output. This Interaction-aware optimization should be interpreted as an exploratory extension of the WER-based approach. It is important to underline that its aim is not maximizing the production of interactional events. Instead, it seeks to reduce their systematic suppression when they are present in the reference transcription, while preserving overall transcription accuracy. This choice reflects a conservative design aimed at testing whether small adjustments at decoding time are sufficient to influence the representation of interactional phenomena. For each segmentation configuration and for each objective function, the optimization process yielded the best trial, which represents the best-performing parameters. They are listed in Appendix D, Table 6.

4.4. Normalization and Error Analysis

Before evaluating the optimized outputs, a preliminary inspection of word-level alignments was conducted to identify systematic mismatches that were not related to genuine recognition errors. Raw substitution patterns and manual alignment checks revealed that several high-frequency errors were due to orthographic variation or tokenization inconsistencies. Normalization therefore addressed three main issues: (i) differences in token boundaries (e.g., apostrophe splitting: gold *all'* + *interno* vs. Whisper *all'interno*), (ii) truncated forms in repair sequences (e.g., *vennero anda-*), and (iii) deleted pause markers ([PAUSE]), which are not modeled by the ASR system. These adjustments aimed to prevent structural penalization of phenomena that Whisper cannot explicitly encode. Additionally, frequent orthographic variation in non-lexical vocalizations (e.g., nasal backchannels and filled pauses such as *m*, *mh*, *eh*) was observed. Tokens composed exclusively of *m* and *h* were treated as equivalent across gold and Whisper outputs. Similarly, *eh* $\rightarrow$ *e* ('and') substitutions were normalized only when functioning as filled pauses, based on their phonetic proximity and high occurrence (50 instances in the WER-based and 49 in the Interaction-aware output). In contrast, since no consistent pattern was found, substitutions such as *eh* $\rightarrow$ *è* ('is') were retained: in this case, normalizing them would have introduced interpretative assumptions rather than simply correcting orthographic variation. The top ten raw substitutions are listed in Appendix F. All these normalization choices reflect a methodological trade-off between evaluation fairness and linguistic fidelity. On the one hand, normalization reduces structural mismatches that would otherwise inflate error rates due to tokenization differences, Whisper encoding limitations, or orthographic variation, thereby improving comparability between ASR output and the reference. On the other hand, it may partially obscure the fine-grained form and variability of interactional phenomena, especially in cases where orthographic variation carries interactional or phonetic significance.

A quantitative error analysis was subsequently conducted on all configurations. Word-level alignments were examined to compute the distribution of insertions (INS), deletions (DEL), substitutions (SUB), and correct matches (OK) for interactional tokens. This analysis allowed error patterns to be evaluated independently from the overall WER, providing an empirical basis for comparing the two optimization strategies. To verify that the observed behavior was not specific to Subset A, a complementary analysis was conducted on an additional held-out subset of conversational data not used during parameter tuning, Subset B. This control analysis was performed only for best-performing configuration for each optimization strategy, therefore Configuration A.

Performance was evaluated at two complementary levels: global transcription accuracy and interactional event preservation. First, overall accuracy was assessed through the WER, computed for each conversation and each decoding strategy. Conversation-level WER values were compared in a paired design, and descriptive statistics (mean, variance, standard deviation) were used to summarize central tendency and dispersion. Visualizations at both conversation and aggregate level facilitated comparison between optimization strategies. Second, to specifically evaluate the interactional phenomena under investigation, a personalized metric denominated Mean Match Ratio was introduced. Each annotated sequence was treated as a single event, including multi-token units. For each event, a match ratio was computed as the proportion of correctly transcribed tokens over evaluable tokens. For each phenomenon type, the Mean Match Ratio was then calculated as the simple average across events, ensuring equal weight regardless of event length. Analyses were conducted at two levels. At the conversation level, paired differences between optimization strategies were computed for each phenomenon. Since Shapiro–Wilk tests indicated non-normal distributions of paired differences, Wilcoxon signed-rank tests were employed. At the interaction level, events were aggregated to provide a comparison across conversational settings. However, due to the limited number of conversations per interaction type, normality testing within each interactional setting was deemed unreliable. Consequently, only inferential and descriptive analyses were conducted at the overall type of interaction level.

5. Results

5.1. Analysis on Optimal Parameters

All best-performing trials converged on the `large-v3` model, therefore the findings should be interpreted as model-specific. Decoding parameters varied considerably across configurations (see Table 6 in Appendix D), with no consistent optimal strategy emerging.

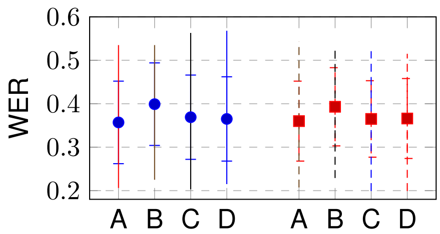


Figure 1: Mean WER with standard deviation (error bars) and min–max whiskers across configurations. Circles: WER-based optimization. Squares: Interaction-aware optimization.

Mean WER values remained comparable across objectives (Figure 1). In the WER-based optimization, mean values ranged from 0.357 (Configuration A) to 0.399 (Configuration B), with relatively similar variability across configurations. In the Interaction-aware optimization, Configurations A (0.360), C (0.365) and D (0.366) performed similarly, while B showed the highest WER mean (0.393). Variability was slightly reduced overall. Configuration A achieved both the lowest composite loss (0.563) and the lowest mean WER, and was therefore selected as the best Interaction-aware setup. Importantly, incorporating event-sensitive components did not degrade global WER.

Event-level error distributions (Table 2) show that deletions remain the dominant error under both objective functions (ranging from 50 to almost 55%), confirming the tendency of a structural suppression of interactional tokens. Introducing the Interaction-aware optimization does not substantially reduce deletion rates. However, minor shifts in distribution are observable: substitutions decrease slightly in some configurations (notably B), correct matches increase marginally, and insertions remain low. This means that the Interaction-aware objective function slightly redistributes errors without altering the overall omission tendency.

ID	type	DEL (%)	SUB (%)	OK (%)	INS (%)
A	WER	53.39	19.78	21.11	5.73
A	I-WER	54.67	20.44	21.39	3.50
B	WER	50.92	23.12	20.57	5.39
B	I-WER	53.04	20.45	20.91	5.61
C	WER	51.72	25.86	18.06	4.36
C	I-WER	50.89	25.13	19.67	4.32
D	WER	51.17	24.91	19.06	4.86
D	I-WER	51.18	24.83	20.12	3.87

Table 2: Distribution of alignment operations for event tokens under WER-based and Interaction-aware (I-WER) optimization.

The control analysis on Subset B (see Appendix E, Table 7) replicates this pattern. Interaction-aware optimization yields a slightly lower WER mean (33.24%) than the WER-based setup (34.23%), confirming that event-sensitive weighting does not harm global accuracy. Deletions, again, remain dominant (58%, approximately), with only minimal differences between strategies. Overall, Interaction-aware optimization introduces small, controlled shifts in error distribution while preserving transcription quality.

The contrast between the highest-WER configuration (Trial 0, value = 0.90) and the lowest-WER configuration (Trial 21, value = 0.35), which can be examined in Table 3, qualitatively illustrates the structural impact of decoding parameters. In the highest-WER output, lexical instability and hallucinations emerge: proper nouns are distorted (*San Luca* → *saluca/saluka*), non-existent forms appear (*l'equiline*, *videa*), and morphologically implausible variants are produced (*non cambiente*, *villana*). These errors may reflect decoding instability rather than simple substitution patterns. By contrast, the optimized configuration preserves referential consistency and syntactic coherence. *San Luca* remains stable, morphologically correct forms are produced (*non cambia niente*), and substitutions are semantically plausible (e.g., *nelle colline* instead of *l'equiline*). Minor infelicities persist (e.g., *sono tutte verde* instead of *sono tutte verdi*), but the transcript remains globally readable and interpretable. The difference is therefore not merely quantitative: suboptimal parameters produce cumulative lexical degradation and phonological drift, whereas optimized decoding preserves discourse continuity and referential stability.

5.2. Analysis on Annotated Data

Conversations from the annotated dataset were transcribed using the best parameters for each optimization strategy, that is, Configuration A for both. Across interaction types (Figure 2), relative rankings between the two types of optimizations remain stable, suggesting that conversation-specific factors (e.g., acoustic conditions, overlap density) may drive most variability rather than optimization strategy. Aggregated statistics confirm this pattern: the Interaction-aware configuration yields a slightly lower mean WER (0.361) compared to the WER-based setup (0.367), with marginally reduced variance (WER 0.0096 vs. IA 0.0071) and standard deviation (WER 0.0979 vs. IA 0.0846). Differences are small and distributional structures remain comparable. Overall, incorporating interaction-sensitive components does not substantially alter global transcription accuracy, but slightly stabilizes performance.

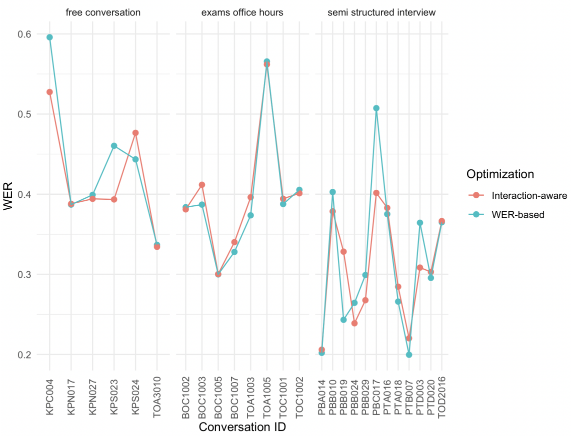


Figure 2: Comparison of conversation-level Word Error Rate (WER) between WER-based and Interaction-aware optimization strategies, grouped by interaction type. Each pair of points corresponds to the same conversation under the two decoding conditions.

Regarding the Mean Match Ratio, Wilcoxon signed-rank tests at the conversation-level revealed no statistically significant differences between the two configurations for any of the investigated phenomena, despite the variability observed in Figure 3. Indeed, all phenomena $p > 0.05$ (BC $= 0.207$, FP $= 0.780$, SR $= 0.197$, OR $= 0.371$). This indicates that the median difference in Mean Match Ratios between the two optimization strategies does not significantly deviate from zero: despite small conversation-level fluctuations, neither configuration demonstrates a systematic advantage over the other in the recognition of annotated interactional phenomena. This result suggests that decoding-time optimization alone may not be sufficient to substantially affect the recognition of interactional phenomena. Given the relatively low number of OR events per conversation, the Wilcoxon test for this phenomenon should be interpreted with caution, as limited sample size may reduce statistical power.

Clear differences emerge across phenomena when analyzed at the interaction-level (Table 4).

Speaker	High-WER output (Trial 0 – value 0.90)	Low-WER output (Trial 21 – value 0.35)
SPEAKER_B	San Luca, quando con una persona che mi ganna, magari sei innamorata, poi la scena quando vai lì, l'equiline, il San Luca è San Luca.	San Luca quando una persona che magari sei innamorata, poi la sera quando vai lì nelle colline, San Luca è San Luca.
SPEAKER_A	Il saluca è saluca. Cioè... Però avevi detto che hai fatto la camminata verso saluca che hai fatto. Sì, se mai non avevi avuto...	San Luca è San Luca. Però mi hai detto che hai fatto la camminata verso San Luca. Sì, sembra una mia vita.
SPEAKER_B	Saluka è sempre là, non cambiente perché tutti vanno alla fine, è sempre desiderata magari da tutti, vanno tutti là e quindi ci sono tutti tutti verdi, bellissimo, ci sono delle videa che rimani senza parole.	San Luca è sempre là, non cambia niente, perché tutti vanno alla fine, è sempre desiderata magari da tutti, vanno tutti là. E quindi sono tutte verde, bellissimo. Ci sono delle vite là che rimani senza parole.
SPEAKER_A	Chissà, se un giorno ormai...	Chissà se un giorno...
SPEAKER_B	per grattavinci compro una villana	Se non ci sono grattaventi comprerò una villa là.

Table 3: Comparison between the highest-WER configuration (Trial 0) and the lowest-WER configuration (Trial 21) of the WER-based optimization. Extracted from PBB010, ParlaBO (Mauri et al., 2024b).

Self-repairs show the highest Mean Match Ratios (≈ 45–56%), suggesting relative robustness. Backchannels display intermediate preservation (≈ 17–23%) and filled pauses remain extremely low across all contexts (≈ 1–4%), confirming their high susceptibility to omission. Other-initiated repairs show relatively high values, however they are based on small samples. That said, phenomenon type appears to be the primary determinant of recognition performance.

interaction type	phen.	n	WER (%)	Event (%)
free conversation	BC	524	16.92	16.89
	FP	174	3.45	4.02
	SR	48	51.80	52.49
	OR	13	46.15	53.85
exams/office hours	BC	636	22.10	21.53
	FP	761	1.38	1.45
	SR	133	46.64	45.12
	OR	18	55.56	58.15
semi-structured interview	BC	1354	21.71	23.01
	FP	776	4.23	4.17
	SR	62	56.11	53.33
	OR	0	–	–

Table 4: Mean match ratio (in percentage) by interaction type and optimization strategy.

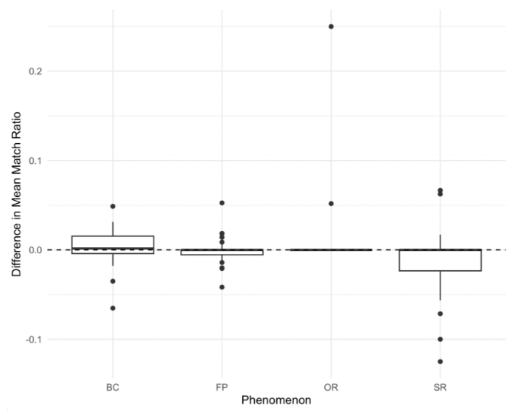


Figure 3: Distribution of differences in Mean Match Ratio, for each phenomenon.

5.3. Qualitative Analysis

Filled Pauses Across the few successful matches, filled pauses tend to be prosodically salient, segmentally isolated, and positioned at the onset of an intonational unit, preceding propositional content (e.g., PTD020, Figure 4). Nasal hesitations (e.g., *ehm*, *mh*) are almost systematically omitted, however *ehm* appears to be recognized more frequently than *mh*. The few matched instances of nasal forms, including the single normalization case (*mh* → *mmm* in KPS024), share similar properties: they are acoustically isolated, not produced in overlap, and occupy structurally well-defined turn positions. The near absence of preserved nasal forms overall suggests that short, low-intensity hesitation markers are particularly vulnerable to deletion, especially when produced in turn-final or non-prominent positions.

```
TOR001: | [cosa fai durante il tuo tempo libero?]
TOI008: |

TOR001: |
TOI008: |          [e:::h è una bella domanda]
event:  |          [FP      ]

TOR001: |
TOI008: | [diciamo che::: mh ]
event:  |                [FP ]
```

Figure 4: A recognized FP from PTD020, ParlaTO (Cerruti and Ballarè, 2020).

Backchannels Qualitative inspection reveals a clear positional pattern in the preservation of backchannels. Successfully recognized cases tend to occur outside overlap, either in the transition space between turns or at the onset of a new turn. In these contexts, backchannels such as *eh no ti capisco*, turn-initial *eh*, or nasal *mh* are sequentially and prosodically salient. Their placement at a recognizable boundary (i.e., immediately after a prior turn completion or as incipient speakership), makes them acoustically less masked and structurally easier to segment. Whisper therefore appears more likely to preserve backchannels when they are integrated into a clearly delimited turn rather than embedded within ongoing speech. Only a handful of preserved cases occur in overlap: however, these instances are relatively limited and typically involve multi-word backchannels (e.g., *eh immagino*, *eh vabbè*) that are longer and more acoustically robust than, for instance, an isolated interjection. The contrast becomes particularly evident in alignment sequences. In BOC1002 (Figure 5), the backchannel *bibliografia sì*, produced in overlap as an aligning response, is not preserved by Whisper. By contrast, the subsequent token *magistrale*, produced at the onset of a new turn and outside direct overlap, is correctly transcribed. This asymmetry suggests that overlap, rather than interactional function per se, constitutes the primary source of suppression.

```
BO087: |[[<bibliografia>],]     [magist..→
BO089: |[[bibliografia (.) sì].]
event: |[BC                ]

BO087: |rale.]
BO089: |      [magistrale (.) ho portato an..→
event: |      [BC        ]

BO087: |
BO089: |anche::: [il reperto..→
```

Figure 5: Example of both an unrecognized and a recognized BC alignment from BOC1002, KIP (Mauri et al., 2019).

Self-repairs Across both decoding strategies, self-repairs are predominantly partially preserved. Fully preserved cases are rare, and complete omissions remain limited but non-negligible. The dominant pattern is linearization, meaning that Whisper tends to retain the repair solution while deleting the repairable, truncations, and hesitation markers. In BOC1002, the sequence *delle opere pubbli- [PAUSE] delle traduzioni pubblicate* is reduced to *delle traduzioni pubblicate*: the initial formulation is removed, and only the corrected portion of the repair survives. Similarly, in TOC1002, *possono eh am- [PAUSE] agire* is simplified to *possono agire*, with the hesitation marker and truncated segment deleted. In TOC1001, in the sequence *che si è contratti [PAUSE] che si è contratto*, the incorrect plural agreement (*che si è contratti*) disappears, while the grammatically correct form in singular (*che si è contratto*) is directly preserved. In these cases, Whisper privileges grammatical coherence over the faithful reproduction of incremental speech production. Truncated forms are not preserved as such: in PTB007, the speaker first produces the truncated *dal millenovec-*, followed by the complete form *nel millenovecento*. In the ASR output, however, the truncated repairable is omitted, while the reformulated and phonologically complete segment (*nel millenovecento*) is retained. The repair trajectory therefore disappears, and only the resolved, well-formed formulation survives. More complex repairs are likewise compressed. In PBB029, the speaker first produces the incorrect clause *non ho mai pesato* and then reformulates it as *non mi è mai pesato*. In the ASR output, the entire first formulation is deleted, and only the second clause is preserved. Fully preserved cases are exceptional. In PBA024, both the initial formulation (*abito praticamente vicino al centro*) and the reformulation (*lavoro vicino al centro*), together with repair markers occurring between them (*cioè, scusa*), are retained. Unlike the previous examples, this sequence (i) contains no abrupt truncations (ii) does not contain a silenced pause and, finally, (iii) contains an explicit repair marker, which is the element that appears to have made the difference for the sequence to be fully preserved.

Other-initiated Repairs The qualitative analysis of other-initiated repairs (OR) reveals a recurrent pattern of structural simplification. Repair initiations are frequently partially preserved or entirely suppressed, especially when short, repeated, or produced in overlap. In BOC1005, the clarification request *che cosa vuol dire?* is produced twice in close succession, reinforcing the repair trajectory. However, Whisper only retains the first occurrence, while the second is omitted. From an interactional perspective, the repetition intensifies the request for clarification and signals persistent trouble. Its deletion attenuates this persistence, reducing the sequence to a single, non-reiterated question. This suggests that Whisper may suppress closely repeated repair initiators, particularly when they occur rapidly and in overlap. A more radical case of suppression is observed in TOA3010, shown in Figure 6. The repair initiation *ma chi?*, targeting referential ambiguity, and the subsequent specification (*Luca*) are both omitted in the ASR output. The entire repair trajectory disappears, leaving the referential problem unresolved. Here, both the initiator and the repair solution are removed, effectively erasing the interactional work performed to restore clarity. This pattern indicates that short, overlapping, and low-intensity turns are especially vulnerable to deletion. Minimal repair-relevant signals show similar behavior. In KPN027, the token *mh?*, functioning as a repair initiator or signal of trouble, is not transcribed. As already observed for filled pauses and backchannels, nasal tokens such as *mh* are systematically suppressed, likely due to their brevity and low acoustic salience.

```
TO086: |[perché lui in genere gli aneddoti non li
TO085: |

TO086: | butta a caso li dice:: per]  [[li le~]   ]
TO085: |                              [[ma chi]?  ]
event: |                              [OR         ]

TO086: | [luca ]
TO085: |
```

Figure 6: Example of an other-initiated repair mismatched in TOA3010, KIP (Mauri et al., 2019).

By contrast, matched cases tend to share clear structural and prosodic properties. In KPS023, the repair initiator *che COSA?* is preserved. Unlike the previous examples, it is prosodically salient (as highlighted by the capital letters signaling high volume in CA format) and occurs outside direct overlap, forming a clearly bounded turn. This suggests that acoustic prominence and sequential independence increase the likelihood of preservation.

6. Conclusion and Future Work

Decoding optimization exerts only a limited influence on overall transcription accuracy (RQ1). Across configurations and subsets, Word Error Rate remains broadly comparable between the WER-based and the Interaction-aware strategies. Incorporating interaction-sensitive components into the objective function does not degrade global accuracy, but it does not substantially improve event preservation either. Variation appears to be driven more by conversation-specific factors (e.g., acoustic conditions, overlap density) than by decoding objective. The absence of statistically significant differences between optimization strategies further supports this interpretation, suggesting that decoding-time adjustments alone may not be sufficient to overcome the structural limitations of ASR systems in representing interactional phenomena.

Recognition patterns are strongly phenomenon-dependent (RQ2). Self-repairs are often preserved in linearized form, with the repair solution retained and the disfluent material deleted. Backchannels show intermediate robustness, especially when produced outside overlap, whereas filled pauses are systematically suppressed. Other-initiated repairs remain particularly vulnerable when short or acoustically weak. Structural and acoustic properties appear to be playing a more decisive role than decoding strategy in determining event survival.

Despite limitations, the study produced a linguistically annotated dataset aligned with ASR output, which provides a valuable resource for further empirical investigation of conversational phenomena and a basis upon which improved evaluation and modeling approaches can be developed. Future work should extend the annotated dataset and explore inter-annotator agreement to strengthen the annotation scheme robustness. Also, a more structurally sensitive computational modeling of interactional events, beyond form-based token lists, would allow evaluation to better reflect sequential function rather than surface preservation. Further research should also explore training-level adaptations, including fine-tuning on interactionally annotated conversational data, to assess whether normalization tendencies can be mitigated beyond inference-time adjustments. Integrating acoustic analysis (e.g., duration, intensity, prosodic prominence) would help clarify whether suppression patterns reflect measurable phonetic vulnerability. Finally, comparative evaluation across different ASR architectures would determine whether the linearization effects documented here are Whisper-specific or characteristic of contemporary end-to-end systems more broadly.

7. Limitations

Despite the methodological care adopted in this study, several limitations must be acknowledged. First, the size of the dataset remains relatively modest. The gold-annotated portion amounts to 8 hours and 40 minutes of speech, while the optimization procedure was conducted on approximately 3 hours of conversational data. Although sufficient for exploratory analysis, this size may limit statistical power, particularly for low-frequency phenomena such as other-initiated repairs. In addition, the data subsets were balanced primarily in terms of duration rather than interactional composition, which may affect the generalizability of the results. Only the first 20 minutes of each conversation were annotated, potentially underrepresenting phenomena that emerge in later stages of interaction. Second, annotation was performed by a single annotator. While consistent criteria were applied, the absence of inter-annotator agreement measures represents a limitation, particularly for perceptually subtle phenomena such as filled pauses and short vocalic items: due to time constraints, multi-annotator validation was not feasible. Third, the weighting parameters of the Interaction-aware objective function were manually defined and not optimized through a systematic search or ablation study. Finally, the analysis is restricted to a single ASR system (Whisper large-v3). Although this model provides strong baseline performance, the findings should be interpreted as model-specific. It remains an open question whether similar biases occur in other Whisper models or ASR systems, or whether they are amplified by large-scale models trained on predominantly non-conversational data.

8. Acknowledgements

We would like to thank Jaka Čibej, who kindly introduced us to INCEpTION and provided essential advice on organizing the annotation scheme.

9. Bibliographical References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio.

Peter Blomsma, Julija Vaitonyte, Gabriel Skantze, and Marc Swerts. 2024. Backchannel behavior is idiosyncratic. *Language and Cognition*, 16:1–24.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2019. pyannote.audio: neural building blocks for speaker diarization.

N. Christenfeld, S. Schachter, and F. Bilous. 1991. Filled pauses and gestures: It's not coincidence. *Journal of Psycholinguistic Research*, 20:1–10.

Eve V. Clark. 2020. Conversational repair and the acquisition of language. *Discourse Processes*, 57(5-6):441–459.

Francisco Cossavella and Jazmín Cevasco. 2021. The importance of studying the role of filled pauses in the construction of a coherent representation of spontaneous spoken discourse. *Journal of Cognitive Psychology*, 33(2):172–186.

Ronald Cumbal, Birger Moell, José Lopes, and Olov Engwall. 2021. "you don't understand me!": Comparing asr results for l1 and l2 speakers of swedish. pages 4463–4467.

Christina Dideriksen, Riccardo Fusaroli, Kristian Tylén, Mark Dingemanse, and Morten H. Christiansen. 2019. Contextualizing conversational strategies: Backchannel, repair and linguistic alignment in spontaneous and task-oriented conversations. In *Annual Meeting of the Cognitive Science Society*.

Mark Dingemanse and N. J. Enfield. 2015. Other-initiated repair across languages: towards a typology of conversational structures. *Open Linguistics*, 1(1).

Mark Dingemanse and N.J. Enfield. 2024. Interactive repair and the foundations of language. *Trends in Cognitive Sciences*, 28(1):30–42.

Tanvi Dinkar, Chloé Clavel, and Ioana Vasilescu. 2023. Fillers in spoken language understanding: Computational and psycholinguistic perspectives.

Paul Drew. 1997. 'open' class repair initiators in response to sequential sources of troubles in conversation. *Journal of Pragmatics*, 28(1):69–101.

Giolo Fele. 2007. *L'analisi della conversazione*. Il Mulino, Bologna.

Jan Gorisch and Thomas Schmidt. 2024. Evaluating workflows for creating orthographic transcripts for oral corpora by transcribing from scratch or correcting ASR-output. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6564–6574, Torino, Italia. ELRA and ICCL.

Gail Jefferson. 2004. Glossary of transcript symbols with an introduction. In Gene H. Lerner, editor, *Conversation Analysis: Studies from the First Generation*, chapter 2, page 13–31. John Benjamins, Amsterdam / Philadelphia.

Dietrich Klakow and Jochen Peters. 2002. Testing the correlation of word error rate and perplexity. *Speech Communication*, 38(1):19–28.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of System Demonstrations of the 27th International Conference on Computational Linguistics*, pages 5–9, Santa Fe, NM.

Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse. 2023. The timing bottleneck: Why timing and overlap are mission-critical for conversational user interfaces, speech recognition and dialogue systems.

Alianda Lopez, Andreas Liesenfeld, and Mark Dingemanse. 2022. 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)*, pages 135–143, Potsdam, Germany. KONVENS 2022 Organizers.

Caterina Mauri, Silvia Ballarè, Eugenio Goria, Massimo Cerruti, and Francesco Suriano. 2019.

KIParla corpus: A new resource for spoken Italian. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, pages 243–249, Bari, Italy. CEUR Workshop Proceedings.

Max Planck Institute for Psycholinguistics. 2025. *ELAN (Version 7.0)*. The Language Archive, Nijmegen. Computer software.

Daniela Mereu, Francesco Cangemi, and Martine Grice. 2024. Backchannels are not always very short utterances. the case of italian multi-unit backchannels. *Journal of Pragmatics*, 228:1–16.

Pilar Pernas and Margarita Borreguero Zuloaga. 2010. Cortesia e scortesia in un contesto di apprendimento linguistico: la gestione dei turni. In Marcello Pettorino, Antonietta Giannini, and Francescamaria Dovetto, editors, *La Comunicazione Parlata 3. Atti Del Congresso Internazionale (Napoli, 23-25 Febbraio 2009)*, volume I, pages 227–247. Università Degli Studi Napoli L'Orientale, Napoli.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Emanuel Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language*, 53:361–382.

Loredana Schettino and Violetta Cataldo. 2019. Lexicalized pauses in italian. pages 189–192.

Lorenzo Spreafico. 2012. Le pause piene nel parlato plurilingue. In *Lessico e lessicologia: atti del XLIV Congresso internazionale di studi della Società di linguistica italiana (SLI)*, number 56 in Pubblicazioni della Società di linguistica italiana, pages 355–368, Roma. Bulzoni.

Muhammad Umair, Julia Mertens, Saul Albert, and J. Ruiter. 2022. Gailbot: An automatic transcription system for conversation analysis. *Dialogue & Discourse*, 13:63–95.

V.N. Vitale, L. Schettino, and F. Cutugno. 2024. Rich speech signal: exploring and exploiting end-to-end automatic speech recognizers' ability to model hesitation phenomena. In *Proc. Interspeech 2024*, pages 222–226.

Nigel Ward and Wataru Tsukahara. 2000. Tsukahara, w.: Prosodic features which cue backchannel responses in english and japanese. journal of pragmatics 23, 1177-1207. *Journal of Pragmatics*, 32:1177–1207.

Hiroyoshi Yamasaki, Jérôme Louradour, Julie Hunter, and Laurent Prévot. 2023. Transcribing and aligning conversational speech: A hybrid pipeline applied to french conversations.

Vicky Zayats, Trang Tran, Richard Wright, Courtney Mansfield, and Mari Ostendorf. 2019. Disfluencies and human speech transcription errors. In *Interspeech 2019*, pages 3088–3092.

10. Language Resource References

Bernasconi, Beatrice and Goria, Eugenio. 2026. *StraParlaTO*. Università degli Studi di Torino, 0.1.0.

Cerruti, Massimo and Ballarè, Silvia. 2020. *Modulo ParlaTO*.

Mauri, Caterina and Ballarè, Silvia and Zucchini, Eleonora. 2024a. *KIPasti*. distributed via ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", 1.1.0. PID http://hdl.handle.net/20.500.11752/OPEN-2124.

Mauri, Caterina and Ballarè, Silvia and Zucchini, Eleonora. 2024b. *ParlaBO*. distributed via ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", 1.1.0. PID http://hdl.handle.net/20.500.11752/OPEN-2126.

Mauri, Caterina and Goria, Eugenio and Ballarè, Silvia. 2019. *Modulo KIP*.

Zucchini, Eleonora and Ballarè, Silvia and Mauri, Caterina. 2026. *StraParlaBO*. Alma Mater Studiorum - Università di Bologna, 0.1.0.

A. Appendix: Annotation Criteria

Table 5 summarizes the annotated phenomena, tags and examples.

Phenomenon	TAG	Function	Example
Backchannels	BC	Continuers, assessments, agreement, incipient speakership	*mhmh*, *va bene*, *ho capito*
Filled pauses	FP	Hesitation markers	*eh*, *ehm*, *mh*
Self-repair	SR	Same turn, transition space, third position	*vennero anda- ... andarono*
Other-initiated repair	OR	Open and restricted initiators	*eh?*, *che cosa?*

Table 5: Overview of annotated phenomena.

Backchannels (BC) Both vocalic and lexical/multi-unit backchannels were annotated, including continuers, assessments, agreement tokens and incipient speakership. Particular attention was given to short and overlapping items, which are known to be vulnerable to omission in ASR output.

Other-initiated repair (OR) Annotation includes open-class initiators (e.g., *eh?*, *prego?*) and restricted requests (e.g., wh-questions). More elaborate repair sequences were excluded, as they are less prone to omission.

Self-repair (SR) Self-repair was annotated when speakers explicitly reformulated their own talk. This includes truncations followed by correction and repetitions reflecting reconstruction of prior material. Interruptions due to hesitation or overlap were excluded unless they resulted in clear reformulation.

Filled pauses (FP) Filled pauses include vocalic hesitation markers (e.g., *eh*, *ehm*, *mh*), identified through auditory inspection. Ambiguous cases were resolved based on prosodic and interactional cues. When occurring within self-repair, they were annotated both as FP and as part of SR.

General considerations It must be taken into consideration the fact that the KIParla corpus is a modular and incremental resource, and, as such, it was not always possible to control and uniform spelling variation with regard to these phenomena. Also, the project has grown over the years and not all transcription conventions were solidly defined from the start.

B. Appendix: INCEpTION Annotation Interface

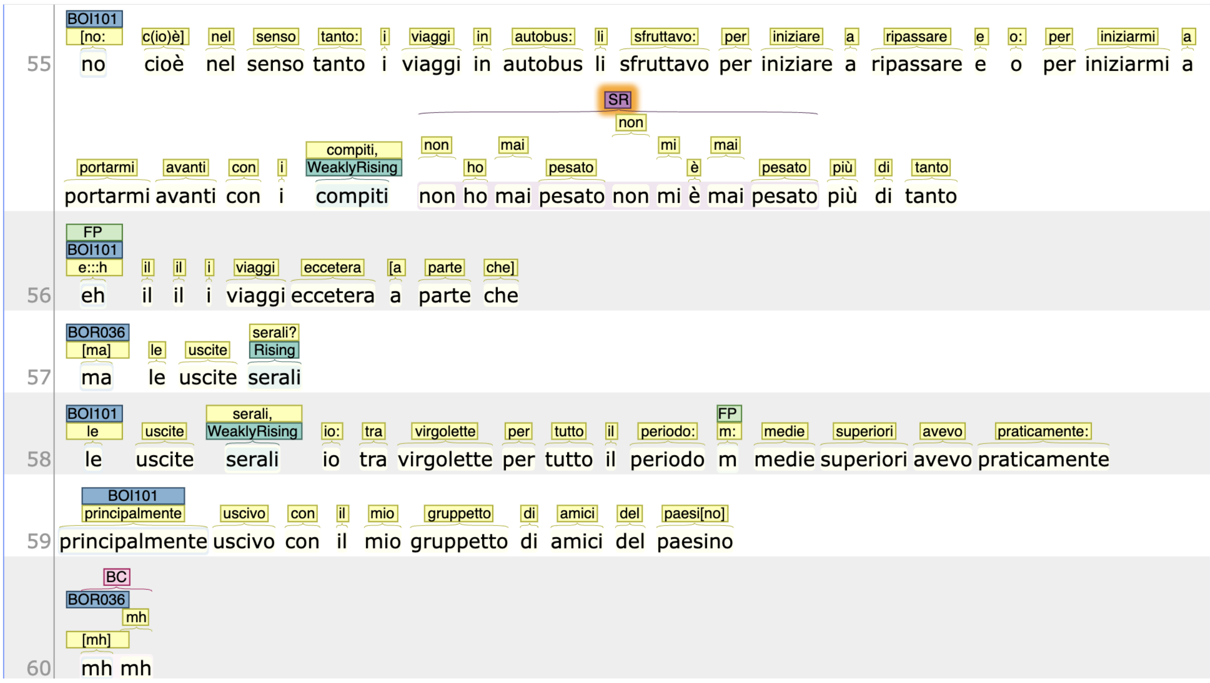


Figure 7: Interface of the INCEpTION annotation environment with annotated phenomena.

C. Appendix: Form-based Event Token List for Optimization

- *eh*
- *ehm*
- *ah*
- *oh*
- *mh*
- *mhmh*
- *mm*
- *okay*
- *sì*
- *bene*
- *esatto*
- *cioè*
- *diciamo*
- *insomma*

D. Appendix: Optimal Parameters

	WER				Interaction-aware WER			
Parameter	**A**	**B**	**C**	**D**	**A**	**B**	**C**	**D**
temperature	0.143	0.636	0.555	0.375	0.053	0.266	0.240	0.158
beam size	3	10	4	4	7	8	8	1
best-of	6	2	5	9	2	1	2	3
no speech threshold	0.203	0.415	0.600	0.249	0.539	0.354	0.617	0.465
compression ratio	2.121	2.501	2.240	2.61	1.734	1.630	2.227	2.492
model	large-v3	large-v3	large-v3	large-v3	large-v3	large-v3	large-v3	large-v3
condition on previous text	true	true	true	false	false	true	false	true
patience	0.592	0.626	0.344	0.217	0.506	0.308	0.705	0.316
length penalty	0.479	0.616	0.700	0.241	0.241	0.295	0.614	0.397

Table 6: Best-performing trials selected through Optuna optimization for each configuration under the WER-based and Interaction-aware objective functions.

E. Control Analysis on Subset B

type	WER (%)	DEL (%)	SUB (%)	OK (%)	INS (%)
WER	34.23	57.92	18.75	20.41	2.90
Interaction-aware	33.24	57.73	19.76	19.75	2.75

Table 7: Comparison summary of WER, deletions, substitutions, matches and insertions for the two decoding strategies on Subset B.

F. Appendix: Top 10 Raw Substitutions

WER-based (raw)			Interaction-aware (raw)		
Gold	**Whisper**	**Freq.**	**Gold**	**Whisper**	**Freq.**
eh	e	50	eh	e	49
mh	grazie	26	mh	grazie	28
eh	è	17	eh	è	17
ehm	e	16	mh	e	13
mh	e	15	ehm	e	12
mh	non	13	mh	non	11
mh	è	12	mh	che	10
sì	grazie	11	sì	grazie	10
mh	che	10	mh	è	9
eh	grazie	9	okay	e	9

Table 8: Top 10 raw substitution patterns across optimization strategies, before optimization.

Not all polar questions are the same: ASR, Humans, and Russian

Maria Onoeva
Charles University and Humboldt-Universität zu Berlin
onoevam@ff.cuni.cz

Abstract

Word Error Rate (WER) remains the standard metric in automatic speech recognition (ASR) evaluation, yet it does not capture higher-level linguistic distinctions such as prosody. This article examines how three state-of-the-art open-source ASR models (Whisper, Meta's MMS, and GigaAM) handle the distinction between Russian polar questions and assertions. Russian is particularly suitable for this investigation because polar questions can be marked either morphologically (*li, razve*) or purely intonationally, without changes in word order. Using audio stimuli from a controlled psycholinguistic experiment, I compare human classification performance in two experimental studies with ASR transcriptions, taking sentence-final punctuation as a proxy for prosodic interpretation. While human participants show near-ceiling accuracy, the ASR models perform inconsistently, especially on intonationally marked questions. Additional contextual cues improve performance in some cases but also reveal instability across conditions. The results demonstrate that evaluating punctuation provides insights beyond WER and allows a more fine-grained view of how current ASR systems encode prosodic and grammatical information.

Keywords: ASR, Russian, polar questions, intonation

1. Introduction

Moving beyond Word Error Rate (WER) as the sole evaluation metric for automatic speech recognition (ASR) systems is essential for a more comprehensive assessment of their performance (Aksënova et al., 2021; Gandhi et al., 2022). While WER captures lexical accuracy, it does not reflect higher-level linguistic properties. In many languages, crucial grammatical distinctions are expressed through intonation rather than through segmental morphology or word order (e.g., focus or question marking; Ladd 2008, p. 5). However, despite operating on acoustic input, most ASR systems are optimized for lexical transcription rather than for the explicit modeling of suprasegmental structure (Renals and King, 2010, p. 814, 829).

In this paper, I investigate the extent to which three off-the-shelf ASR models encode prosodic distinctions in a low-resource experimental setting. Following earlier work that considers punctuation as an additional evaluation dimension (Meister et al., 2023; Gris et al., 2023), I use sentence-final punctuation as a proxy for prosodic interpretation. While punctuation does not map to speech in a one-to-one manner, it provides a measurable way to probe whether ASR systems differentiate between sentence types that are structurally identical but prosodically distinct.

Using audio stimuli from a psycholinguistic eye-tracking study designed to examine the processing of Russian polar (yes/no) questions and assertions (Razguliaeva et al., to appear), I compare human classification performance with the transcriptions produced by three state-of-the-art open-source models —OpenAI's Whisper (Radford et al., 2022), Meta's Massively Multilingual Speech (MMS) (Pratap et al., 2023), and GigaAM (Kutsakov et al., 2025). Russian offers a particularly suitable testing ground, as polar questions may be marked either morphologically or purely intonationally, without changes in word order. This allows us to directly assess whether current ASR systems capture distinctions that rely exclusively on prosody.

This paper is organized as follows. Section 2 introduces the relevant properties of Russian polar questions and assertions. Section 3 presents the ASR models and experimental results. Section 4 discusses the findings, and Section 5 concludes.

2. Russian PQs and assertions

Classified as a "Question Particle" language in Dryer (2013), Russian, in fact, exhibits two polar question (henceforth, PQ) strategies, both of which allow negation (Restan 1972; Zanon 2024; Korotkova 2023; Šimík to appear, a.m.o.): (i) overt particle marking as in (1) or (2), and (ii) intonational marking as in (3). The particle *li* attaches to a fronted verb in canonical matrix PQs as in (1) (King, 1994). PQs with another question particle *razve*, as in (2), are dubbed biased and approximately translated to English PQs with 'really' (Geist and Repp 2023, cf. Korotkova submitted).

(1) (Ne) Zažëg li Miša večerom svečku?
NEG lit LI Miša evening candle
'Did Miša (not) light a candle in the evening?'

(2) Razve Kira (ne) xodila segonja v školu?
RAZVE Kira NEG went today in school
'Did Kira really (not) go to school today?'

Proceedings of SPEAKABLE @ LREC 2026, pages 31–38
11 May 2026.

The latter intonation strategy in (3) preserves declarative SVO order with questionhood indicated exclusively by prosody, placing a special nuclear pitch accent L+H* on the inflected verb (Meyer 2004; Esipova 2025).[1] Compare an intonPQ in (3) that carries the prominence on the verb *zažëg* 'he lit' (visual representation in Figure 1) with an assertion in (4) (also in Figure 2).[2] Meyer and Mleinek (2006, p. 1616) point out that this type of PQs might sound impolite or rough to English or German ears; however, on par with liPQs, intonPQs are attested in out-of-the-blue contexts, i.e., they are deemed canonical in Russian. But contrary to liPQs, intonPQs are more frequent in spoken speech and corpora (Restan 1972; Bryzgunova 1975; Esipova 2025; King 1994; Onoeva and Staňková 2025).

(3) Miša (ne) zažë$_{L+H*}$g večerom svečku?
Miša NEG lit evening candle
'Did Miša (not) light a candle in the evening?'

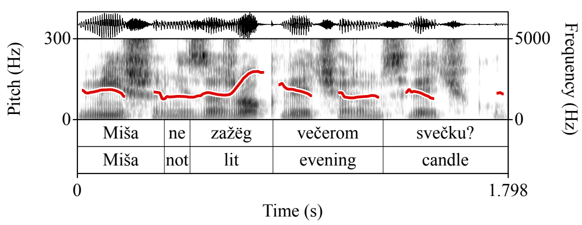


Figure 1: intonPQ in (3)

(4) Miša (ne) zažëg večerom lampočku.
Miša NEG lit evening lightbulb
'Miša did (not) light a lightbulb in the evening.'

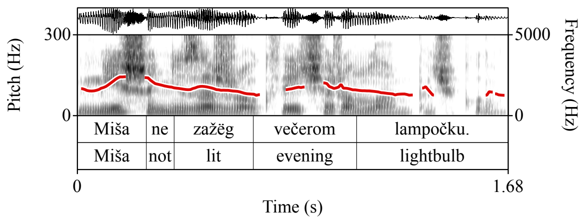


Figure 2: ASSERTION in (4)

As expected, native Russian speakers presented with just auditory cues (≈ 1 to 2 seconds recorded by a male speaker) detect the differences between the two sentence types at a very high level. During the eye-tracking visual world paradigm experiment reported in Razguliaeva et al. (to appear), we were able to observe the differences in processing between assertions like (4) (16 positive and 16 negative sentences) and PQs (32 positive and 32 negative for each liPQs in (1) and intonPQs in (3)): while in assertions, participants (N = 52) focused their attention on the picture corresponding to the expressed polarity, i.e., to a picture with the unlit lightbulb for (4) with negation and to the lit one with a positive assertion. Fixations in questions became concentrated on the positive picture (i.e., the lit candle for (1) and (3)), largely independently of the polarity expressed in the stimulus. The results were further replicated in a forced-choice task conducted with the same audio stimuli. The task was to listen to an audio recording and select one of the two presented options: for the sentences as in (1), (2), and (3), participants (N = 54) were expected to press "question," while for (4) they were to press "assertion". Participants showed near-ceiling performance in utterance classification; the descriptive results are in Figure 3 (means $= 0.97$-1, SD $=$ 0.00-0.16, with little variance for statistical models to detect group differences).[3]

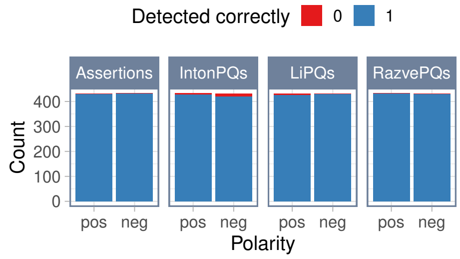


Figure 3: Plain audios results: Human accuracy.

The same task, when performed on the identical set of audio files by the three ASR models, proved to be considerably more problematic, as shown below.

3. ASR and Russian PQs

The evaluated ASR systems differ in architecture and decoding strategy. Whisper (Radford et al., 2022) is an autoregressive encoder–decoder model, `large_v3` was used for the experiments; transcriptions were generated using greedy decoding (temperature $= 0.0$). MMS by Meta (Pratap et al., 2023) is a CTC-based end-to-end model decoded via standard greedy CTC decoding. Since MMS does not generate punctuation, I restored punctuation in a post-processing step using the

[1] Alternatively, this nuclear pitch accent can be placed on the linearly last stressed syllable, resulting in an explanation-seeking question with a higher Question Under Discussion (Esipova and Romero, 2023; Esipova, 2025). It is not examined here and is left for future research.

[2] The graphics are compiled in Praat (Boersma and Weenink, 2009).

[3] All statistical analyzes and visualizations were performed in R (R Core Team, 2021).

Table 1: Plain audios: Humans and ASR models accuracy (%)

	ASSERTIONS		INTONPQs		LIPQs		RAZVEPQs	
	neg	pos	neg	pos	neg	pos	neg	pos
Humans (N=54)	100	99.8	97.2	99.1	99.8	98.6	99.8	100
Meta's MMS	100	100	0	0	18.8	18.8	68.8	68.8
GigaAM (CTC)	100	100	96.9	56.2	100	93.8	100	100
GigaAM (RNNT)	100	100	100	34.4	100	90.6	100	100
Whisper	100	100	62.5	6.2	96.8	84.4	100	100

external Silero Text Enhancement model (Silero Team, 2026). GigaAM (Kutsakov et al., 2025), designed specifically for Russian, was evaluated in two variants: (i) a CTC model decoded with greedy CTC decoding and (ii) an RNNT model using its standard streaming decoding. Except for Silero, no external language models were applied, ensuring comparability under low-resource conditions.

For Russian speech, the WER reported for the base Whisper `large_v3` model is $5.8\,\%$ on Common Voice and $5\,\%$ on FLEURS (OpenAI, 2022). The end-to-end variant of GigaAM-v3, which was used for both CTC and RNNT, won $70\,\%$ of pairwise side-by-side comparisons against Whisper (Salute Developers, 2023). MMS 1B achieves approximately $15\,\%$ WER on Russian in the multilingual FLEURS evaluation (Pratap et al., 2023). However, for the present study, punctuation, namely, a period for assertions and a question mark for questions, was used as a dependent variable in the ASR experiments.

3.1. Plain audios – results

Accuracy for the ASR models was defined as the proportion of utterances whose sentence-final punctuation matched the intended sentence type (question mark for questions, period for assertions). All other outputs were counted as incorrect. Similar to humans, detecting plain assertions and RAZVEPQs posed no difficulty for the ASR models, with near-ceiling accuracy across polarity conditions, see Table 1 and Figure 4 for the results. The only exception was MMS, which reached $68.8\,\%$ for RAZVEPQ. A larger variation emerged for LIPQs, while MMS showed low accuracy, GigaAM (CTC and RNNT) performed strongly on these questions, and Whisper remained slightly below ceiling.

The most striking differences appeared for INTONPQs: unlike top-performing humans, the models often fail to differentiate between the sentence types. MMS did so completely with $0\,\%$ accuracy, meaning it placed the period instead of the question in all INTONPQs. GigaAM models showed different results on the same set of audios: while RNNT reached $100\,\%$ accuracy in the negative condition, it dropped sharply in the positive one ($34.4\,\%$), resulting in a stronger polarity asymmetry compared to

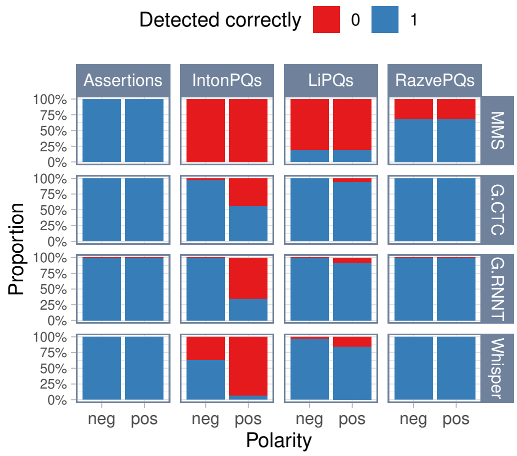


Figure 4: Plain audios results: ASR accuracy.

CTC ($96.9\,\%$ vs. $56.2\,\%$). Whisper also performed inconsistently, detecting only $6.2\,\%$ of positive INTONPQs and $62.5\,\%$ of negative. Overall, models handled morphologically marked questions better than purely intonational ones. To further investigate whether transcription accuracy can be improved, I conducted additional tests with the same audio stimuli enriched by some context.

3.2. Enriched audios – results

The original audio files recorded by a native male Russian speaker were enriched with nine additional contexts spoken by a female native speaker. The contexts were either related to questioning or unrelated, and were added before or after the plain utterances. Question-related contexts included polar replies such as *Ja dumaju, čto da/net* 'I think that yes/no' appended after a sentence, and preposed cues such as *On sprosil* 'He asked' and *Sledujuščee predloženie — èto vopros* 'The next sentence is a question'. Unrelated contexts included the prepended *On skazal* 'He said', the word *čerepaxa* 'turtle', and excerpts from Pushkin's *Ruslan and Ljudmila* (R&L), added either before or after the original audio. Figure 5 summarizes the results of all ASR experiments across four models.

Each heatmap shows accuracy (%), with darker shades indicating higher performance. Rows rep-

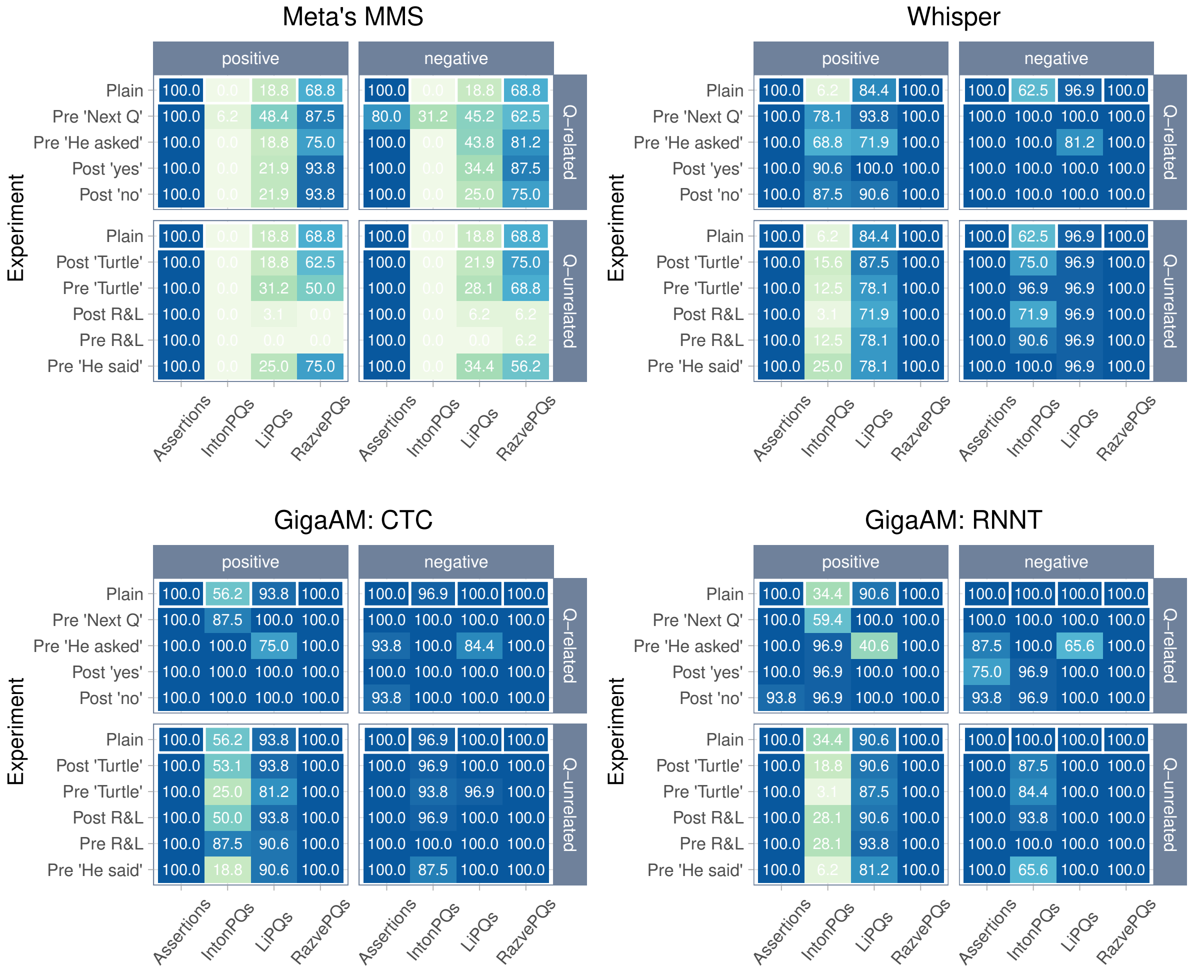


Figure 5: All ASR experiments: darker color = higher accuracy (%). The first row condition PLAIN had no context and added to each facet as a baseline comparison.

resent context manipulations, while columns distinguish between sentence types with and without negation. The first row in each facet corresponds to plain audio (i.e., from Table 1) and serves as a no-context baseline. Compared to the plain audios, the additional utterances had virtually no effect on assertions and RAZVEPQs (the first and last columns, respectively), which remained at or near ceiling across models and contextual manipulations. Once again, except for Meta's MMS in the preceding and added after lines from R&L, which, however, already had low accuracy in the baseline.

The results for positive INTONPQs (the left top and bottom facets) show a similar pattern in GigaAM CTC, GigaAM RNNT, and Whisper. Performance for the enriched audio stimuli with contexts related to question asking (the top left facet) is generally stable and, in some cases, slightly improves over the plain condition, although the gains are rather limited (only GigaAM RNNT does not exceed $60\,\%$ in Pre 'Next Q'). In contrast, the bottom left part of the plots, corresponding to conditions unrelated to questioning, shows a consistent decline, often reaching or falling below the plain baseline. This decrease is particularly visible for Whisper and GigaAM RNNT, where accuracy drops substantially compared to their baseline performance, while GigaAM CTC remains comparatively more robust, with a notable exception in the Pre R&L condition ($87.5\,\%$). The bootstrap results (provided in Table 2 in the Appendix) support the patterns observed in Figure 5, confirming that improvements for positive PQs are primarily associated with question-related contexts, while unrelated contexts tend to yield no effects or worsen detection. For negated PQs, Whisper shows better performance as well, but for the GigaAM models, the effect of context is minimal, with performance remaining largely stable across conditions.

Notably, there is a decrease in performance for Pre 'He asked' in LIPQs. In addition to matrix polar questions, the particle *li* also occurs in embedded contexts such as (5), resembling English *whether* or German *ob* (Restan, 1972; King, 1994).

(5) Mama sprosila, kupil li Maks xleb.
Mom asked bought LI Maks bread
'Mom asked whether Max bought bread.'

The drop is particularly visible for GigaAM RNNT: when the prepended context 'He asked' is combined with the presence of the *li* particle, the model tends to interpret the utterance as an embedded clause, whereas previously it placed a question mark. Although this is a possible interpretation, in our set it was still counted as incorrect, as two sentences were produced by a female and a male, meaning the models failed to recognize two different speakers.

4. Discussion

As rightly noted by the two anonymous reviewers, the experiment with Meta's MMS has a methodological limitation: the model itself does not predict punctuation but relies on an additional Silero component, which has no access to prosody. However, I decided to retain these results for three reasons. First, they point to a clear direction for improvement, namely integrating prosody and punctuation prediction. Second, MMS performs consistently poorly across conditions; i.e., it is stable in its errors, which makes its behavior interpretable rather than noisy. Third, this case also highlights a limitation of WER: while MMS may achieve a relatively good WER score, this metric does not capture its failure to detect sentence type.

When it comes to GigaAM and Whisper, the improved transcriptions observed when question-related utterances were added can plausibly be attributed to the architectures of the models. Whisper employs a transformer encoder–decoder model, while GigaAM is based on a Conformer architecture with self-attention. In both cases, attention mechanisms allow the models to make use of broader contextual information, which appears to improve question detection.

While this is encouraging, it also highlights a limitation of current ASR systems for Russian PQs. As is clear from above, prosody alone often signals sentence type and is reliably detected by native speakers without relying on additional contextual cues. At the same time, as rightly pointed out by an anonymous reviewer, human performance may also depend on factors such as language proficiency or sensitivity to contextual cues. However, all participants in the present study were native speakers of Russian, which minimizes the variability of the former issue. It is possible that non-native speakers would show lower accuracy or rely more on contextual information. For the latter issue, it is also conceivable that enriched audio could mislead human participants; for instance, by biasing them to interpret assertions as questions when preceded by cues such as *He asked* or *The next sentence is a question*. However, even in such cases, the effect would arise at the level of pragmatic expectations, rather than from an inability to process prosodic structure. By contrast, the ASR systems appear to rely on such external pragmatic cues in order to approximate distinctions that human listeners can derive directly from the acoustic signal.

Moreover, the models behave inconsistently across different contextual manipulations, which points to instability in their performance. This variability is not captured by standard WER, highlighting the need for more fine-grained evaluation measures sensitive to sentence type interpretation. While future work could extend the comparison to proprietary ASR systems or explore parameter settings such as temperature and beam size, it should no longer be treated as low-resource. Taken together, these results suggest that, for Russian, current ASR systems still fall short of human listeners in their ability to interpret sentence type from prosody alone.

Another crucial result from the experiment is that negation noticeably improves INTONPQs detection for multilingual Whisper and Russian-only GigaAM (accuracy is never below $60\,\%$), which, in turn, points to the fact that the models do rely on prosody somehow. Negative PQs might occur relatively often in the training data for Russian, so the models must have picked the combination of the nuclear pitch on the verb and negation as a question marker. I suggest that it is the combination of the two because (i) the models struggle with positive INTONPQs, i.e., with just verbal prominence, but (ii) for negative assertions, i.e., with no prominence on the verb, they perform with $100\,\%$ accuracy. It is unexpected because, cross-linguistically, negative PQs are considered to be marked (or biased; see, e.g., Gärtner and Gyuris 2017; Goodhue 2022) and used in specific contexts; thus, they should occur much less frequently than positive ones (for English, see Keisanen 2006). This also contradicts the results for Russian from the spoken corpus (Onoeva and Staňková, 2025): out of 500 randomly collected PQs, only 79 were negative ($15.8\,\%$). On the other hand, the so-called expletive negation is widely attested in Russian PQs (Brown and Franks, 1995; Abels, 2005; Zanon, 2024). Semantically, it is interpreted as having no negative force, but it might contribute a different meaning flavor in PQs, e.g., the inquirer's attitude towards a possible answer (see a similar idea for Czech in Šimík to appear). Mills (1992) brings further support for that claim, as she links negation in Russian PQs to politeness.

From an applied perspective, these findings raise questions about how current ASR systems handle prosodic information in downstream applications, especially in tasks where sentence type is relevant, such as dialog systems or speech interfaces. At the same time, the present experiments show how controlled manipulations of input can be used to examine model behavior more systematically, in particular, their sensitivity to contextual cues.

5. Conclusion

In this article, I pursued the goal of going beyond WER as the primary metric in ASR evaluation and gaining insight into the "black box" behavior of state-of-the-art models. Punctuation proved to be a useful proxy in low-resource settings, allowing assessment of sentence type distinctions. The results suggest that standard WER-based evaluation would be insufficient here, as it does not capture systematic errors in prosodic interpretation. The experiments further indicate that current ASR systems rely on contextual cues and struggle to generalize from prosody alone, as reflected in their variability across conditions and the asymmetry between positive and negative INTONPQs. At the same time, their relatively strong performance on negative INTONPQs suggests that they do have access to prosodic information. However, reducing speech to text still overlooks other prosodic features, such as focus marking. Overall, this approach makes it possible to probe model behavior in a more fine-grained way.

6. Acknowledgments

I wish to express my gratitude to my co-authors of Razguliaeva et al. (to appear): Mariia Razguliaeva, Radek Šimík, Roland Meyer, and Kateřina Hrdinková. I also thank the organizers and committee of the SPEAKABLE workshop and the anonymous reviewers for their comments that improved the article. The study was funded by a one year grant from DAAD and CRC 1412 *Register* fellowship awarded to me in 2025–2026.

7. Data and code availability

Data and code are available on OSF: `https://osf.io/4vcya/overview?view_only=d5bdad3e77d04e89981300236f634fbf`

8. Ethics

Ethical approval was received for the QueSlav project (funded jointly by the Czech Science Foundation and the German Research Foundation). Human participants were reimbursed for the forced-choice task with €3.20. One participant was removed because they did not pass a reliability test (detect RAZVEPQs as PQs in $95\,\%$).

9. Bibliographical References

Klaus Abels. 2005. "Expletive Negation" in Russian: A Conspiracy Theory. *Journal of Slavic Linguistics*, 13(1):5–74.

Alëna Aksënova, Daan van Esch, James Flynn, and Pavel Golik. 2021. How might we create better benchmarks for speech recognition? In *Proceedings of the 1st workshop on benchmarking: Past, present and future*, pages 22–34.

Paul Boersma and David Weenink. 2009. Praat: doing phonetics by computer (Version 5.1.13).

Sue Brown and Steven Franks. 1995. Asymmetries in the Scope of Russian Negation. *Journal of Slavic Linguistics*, 3(2):239–287.

Elena A. Bryzgunova. 1975. The declarative-interrogative opposition in russian. *The Slavic and East European Journal*, 19(2):155.

Matthew S. Dryer. 2013. Polar Questions (v2020.3). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

Maria Esipova. 2025. Prosody across sentence types. *Semantics and Linguistic Theory*, pages 68–87.

Maria Esipova and Maribel Romero. 2023. Prejacent truth in rhetorical questions: Lessons from Russian. Talk at *Formal Approaches to Slavic Linguistics 32*.

Sanchit Gandhi, Patrick von Platen, and Alexander M. Rush. 2022. ESB: A Benchmark For Multi-Domain End-to-End Speech Recognition.

Hans-Martin Gärtner and Beáta Gyuris. 2017. On delimiting the space of bias profiles for polar interrogatives. *Linguistische Berichte*, 251:293–316.

Ljudmila Geist and Sophie Repp. 2023. Responding to negative biased questions in Russian. In Petr Biskup, Marcel Börner, Olav Mueller-Reichau, and Iuliia Shcherbina, editors, *Advances in formal Slavic linguistics 2021*, Open Slavic Linguistics. Language Science Press, Berlin.

Daniel Goodhue. 2022. Isn't there more than one way to bias a polar question? *Natural Language Semantics*, 30.

Lucas Rafael Stefanel Gris, Ricardo Marcacini, Arnaldo Candido Junior, Edresson Casanova, Anderson Soares, and Sandra Maria Aluísio. 2023.

Evaluating OpenAI's Whisper ASR for Punctuation Prediction and Topic Modeling of life histories of the Museum of the Person.

Tiina Keisanen. 2006. *Patterns of stance taking: Negative yes/no interrogatives and tag questions in American English conversation*. @Acta Universitatis Ouluensis. Oulun Yliopisto, Oulu.

Tracy Holloway King. 1994. Focus in Russian Yes-No Questions. *Journal of Slavic Linguistics*, 2(1):92–120.

Natasha Korotkova. 2023. Conversational dynamics of Russian questions with *razve*. In *Proceedings of Sinn und Bedeutung 27*, Prague. Institute of Czech Language & Linguistic Theory, Faculty of Arts, Charles University.

Natasha Korotkova. submitted. A new perspective on negative bias in polar questions: The view from Russian. In G. Walkden Eckardt, R. and N. Dehé, editors, *The Oxford Handbook of Non-Canonical Questions*.

Aleksandr Kutsakov, Alexandr Maximenko, Georgii Gospodinov, Pavel Bogomolov, and Fyodor Minkin. 2025. GigaAM: Efficient Self-Supervised Learner for Speech Recognition. In *Interspeech 2025*, pages 1213–1217.

D. Robert Ladd. 2008. *Intonational Phonology*. Cambridge University Press.

Aleksandr Meister, Matvei Novikov, Nikolay Karpov, Evelina Bakhturina, Vitaly Lavrukhin, and Boris Ginsburg. 2023. LibriSpeech-PC: Benchmark for Evaluation of Punctuation and Capitalization Capabilities of End-to-End ASR Models. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–7. IEEE.

Roland Meyer. 2004. Prosody, Mood, and Focus. A Study of so-called "intonationally marked" Yes-No Questions in Russian. *Formal Approaches to Slavic Linguistics 12*, pages 333–352. The Ottawa Meeting.

Roland Meyer and Ina Mleinek. 2006. How prosody signals force and focus—A study of pitch accents in Russian yes–no questions. *Journal of Pragmatics*, 38(10):1615–1635.

Margaret H. Mills. 1992. Conventionalized politeness in Russian requests: A pragmatic view of indirectness. *Russian Linguistics*, 16(1).

Maria Onoeva and Anna Staňková. 2025. Polar questions in Czech and Russian: An exploratory corpus investigation.

OpenAI. 2022. Whisper. `https://github.com/openai/whisper`. GitHub repository, accessed February 2026.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision.

Mariia Razguliaeva, Maria Onoeva, Radek Šimík, Roland Meyer, and Kateřina Hrdinková. to appear. Processing of Russian and Czech polar questions: Evidence for the effect of question bias. *Glossa Psycholinguistics*.

Steve Renals and Simon King. 2010. Automatic Speech Recognition.

Per Restan. 1972. *Sintaksis voprositel'nogo predloženija: obščij vopros [Syntax of interrogative sentences: polar questions]*. Universitetsforlaget, Oslo.

Salute Developers. 2023. GigaAM: Large-Scale Russian Automatic Speech Recognition. `https://github.com/salute-developers/GigaAM`. GitHub repository, accessed February 2026.

Silero Team. 2026. Silero Models: Pre-trained enterprise-grade STT/TTS models and benchmarks. GitHub repository, accessed February 2026.

Radek Šimík. to appear. *Polar question semantics and bias: Lessons from Slavic/Czech*.

Ksenia Zanon. 2024. Expletive Negation revisited: on some properties of negative polar interrogatives in Russian. *Journal of Slavic Linguistics*.

10. Appendices

Table 2: Performance across the ASR models (excluding Meta's MMS) for INTONPQs only under different contextual conditions. Accuracy (Acc.), improvement (Imp.) over baseline, and $95\,\%$ bootstrap confidence intervals (CI) are reported separately for positive and negative INTONPQs. Improvements are expressed in percentage points relative to the baseline (plain audio) condition from Table 1. Effects are classified as significant when the confidence interval does not include zero. The rest of the results are available on OSF.

Context	Positive INTONPQs				Negative INTONPQs			
	Acc.	Imp.	CI	Effect	Acc.	Imp.	CI	Effect
Whisper								
Pre 'Next Q'	78.10	71.90	[56.25, 87.50]	Improves	100.00	37.50	[21.88, 53.12]	Improves
Pre 'He asked'	68.80	62.50	[43.75, 78.12]	Improves	100.00	37.50	[21.88, 53.12]	Improves
Post 'yes'	90.60	84.40	[71.88, 96.88]	Improves	100.00	37.50	[21.88, 56.25]	Improves
Post 'no'	87.50	81.20	[65.62, 93.75]	Improves	100.00	37.50	[21.88, 53.12]	Improves
Pre 'Turtle'	12.50	6.20	[−6.25, 18.75]	No effect	96.90	34.40	[18.75, 50.00]	Improves
Post 'Turtle'	15.60	9.40	[−3.12, 21.88]	No effect	75.00	12.50	[3.12, 25.00]	Improves
Pre R&L	12.50	6.20	[−6.25, 18.75]	No effect	90.60	28.10	[12.50, 43.75]	Improves
Post R&L	3.10	−3.10	[−9.38, 0.00]	No effect	71.90	9.40	[−6.25, 25.00]	No effect
Pre 'He said'	25.00	18.80	[6.25, 34.38]	Improves	100.00	37.50	[21.88, 53.12]	Improves
GigaAM: RNNT								
Pre 'Next Q'	59.40	25.00	[3.12, 46.88]	Improves	100.00	0.00	[0.00, 0.00]	No effect
Pre 'He asked'	96.90	62.50	[46.88, 78.12]	Improves	100.00	0.00	[0.00, 0.00]	No effect
Post 'yes'	96.90	62.50	[46.88, 78.12]	Improves	96.90	−3.10	[−9.38, 0.00]	No effect
Post 'no'	96.90	62.50	[46.88, 78.12]	Improves	96.90	−3.10	[−9.38, 0.00]	No effect
Pre 'Turtle'	3.10	−31.20	[−46.88, −15.62]	Worsens	84.40	−15.60	[−28.12, −3.12]	Worsens
Post 'Turtle'	18.80	−15.60	[−31.25, 0.00]	No effect	87.50	−12.50	[−25.00, −3.12]	Worsens
Pre R&L	28.10	−6.20	[−28.12, 15.62]	No effect	100.00	0.00	[0.00, 0.00]	No effect
Post R&L	28.10	−6.20	[−25.00, 12.50]	No effect	93.80	−6.20	[−15.62, 0.00]	No effect
Pre 'He said'	6.20	−28.10	[−43.75, −12.50]	Worsens	65.60	−34.40	[−50.00, −18.75]	Worsens
GigaAM: CTC								
Pre 'Next Q'	87.50	31.20	[12.50, 50.00]	Improves	100.00	3.10	[0.00, 9.38]	No effect
Pre 'He asked'	100.00	43.80	[28.12, 59.38]	Improves	100.00	3.10	[0.00, 9.38]	No effect
Post 'yes'	100.00	43.80	[28.12, 62.50]	Improves	100.00	3.10	[0.00, 9.38]	No effect
Post 'no'	100.00	43.80	[28.12, 59.38]	Improves	100.00	3.10	[0.00, 9.38]	No effect
Pre 'Turtle'	25.00	−31.20	[−46.88, −15.62]	Worsens	93.80	−3.10	[−9.38, 0.00]	No effect
Post 'Turtle'	53.10	−3.10	[−18.75, 12.50]	No effect	96.90	0.00	[0.00, 0.00]	No effect
Pre R&L	87.50	31.20	[15.62, 46.88]	Improves	100.00	3.10	[0.00, 9.38]	No effect
Post R&L	50.00	−6.20	[−21.88, 9.38]	No effect	96.90	0.00	[0.00, 0.00]	No effect
Pre 'He said'	18.80	−37.50	[−53.12, −21.88]	Worsens	87.50	−9.40	[−21.88, 0.00]	No effect

Quantizing Whisper: How Design Choices Affect ASR Performance

Arthur Söhler[1], Julian Irigoyen[2], Andreas Søeborg Kirkedal[3]
[1] Copenhagen Business School, Copenhagen, Denmark
[2] Danske Bank, Copenhagen, Denmark
[3] Jabra (GN Group), Copenhagen, Denmark
arthursoehler@gmail.com, jiri@danskebank.dk, askirkedal@jabra.com

Abstract

Large speech recognition models like OpenAI's Whisper achieve high accuracy but are difficult to deploy in resource-constrained environments due to their high memory and computational demands. This matters for low-resource and on-device settings, where compute and memory constraints often limit the practical use and evaluation of ASR systems. To address this, we present a unified, cross-library evaluation of post-training quantization (PTQ) on Whisper-small, comparing supported configurations across quantization scheme, method, granularity, and bit-width. Our study is based on four libraries—*PyTorch*, *Optimum-Quanto*, *HQQ*, and *bitsandbytes*. Experiments on *LibriSpeech test-clean* and *test-other* show that dynamic *int8* quantization with *Optimum-Quanto* offers the best trade-off, reducing model size by 57% while lowering Word Error Rate below the baseline. Additional experiments on Whisper-base and Whisper-tiny confirm these trends, though with more pronounced degradation at lower bit-widths. Static quantization performed worse, likely due to the absence of efficient low-bit implementations for operations such as LayerNorm and Softmax. More aggressive formats (e.g., *nf4*, *int3*) achieved up to 71% compression at the cost of accuracy in acoustically challenging conditions. Our results demonstrate that carefully chosen PTQ methods can substantially reduce model size and inference cost without retraining, enabling efficient deployment of Whisper on constrained hardware.

Keywords: automatic speech recognition, neural network quantization, model compression

1. Introduction

Automatic speech recognition (ASR) has advanced rapidly with large-scale Transformer models (Vaswani et al., 2017) such as the Whisper family (Radford et al., 2023), which deliver state-of-the-art transcription accuracy across diverse languages and domains. However, this accuracy comes at a cost: models with hundreds of millions of parameters are difficult to deploy on edge devices, embedded systems, or latency-sensitive applications (Gholami et al., 2021; Wei et al., 2024). This challenge is especially pronounced in low-resource and on-device settings, where limited compute budgets and strict memory constraints can prevent both deployment and systematic evaluation. Post-training compression methods such as post-training quantization (PTQ) (Gholami et al., 2021; Nagel et al., 2021) therefore offer a practical path to making strong ASR models usable in settings where retraining and specialized hardware are not feasible.

Model compression techniques address this challenge by reducing computational cost in deep neural networks (Vanhoucke et al., 2011). Among them, *quantization* has emerged as one of the most practical and hardware-friendly approaches. By mapping 32-bit floating point weights and activations to lower-precision formats, quantization can shrink memory footprint and accelerate inference (Gray and Neuhoff, 1998). The choice of method—such as PTQ, which requires no retraining, or quantization-aware training (QAT), which incorporates quantization effects during training—determines the balance between efficiency and accuracy. While QAT often yields superior results under aggressive compression, PTQ is attractive for its speed and ease of deployment (Wu et al., 2020).

Recent years have seen rapid progress in low-bit quantization, including sub-8-bit formats (e.g., *int4*, *nf4*) and mixed-precision strategies designed to handle outlier activations (Dettmers et al., 2022; Shen et al., 2024). These methods have been evaluated extensively in computer vision and large language models, but their impact on ASR—and particularly on large-scale architectures like Whisper—remains underexplored.

In this paper, we take Whisper-small as a case study to evaluate post-training quantization across multiple libraries, methods, schemes, and bit-widths. We present a comparative study of Whisper PTQ in which all configurations are evaluated under a unified measurement protocol on both CPU and GPU, highlighting practical trade-offs between latency, accuracy, and compression across devices and tools. Our goal is to uncover how specific PTQ design choices translate into deployment behavior (efficiency and robustness) in realistic scenarios, where library support and overheads can materially affect outcomes. We additionally confirm our findings across model sizes by performing experiments on Whisper-base and Whisper-tiny, the two smallest variants of the Whisper family.

Proceedings of SPEAKABLE @ LREC 2026, pages 39–46

Our contributions are:

(1) Unified evaluation framework: Systematic cross-library comparison of PTQ for Whisper under a controlled, unified evaluation protocol, enabling direct comparison of library-specific trade-offs.

(2) Comparative analysis of PTQ design choices: We compare the effects of quantization scheme (dynamic vs. static), method (symmetric vs. asymmetric), granularity (per-tensor vs. per-channel), and bit-width (*int8* to *int3*/*nf4*) across representative libraries.

(3) Device-specific deployment insights: Parallel evaluation on CPU and GPU revealing different optimal configurations per device and acoustic condition (clean vs. acoustically challenging).

(4) Cross-model validation: Verification that the main findings hold across Whisper-small, Whisper-base, and Whisper-tiny, demonstrating robustness of deployment conclusions across model sizes.

2. Related Work

While its theoretical foundations date back decades (Gray and Neuhoff, 1998), modern neural network quantization has evolved rapidly, with several surveys providing systematic overviews (Gholami et al., 2021; Nagel et al., 2021; Wei et al., 2024). These works classify methods into post-training quantization and quantization-aware training, and outline factors such as bit-width, quantization granularity, and calibration strategies that strongly influence performance.

In Transformer-based architectures, Kim et al. (2021) proposed I-BERT, an integer-only Transformer that approximates nonlinear operations for complete deployment on integer hardware. Kim et al. (2022) further introduced a zero-shot static quantization approach for ASR models such as Conformer and QuartzNet, achieving $4\times$ compression without real training data by synthesizing calibration inputs. Wu et al. (2020) demonstrated that with per-channel weight quantization and entropy-based activation calibration, PTQ can maintain accuracy within 1% of full precision.

Recent efforts have focused on stabilizing quantization for large Transformer models by handling outlier activations. Dettmers et al. (2022) introduced LLM.int8(), a mixed-precision method that preserves high precision for outlier dimensions, enabling robust 8-bit inference for models up to 175B parameters. Shen et al. (2024) explored low-precision floating-point formats such as *fp8*, reporting improved accuracy and flexibility compared to *int8* in outlier-heavy models.

Prior work has also examined Whisper quantization in specific toolchains and limited method sets. For example, Andreyev (2025) evaluates different bit-widths using *whispercpp* on *LibriSpeech* corpus (Panayotov and Povey, 2015) and reports latency and accuracy trade-offs for edge deployment scenarios. In contrast, our work focuses on a controlled, cross-library comparison spanning multiple PTQ libraries and design choices (scheme, granularity, and bit-width) under one measurement protocol across CPU and GPU.

More recent work has explored low-bit quantization specifically for speech and audio models. Feng et al. (2025) propose Edge-ASR, a comprehensive benchmark of post-training quantization methods applied to Whisper and other edge-ASR models. Kang and Kim (2025) propose GenPTQ, a mixed-precision post-training quantization method that performs efficient layer-wise bit allocation based on gradient-driven sensitivity analysis, achieving strong compression with minimal WER degradation. Beyond speech recognition, Khandelwal and Fuentes (2025) study post-training quantization for audio diffusion transformers, highlighting challenges related to activation variability and outliers in transformer-based audio models. These works primarily focus on model-specific optimization strategies, whereas our study provides a cross-library evaluation of quantization approaches under a unified experimental framework.

While most prior work has focused on vision and text models, the techniques—especially sub-8-bit quantization, mixed-precision schemes, and outlier handling—are directly relevant to ASR. However, to our knowledge, there is no comparative evaluation of quantization libraries that jointly contrasts (i) activation method (dynamic vs. static), (ii) granularity (per-channel vs. per-tensor), (iii) weight-only vs. weight+activation quantization, and (iv) multiple bit-widths, while evaluating accuracy, latency, and compression on both CPU and GPU under a single protocol. This work addresses that gap by assessing PTQ strategies and their trade-offs for Whisper-small, Whisper-base, and Whisper-tiny.

3. Background

Quantization reduces the memory and computational cost of neural networks by mapping high-precision values (typically 32-bit floating point) to lower-precision representations. In practice, this means storing weights and activations in compact formats such as *int8* or *fp8*, with memory usage scaling approximately as $b/32$ relative to *fp32*, yielding a compression ratio of $32/b$ where b is the bit-width (Gholami et al., 2021; Nagel et al., 2021).

Formally, a quantizer can be written as

$$Q(x) = \text{clip}\Big(\text{round}\big(\tfrac{x}{s}\big), q_{\min}, q_{\max}\Big), \qquad (1)$$

where s is a scaling factor, and $[q_{\min}, q_{\max}]$ defines the representable integer range for a given bit-width.

Subset	Hours	Minutes/ speaker	Female speakers	Male speakers	Total speakers
test-clean	5.4	8	20	20	40
test-other	5.1	10	17	16	33

Table 1: Dataset statistics for *test-clean* and *test-other*.

The effectiveness of this mapping depends not only on the chosen bit-width, but also on how scales are applied across the network.

This is captured by the notion of *granularity*. In *per-tensor* quantization, a single scale is shared across an entire tensor, which is efficient but may distort values when distributions vary widely. *Per-channel* quantization instead assigns a scale to each output channel, better capturing local statistics and improving accuracy in deep Transformers, though at higher storage and compute cost (Gholami et al., 2021). A middle ground is *per-group* quantization, which clusters channels into groups with shared scales to balance efficiency and accuracy (Yao et al., 2022).

Beyond granularity, quantization can be applied either post-training (PTQ), which avoids retraining but may reduce accuracy at low bit-widths, or during training (QAT), which improves robustness at the cost of additional training (Han et al., 2015; Gholami et al., 2021).

Additionally, quantization methods differ on how they distribute values within the quantized range. Symmetric quantization centers values around zero, whereas asymmetric quantization shifts the zero-point to better match non-zero-centered distributions of weights (Nagel et al., 2021).

Finally, schemes differ in how scaling factors and zero-points are determined. *Static quantization* fixes them in advance using calibration data, while *dynamic quantization* computes them at runtime.

These design choices become especially important for Transformer-based ASR models such as Whisper-small. Operations like *LayerNorm*, *Softmax*, and *GELU* are highly sensitive to reduced precision, and activations often exhibit heavy-tailed distributions (Dettmers et al., 2022; Shen et al., 2024). As a result, quantization can compromise robustness, underscoring the need for careful selection of quantization strategies in deployment.

4. Data

We conduct our experiments on the English-language subsets *test-clean* and *test-other* of the *LibriSpeech* corpus (Panayotov and Povey, 2015). Dataset statistics are shown in Table 1.

These test sets have been used to evaluate ASR systems for many years. *test-clean* represents an easy evaluation task, whereas *test-other* is a more challenging dataset. The original *LibriSpeech* data consists of read-aloud books and was divided into a *clean* and an *other* partition based on Word Error Rate (WER) scores produced by a hybrid ASR system with an acoustic model trained on a subset of the Wall Street Journal corpus and a bigram LM estimated from the text of the respective books. The speakers were ranked according to WER and the data partitioned into two sets of roughly equal sizes. *test-clean* consists of randomly selected speakers from the *clean* partition. *test-other* specifically consists of speakers from the 3rd quartile according to WER. This sampling is intended to create a more challenging test dataset. Gender balance is ensured at the speaker level (Panayotov et al., 2015).

5. Methods

We evaluate PTQ on Whisper-small, a 244M-parameter Transformer-based ASR model pre-trained for multilingual and multitask speech recognition (Radford et al., 2023).

Depending on library support, we apply quantization across a range of bit-widths (*int8*, *int4*, *int3*, *nf4*, and *fp8*) and compare four widely used PTQ libraries: *PyTorch*, offering native dynamic *int8* quantization on CPU; *Optimum-Quanto* (hereafter *Quanto*), supporting both integer and low-precision floating formats across CPU, GPU, and MPS backends; *HQQ*, a quantization library based on half-quadratic optimization with configurable group sizes; and *bitsandbytes (BNB)*, which provides GPU-only implementations of normalized formats such as *nf4* with optional double quantization. Together, these libraries cover a diverse spectrum of formats, methods, and quantization schemes, making them representative of the practical choices available to users when designing model compression pipelines.

Performance is evaluated along three dimensions. Accuracy is measured using WER and Character Error Rate (CER). Efficiency is captured through the Real-Time Factor (RTF), quantifying inference speed relative to audio duration. Finally, we report model size reduction relative to the full-precision *fp32* baseline. To ensure comparability, all configurations were evaluated under the same preprocessing pipeline, dataset splits, batch size, decoding procedure, hardware selection, and timing protocol. RTF was computed as total timed generation time divided by total audio duration over the evaluation set. Because all measurements were collected on the HPC infrastructure described in Section 10, the reported RTF values should be interpreted primarily as relative comparisons across configurations under a fixed setup, rather than as direct estimates of edge-device latency. Inference time was measured only around the model genera-

Device/Method	WER_c	WER_o	CER_c	CER_o	RTF	Size Red.
			CPU			
Baseline (fp32)	3.48	11.88	1.02	3.62	0.121	–
PyTorch int8 (dyn.)	3.72	13.67	1.11	4.21	0.077	57%
HQQ int4 (dyn.)	3.52	14.09	1.09	4.38	0.155	69%
Quanto int8/fp8 (stat.)	5.95	15.92	1.83	5.03	0.169	57%
			GPU			
Baseline (fp32)	3.48	11.88	1.02	3.62	0.006	–
Quanto int8 (dyn.)	3.41	10.65	0.97	3.29	0.008	57%
BNB nf4 (dyn.)	3.54	13.49	1.05	4.05	0.008	70%
HQQ int3 (dyn.)	4.11	12.93	1.22	3.77	0.019	71%

Table 2: Selected best-performing dynamic (dyn.) and static (stat.) quantization configurations for Whisper-small on *LibriSpeech test-clean* (c) and *test-other* (o).

Device/Method	WER_c	WER_o	CER_c	CER_o	RTF	Size Red.
			Whisper-base			
Baseline (fp32)	5.08	12.87	1.93	6.76	0.0022	–
Quanto int8 (dyn.)	5.10	14.72	1.93	7.60	0.0037	36.2%
BNB nf4 (dyn.)	6.05	18.09	2.31	10.16	0.0036	52.1%
			Whisper-tiny			
Baseline (fp32)	7.60	23.69	2.98	12.84	0.0015	–
Quanto int8 (dyn.)	7.64	24.64	2.99	13.24	0.0025	19.3%
BNB nf4 (dyn.)	11.16	32.19	4.78	20.20	0.0026	37.5%

Table 3: Quantization results for Whisper-base and Whisper-tiny using the strongest dynamic (dyn.) configurations on *LibriSpeech test-clean* (c) and *test-other* (o).

tion step, with CUDA synchronization on GPU and warm-up runs before timed evaluation. Audio inputs were converted to model input features using the standard WhisperProcessor from the Transformers library. In the adopted Transformers implementation, Whisper used its default fixed 30 s front end; accordingly, shorter utterances were zero-padded to the 30 s window. The reported RTF values therefore reflect this fixed-window inference setup rather than variable-length audio processing. Relative RTF differences nevertheless remain directly comparable across quantization settings because all configurations were evaluated under the same protocol.

For static quantization, calibration used a randomly selected 10% subset of the full evaluation data, processed with the same feature-extraction pipeline as the test data.

While the main study focuses on Whisper-small, we also run a limited follow-up study on Whisper-tiny and Whisper-base under the same dataset, settings, and measurement procedure. The cross-model follow-up is intended as a validation of the main deployment trends, not as a second exhaustive benchmark. We therefore restrict Whisper-tiny and Whisper-base to the strongest dynamic configurations from the Whisper-small study. Static quantization is omitted in this follow-up because it underperformed on Whisper-small and would substantially increase experimental volume without changing the main practical conclusion.

6. Results

6.1. Quantizing Whisper-small

Table 2 summarizes the best-performing quantized models relative to the full-precision baseline.

On **CPU**, *PyTorch* dynamic *int8* delivered the fastest inference (RTF 0.077; 36.4% faster than the 0.121 baseline) with only a small accuracy drop. *HQQ* dynamic *int4* preserved accuracy on clean speech while achieving the largest compression (69%). In contrast, static *int8*/*fp8* quantization with *Quanto* degraded performance substantially, confirming that Whisper's architecture is ill-suited to static quantization.

On **GPU**, *Quanto* dynamic *int8* not only matched but slightly outperformed the baseline on the more challenging *test-other* split (WER_{other} = 10.65, CER_{other} = 3.29), while reducing model size by 57%. *BNB nf4* offered a 70% size reduction with minimal accuracy loss on *test-clean*, but suffered on *test-other*. *HQQ int3* achieved the smallest size (71% reduction) but at the cost of higher error rates.

Overall, dynamic quantization consistently outperformed static quantization in both accuracy and speed. *int8* proved the most reliable setting across devices, while more aggressive formats (*nf4*, *int3*) enabled extreme compression but compromised robustness, particularly under acoustically challenging conditions.

6.2. Scaling Across Whisper Model Sizes

To assess whether the main conclusions are specific to Whisper-small or stable within the Whisper model family, we evaluate Whisper-tiny and Whisper-base on *LibriSpeech* splits using the strongest quantization configurations from the Whisper-small study. Table 3 shows that the overall trends remain consistent across model sizes: dynamic *int8* remains the most robust operating point, while more aggressive formats trade robustness (especially on *test-other*) for additional compression.

7. Discussion

7.1. Trade-offs Between Different Quantization Methods

On CPUs, *PyTorch* dynamic *int8* consistently achieved the fastest inference. Its advantage likely stems from using a per-tensor asymmetric scheme, which applies a single scale across an entire tensor. This approach simplifies computation and reduces the overhead of quantization and dequantization, explaining the strong runtime performance. The trade-off, however, is lower representational flexibility with quantization parameters only calculated per-tensor, which contributed to weaker robustness on the acoustically diverse *test-other* dataset. However, with a 57% size reduction, it still offers clear advantages over the baseline.

On the GPU, *Quanto* dynamic *int8* took a different approach. It prioritizes accuracy over speed. It was slower but outperformed the *fp32* baseline on WER and CER on both *test-clean* and *test-other*, with the most striking improvement seen under acoustically challenging conditions. One plausible explanation is its symmetric, per-channel scheme, which aligns well with Whisper's near-zero-centered weight distributions and assigns independent scales to each channel. This granularity can better preserve fine-grained variation across channels, yielding higher accuracy but at the cost of additional computational overhead. Similar to the *PyTorch* model, this model comes with a 57% size reduction, a strong compression relative to the baseline.

Libraries enabling more aggressive formats, such as *HQQ* (*int4*, *int3*) and *BNB* (*nf4*), pushed compression further—up to 70%—but consistently degraded robustness, especially in acoustically challenging environments. These results highlight that precision below 8-bit remains less reliable in real-world ASR applications.

Taken together, these comparisons show that library differences are largely driven by the schemes and bit-widths they implement. Per-tensor quantization favors efficiency, making *PyTorch int8* attractive when low latency is the main deployment priority. Per-channel quantization favors accuracy, as shown by *Quanto int8*, which is more suitable when robustness is critical. More aggressive low-bit formats are best reserved for deployments where extreme compression outweighs the need for reliability.

The additional results on Whisper-base and Whisper-tiny (Table 3) follow the same qualitative pattern: dynamic *int8* remains the most robust operating point, while *nf4* increases compression but incurs a larger error increase—especially on *test-other*. This suggests that the robustness–compression trade-off observed for Whisper-small is stable across smaller Whisper variants, even though absolute WER/CER differs by model size.

7.2. Dynamic vs. Static Quantization

In theory, static quantization should reduce runtime overhead by fixing scales in advance, trading a small loss in accuracy for faster inference. Surprisingly, in Whisper-small we observed the opposite: static quantization was both slower and less accurate.

One possible explanation is that operations such as *LayerNorm* and *Softmax* lack efficient low-bit implementations, forcing repeated dequantization that cancels the expected speed gains of static quantization. On top of this, as expected, fixed calibration scales limited robustness under shifting distributions, especially in acoustically challenging conditions.

By contrast, dynamic quantization adapted better at runtime, preserving accuracy and delivering faster inference. This makes dynamic quantization the most reliable choice in our evaluation for Whisper-small, despite theoretical expectations to the contrary.

7.3. Clean vs. Acoustically Challenging Speech

Across nearly all configurations, quantized models suffered larger accuracy drops on *test-other* than on *test-clean*, indicating reduced robustness under the more challenging conditions represented by the *test-other* split. As shown in Figure 1, lower-bit formats such as *BNB nf4* and *HQQ int3* showed the steepest increases in WER, underlining how aggressive compression disproportionately reduces robustness.

Interestingly, *Quanto* dynamic *int8* was an exception: it not only matched the baseline on *test-clean* but also outperformed it on *test-other*. This suggests that under certain conditions, quantization can act as a form of regularization, stabilizing predictions in acoustically challenging environments. However, this effect appears limited to moderate

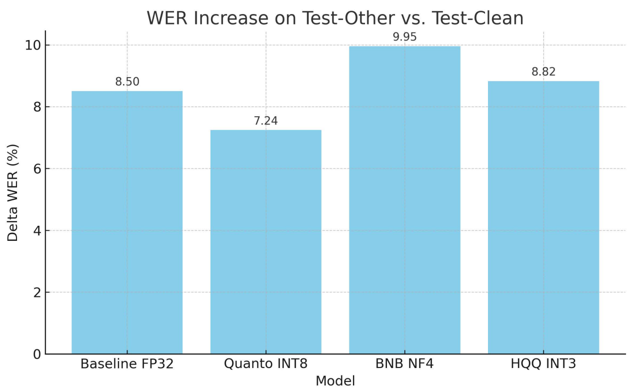


Figure 1: WER increase on *test-other* relative to *test-clean* for selected quantized models of Whisper-small. Lower-bit-width configurations (*nf4*, *int3*) show larger deltas, highlighting the trade-off between compression and robustness.

Priority	Best Model	Trade-off
Fastest Inference	PyTorch Dynamic int8	Higher WER compared to baseline
Best Accuracy	Quanto Dynamic int8	Higher RTF compared to baseline
Best Size Reduction	HQQ Dynamic int3	Higher WER and RTF compared to baseline

Table 4: Comparative Summary of Quantization Methods - Optimal Approaches Based on Deployment Priorities

settings (e.g., *int8*); once precision drops further, errors compound rapidly under adverse acoustic conditions. A similar robustness gap between *test-clean* and *test-other* is also evident for Whisper-base and Whisper-tiny, where *nf4* degrades more strongly than *int8* on *test-other*.

For deployment, this highlights that 8-bit precision is a safe minimum in real-world scenarios, while lower-bit formats should be applied selectively and only when acoustic variability is less of a concern.

7.4. Deployment Implications

Our results point to clear deployment strategies for Whisper-small and its variants under different constraints. Table 4 summarizes these strategies as rule-of-thumb guidance derived from our evaluation protocol. On CPUs, *PyTorch* dynamic *int8* offers the best speed–accuracy trade-off, while *HQQ* dynamic *int4* offers stronger compression when memory is the tighter constraint. On GPUs, *Quanto* dynamic *int8* is the most robust choice, delivering a 57% size reduction while preserving or even improving accuracy, including in acoustically challenging conditions. More aggressive formats such as *HQQ int3* or *BNB nf4* are suitable only in memory-constrained scenarios where some accuracy loss is acceptable, and ideally applied selectively to less critical layers. Finally, in acoustically diverse environments, at least 8-bit precision should be maintained for sensitive components such as attention and *LayerNorm* to avoid robustness degradation. These guidelines emphasize that quantization choices must be tailored not only to model architecture, but also to deployment context, device, and acoustic variability.

8. Conclusion

This study evaluated post-training quantization of Whisper-small across four libraries and multiple bit-widths, providing a controlled cross-library comparison of supported scheme, method, and granularity choices. On *LibriSpeech test-clean* and *test-other*, dynamic quantization consistently outperforms static quantization in both accuracy and inference speed for this architecture. *Quanto* dynamic *int8* emerged as the best overall configuration for GPU deployment, achieving 57% model-size reduction while matching baseline accuracy on *test-clean* and even exceeding it on *test-other*. On CPU, *PyTorch* dynamic *int8* delivered the fastest inference in our experiments, while *HQQ* dynamic *int4* offers a strong compromise when memory is limited.

From a high-level perspective, our results indicate that within this model family and library-supported configurations, dynamic quantization was more reliable than static quantization. Furthermore, the most accurate results were obtained with configurations using symmetric, per-channel quantization rather than asymmetric, per-tensor approaches, especially under acoustically variable conditions. In practice, *int8* is the most reliable precision across devices; more aggressive formats (e.g., *nf4*, *int3*) provide larger compression but degrade robustness on *test-other*. We also observe a mild regularization effect: dynamic *int8* can match or even exceed *fp32* on acoustically challenging speech, but this benefit does not persist below 8-bit precision. Taken together, the *scheme*, *method*, and *granularity* materially shape the accuracy–speed–size trade-off and should be chosen to match device constraints and expected acoustic conditions. Ultimately, our findings show that thoughtful quantization design enables deployment of more advanced speech recognition models in resource-constrained environments.

9. Limitations and Future Work

First, this work is designed as a controlled study to compare cross-library PTQ effects under a shared evaluation protocol. Because the study is constrained by library support, not all combinations of quantization scheme, method, granularity, and bit-width could be evaluated; the comparison is therefore structured rather than exhaustive. We restrict evaluation to the *LibriSpeech test-clean* and *test-other* splits, which provide a reproducible benchmark but do not capture the full diversity of domains, accents, and spontaneous speech encountered in practice. Extending the evaluation to broader domains, including multilingual data, is an important next step.

Second, we focus on PTQ to reflect scenarios where retraining is infeasible; quantization-aware training may achieve higher accuracy at more aggressive bit-widths. In addition, ONNX Runtime is excluded because it relies on a different execution stack involving model export and graph-level optimizations, introducing additional variables that make its quantization and runtime behavior not directly comparable to the native-library workflows evaluated in this study. Both represent complementary directions for future work.

Third, our analysis targets the smaller Whisper variants as representative Transformer-based ASR models. Larger variants such as Whisper-medium and Whisper-large are excluded from this study, as our focus is on smaller models that are most relevant for deployment in resource-constrained environments. Results may differ for the larger Whisper models or for other ASR architectures.

Future work will expand evaluation to additional datasets and model families, and investigate mixed-precision and layer-wise strategies that apply aggressive quantization selectively while preserving accuracy in sensitive components.

10. Acknowledgements

We thank Jabra and GN Group for supporting this research. Computational experiments were performed on the Danish e-Infrastructure Consortium (DeiC) National HPC facilities, utilizing Lenovo ThinkSystem SR675 V3 nodes equipped with dual AMD EPYC 9454 processors (2.75 GHz, 192 vCPUs total), 768 GB DDR5-4800 memory, and four NVIDIA Hopper H100-SXM5 GPUs (80 GB HBM3) per node.

11. Bibliographical References

Allison Andreyev. 2025. Quantization for OpenAI's Whisper Models: A Comparative Analysis. *arXiv preprint arXiv:2503.09905*.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. GPT3.int8(): 8-bit Matrix Multiplication for Transformers at Scale. In *Advances in Neural Information Processing Systems*, volume 35, pages 30318–30332. Curran Associates, Inc.

Chen Feng, Yicheng Lin, Shaojie Zhuo, Chenzheng Su, Ramchalam Kinattinkara Ramakrishnan, Zhaocong Yuan, and Xiaopeng Zhang. 2025. Edge-ASR: Towards Low-Bit Quantization of Automatic Speech Recognition Models. *arXiv preprint arXiv:2507.07877*.

Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W. Mahoney, and Kurt Keutzer. 2021. A Survey of Quantization Methods for Efficient Neural Network Inference. *arXiv preprint arXiv:2103.13630*.

R.M. Gray and D.L. Neuhoff. 1998. Quantization. *IEEE Transactions on Information Theory*, 44(6):2325–2383.

Song Han, Jeff Pool, John Tran, and William Dally. 2015. Learning both Weights and Connections for Efficient Neural Networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Beom Jin Kang and Hyun Kim. 2025. GenPTQ: Green Post-Training Quantization for Large-Scale ASR Models with Mixed-Precision Bit Allocation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10704–10718, Suzhou, China. Association for Computational Linguistics.

Tanmay Khandelwal and Magdalena Fuentes. 2025. Post-training quantization for audio diffusion transformers. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2025, Tahoe City, CA, USA, October 12-15, 2025*, pages 1–5. IEEE.

Sehoon Kim, Amir Gholami, Zhewei Yao, Nicholas Lee, Patrick Wang, Aniruddha Nrusimha, Bohan Zhai, Tianren Gao, Michael W. Mahoney, and Kurt Keutzer. 2022. Integer-Only Zero-Shot Quantization for Efficient Speech Recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4288–4292.

Sehoon Kim, Amir Gholami, Zhewei Yao, Michael W. Mahoney, and Kurt Keutzer. 2021. I-BERT: Integer-only BERT Quantization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5506–5518. PMLR.

Markus Nagel, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart van Baalen, and Tijmen Blankevoort. 2021. A White Paper on Neural Network Quantization. *arXiv preprint arXiv:2106.08295*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. LibriSpeech: an ASR corpus based on public domain audio books. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5206–5210. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Haihao Shen, Naveen Mellempudi, Xin He, Qun Gao, Chang Wang, and Mengni Wang. 2024.

Efficient Post-training Quantization with FP8 Formats. In *Proceedings of Machine Learning and Systems*, volume 6, pages 483–498.

Vincent Vanhoucke, Andrew Senior, and Mark Z. Mao. 2011. Improving the speed of neural networks on CPUs. In *Deep Learning and Unsupervised Feature Learning Workshop, NIPS 2011*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Lu Wei, Zhong Ma, Chaojie Yang, and Qin Yao. 2024. Advances in the Neural Network Quantization: A Comprehensive Review. *Applied Sciences*, 14(17).

Hao Wu, Patrick Judd, Xiaojie Zhang, Mikhail Isaev, and Paulius Micikevicius. 2020. Integer Quantization for Deep Learning Inference: Principles and Empirical Evaluation. *arXiv preprint arXiv:2004.09602*.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. 2022. ZeroQuant: Efficient and Affordable Post-Training Quantization for Large-Scale Transformers. In *Advances in Neural Information Processing Systems*, volume 35, pages 27168–27183. Curran Associates, Inc.

12. Language Resource References

Panayotov, Vassil and Povey, Daniel. 2015. *LibriSpeech ASR corpus*. OpenSLR. PID https://www.openslr.org/12.

"OK Aura, Be Fair With Me": Demographics-Agnostic Training for Bias Mitigation in Wake-up Word Detection

Fernando López[1,2], **Paula Delgado-Santos**[1]
Pablo Gómez[1], **David Solans**[1], **Jordi Luque**[1]
[1]Telefónica Innovación Digital, Madrid, Spain
[2]Universidad Autónoma de Madrid, Madrid, Spain
{fernando.lopez, paula.delgadodesantos, pablo.gomezguerrero,
david.solansnoguero, jordi.luque}@telefonica.com

Abstract

Voice-based interfaces are widely used; however, achieving fair Wake-up Word detection across diverse speaker populations remains a critical challenge due to persistent demographic biases. This study evaluates the effectiveness of demographics-agnostic training techniques in mitigating performance disparities among speakers of varying sex, age, and accent. We utilize the OK Aura database for our experiments, employing a training methodology that excludes demographic labels, which are reserved for evaluation purposes. We explore (i) data augmentation techniques to enhance model generalization and (ii) knowledge distillation of pre-trained foundational speech models. The experimental results indicate that these demographics-agnostic training techniques markedly reduce demographic bias, leading to a more equitable performance profile across different speaker groups. Specifically, one of the evaluated techniques achieves a Predictive Disparity reduction of 39.94% for sex, 83.65% for age, and 40.48% for accent when compared to the baseline. This study highlights the effectiveness of label-agnostic methodologies in fostering fairness in Wake-up Word detection.

Keywords: Wake-up word, fairness, bias, demographics-agnostic

1. Introduction

Voice-based interfaces are now central to human-computer interaction, enabling virtual assistants, hands-free messaging, and applications such as customer support and clinical/legal transcription. The entry point to most of these systems is a Wake-up Word (WuW); a predefined trigger phrase that, once detected by an always-on lightweight acoustic model, activates the device and initiates interaction with the user (Këpuska and Breitfeller, 2006; Këpuska, 2011; López et al., 2023; López-Espejo et al., 2021). However, speech systems often exhibit performance disparities across demographic groups such as sex, age, and accent (Attanasio et al., 2024). Acoustic variability can systematically affect and raise fairness concerns (Fuckner et al., 2023). In always-on settings, these disparities manifest not only as aggregate error-rate gaps but as unequal *interactional burdens* (Choi and Choi, 2025): some users must repeat commands or alter their speech more often than others to obtain the same functionality. WuW detection is particularly susceptible because decisions rely on short speech segments with limited context, which can amplify speaker-dependent variability and reduce reliability for children, older adults, and regional or non-native speakers.

These concerns are consistent with prior work documenting demographic bias across speech tasks, including speaker identification, phoneme recognition, intent classification, keyword spotting (KWS), and emotion recognition (ER) (Meng et al., 2022; Hutiri et al., 2023; Slaughter et al., 2023). In Automatic Speech Recognition (ASR), higher Word Error Rates (WER) are repeatedly reported for speakers with regional or non-native accents, with additional disparities linked to sex, age, and intersectional factors (Garg et al., 2018; Zolnoori et al., 2024; Harris et al., 2024; Feng et al., 2021; Martin and Wright, 2023). Evaluations of large foundation models such as Whisper (Radford et al., 2023) further corroborate persistent racial, sex, and dialect biases, frequently favoring majority or privileged groups (Fuckner et al., 2023; Slaughter et al., 2023; Hutiri et al., 2023). Similar patterns have been observed in KWS and ER, where systems often underperform for children, elderly speakers, and nonstandard accents (Mujtaba et al., 2024; Hutiri et al., 2023; Feng et al., 2021; Martin and Wright, 2023). Recent benchmark efforts such as Fair-Speech for ASR (Veliche et al., 2024) and FaiST for broader speech technology (Jahan et al., 2025) further document systematic performance gaps across multiple demographic attributes, underscoring the need for dedicated fairness analyses in speech interfaces.

Several methodological tools have been proposed to diagnose and mitigate these disparities. For instance, DivExplorer can automatically identify attribute combinations (e.g., sex, age, accent) associated with large performance gaps (Pastor

Proceedings of SPEAKABLE @ LREC 2026, pages 47–58
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

et al., 2021). Building on such diagnostics, mitigation frameworks such as CLUES use discovered subgroups to guide contrastive learning and reduce disparities by targeting underperforming cohorts in the representation space (Koudounas et al., 2024). In parallel, Slaughter et al. (2023) demonstrated that embeddings from pre-trained speech models, including Whisper, wav2vec 2.0, WavLM, and HuBERT, can encode and amplify social biases. To measure this directly within the embedding spaces, they developed the Speech Embedding Association Test (SpEAT). Building on this area of research, Lin et al. (2024) examined how specific architectural and data choices in self-supervised learning (SSL) impact social biases in downstream tasks.

Beyond specific methods and datasets, recent work argues that speech recognition fairness is inherently context-dependent and multi-metric rather than "one-size-fits-all", calling for benchmarks that consider task requirements, deployment constraints, and stakeholder needs (ElGhazaly et al., 2025; Veliche et al., 2024).

Nonetheless, many mitigation strategies depend on explicit demographic labels, which are often unavailable, incomplete, or privacy-sensitive in real-world deployments, and data scarcity for underrepresented groups remains a persistent obstacle (Dheram et al., 2022; Barocas and Selbst, 2016). Other approaches pursue fairness via personalization, for instance, by conditioning KWS models on speaker-specific embeddings to improve performance for underrepresented users (Labrador et al., 2025). While effective, such methods typically require additional user data and enrollment procedures, and may not be feasible for compact, always-on WuW detectors operating entirely on device.

Motivated by these limitations, we study demographic bias in WuW detection and develop mitigation strategies that do not require demographic labels during training. We (i) quantify demographic disparities across sex, age, and accent using groupwise analyses and fairness metrics such as Predictive Disparity (PD) and Disparate Impact (DI), and (ii) investigate demographics-agnostic training methods based on generalization-oriented data augmentation and knowledge distillation/transfer from large self-supervised foundation speech models. Experiments on a real-world Spanish WuW dataset ("OK Aura") show that these label-free strategies substantially reduce demographic bias while preserving overall detection performance.

2. Mitigation Methodology

We implement a mitigation pipeline that remains demographics-agnostic during training, reserving demographic labels strictly for post-hoc bias evaluation. First, we identify bias within the dataset to identify demographic groups underrepresented in the training and validation phases. Subsequently, we assess bias reflected in WuW classifier predictions. Then, we adopt demographics-agnostic training methodologies intended to alleviate those biases.

This choice is motivated by the high cost and practical barriers of collecting additional data for underrepresented groups (e.g., privacy and limited access). Even with balanced data, disparities can persist due to design choices (Hutiri et al., 2023) or feature selection (Bailey and Plumbley, 2021); while demographics-aware methods can reduce bias (Dheram et al., 2022), we target mitigation without explicit demographic conditioning.

Our methodology employs two demographics-unaware training strategies. First, we hypothesize that modulating or partially removing frequency information during training discourages the model from relying on demographic-correlated acoustic cues. Given that sex, age, and accent are known to correlate with F0 (fundamental frequency) and formant structure (Vorperian et al., 2019), spectral envelope (Harnsberger et al., 2008), and prosody (Piat et al., 2008), respectively. By disrupting these cues, the model could be encouraged to learn more invariant representations (Vandenberghe et al., 2023). To this end, we explore data augmentation techniques applied at the spectrum level (Section 2.1). Second, large pre-trained SSL models, trained on diverse audio, have been shown to suppress speaker identity in their upper layers (Mohamed et al., 2022). Furthermore, as some models have been scaled to encompass over 4 million hours of training data (Barrault et al., 2023), we hypothesize that they capture demographically robust representations. Therefore, we investigate using such models as teachers to train a compact, robust student model (Chai et al., 2022) (Section 2.2).

2.1. Data augmentation techniques

We consider both time-domain and time-frequency-domain augmentations. Given an input waveform $x \in \mathbb{R}^N$, we compute a time–frequency representation via the Short-Time Fourier Transform (STFT) and use its magnitude to form a spectrogram,

$$X = \mathrm{Spectrogram}(x) = |\mathrm{STFT}(x)|^2, \qquad (1)$$

where $X \in \mathbb{R}^{T \times F}$, T denotes the number of time frames, and F the number of frequency bins. For augmentations applied in the time-frequency domain, we modify the magnitude to obtain X' while preserving the original phase, and then reconstruct an augmented waveform x' using the inverse STFT (ISTFT).

FreqMixStyle: it mixes frequency-wise feature statistics between samples to promote domain-invariant representations. It is motivated by frequency-wise instance normalization analyses for audio domain generalization (Kim et al., 2022). In practice, we normalize a spectrogram X_i along the frequency axis and re-scale it using mixed statistics from another randomly selected spectrogram X_j:

$$\mu_{\text{new}} = \lambda\mu_i + (1-\lambda)\mu_j, \quad \sigma_{\text{new}} = \lambda\sigma_i + (1-\lambda)\sigma_j \tag{2}$$

where $\lambda \sim \mathrm{Beta}(\alpha, \alpha)$ controls the interpolation strength.

FilterAugment: simulates acoustic filtering by applying smooth frequency-dependent gains rather than masking entire bands (Nam et al., 2022). Given a spectrogram X, we apply a multiplicative weighting mask

$$X' = X \odot W_{\mathrm{FA}}, \tag{3}$$

where $W_{\mathrm{FA}} \in \mathbb{R}^{T \times F}$ contains frequency-dependent weights and $\odot$ denotes element-wise multiplication. We use the linear variant, which linearly interpolates gains across frequency to avoid abrupt discontinuities (Nam et al., 2022).

Frequency masking: as a strong baseline augmentation, we also apply frequency masking from SpecAugment (Park et al., 2019). We sample the mask width $f \sim \mathcal{U}(0, W_F)$ and starting index $f_0 \sim \mathcal{U}(0, \nu - f)$, where ν is the number of mel channels, and set the band $[f_0, f_0 + f)$ to zero:

$$X' = \mathrm{FreqMask}(X). \tag{4}$$

This technique has been shown to improve model robustness by forcing the network to learn from partial spectrograms, thus improving generalization. It is especially effective in situations where the model must handle varying acoustic conditions or incomplete audio inputs (Kim et al., 2021).

Device Impulse Responses (DIR). Impulse responses model how a capture device filters an input signal. Originally, it was presented for device generalization by simulating microphone characteristics. Nonetheless, it modulates frequencies by convolving each training utterance with a sampled device impulse response h_{dir} (Morocutti et al., 2023):

$$x' = x * h_{\mathrm{dir}}. \tag{5}$$

To keep input dimensions consistent, we truncate the convolved signal to match the original length.

2.2. Speech Self-Supervised Learning models

SSL has become a key approach for learning robust speech representations from large amounts of unlabeled audio (Mohamed et al., 2022; Chen et al., 2022; Han et al., 2025; Wang et al., 2024). In WuW settings, SSL models can be particularly useful under limited labeled data and challenging acoustic conditions (Yu et al., 2023; Mørk et al., 2024).

We leverage a large SSL encoder to build a high-capacity teacher classifier (`w2v-BERT2-kws`) and then distill its knowledge into a compact WuW student model. Specifically, we use the w2v-BERT 2.0 pre-trained model (Barrault et al., 2023), a Conformer-based multilingual speech encoder trained on 4.5M hours of unlabeled audio. The `w2v-BERT2-kws` architecture is depicted in Figure 1. We leverage the w2v-BERT 2.0 encoder frozen and train a lightweight classification head. Motivated by evidence that different transformer layers encode complementary information (Pasad et al., 2023), we compute a learnable weighted sum over the 24 layerwise hidden states. The resulting sequence representation is then processed with Multi-Head Factorized Attention (MHFA), a parameter-efficient variant of multi-head attention that factorizes the attention projections so each head operates in a lower-dimensional subspace (Peng et al., 2025). Finally, the obtained result is summarized via attentive pooling before a final linear layer (Peng et al., 2025; Roncel Díaz et al., 2024). The teacher is trained with cross-entropy and used exclusively for distillation; it is not intended for real-time, on-device inference.

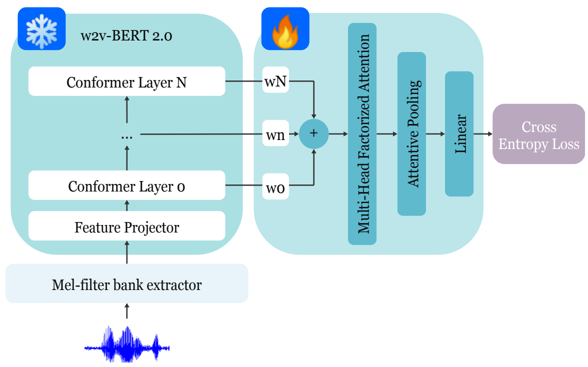


Figure 1: `w2v-BERT2-kws` architecture. Raw audio is converted to 80-channel Mel filterbanks, then passed through convolutional subsampling and a linear projection before a 24-layer Conformer encoder. Layerwise hidden states are combined via a learnable weighted sum, followed by Multi-Head Factorized Attention (MHFA), attentive pooling over time, and a linear classifier. The w2v-BERT 2.0 encoder is frozen; only the layer weights and the classification head are trained with cross-entropy.

Knowledge distillation (KD): after training the teacher, we freeze it and train the student via logit matching. The student minimizes a weighted combination of the standard cross-entropy (CE) loss with respect to ground-truth labels and a Kullback–Leibler (KL) divergence term between temperature-scaled teacher and student predictions:

$$L_{\text{KD}} = \delta\, L_{\text{CE}}(p_{\text{student}}, y_{\text{true}}) + (1-\delta)\,\tau^2\, D_{\text{KL}}(p^{\tau}_{\text{teacher}} \,\|\, \log p^{\tau}_{\text{student}}) \tag{6}$$

where L_{KD} is the total loss function, and $\delta \in [0, 1]$ is a weighting factor that controls the balance between the two loss components. τ is the temperature parameter that controls the sharpness of the probability distribution, and $L_{\text{CE}}(p_{\text{student}}, y_{\text{true}})$ is the cross-entropy loss, defined as:

$$L_{\text{CE}}(p_{\text{student}}, y_{\text{true}}) = -\sum_i y_{\text{true},i} \log p_{\text{student},i} \tag{7}$$

where the term y_{true} represents the ground truth label and p_{student} is the probability output from the student WuW model. Here i refers to each specific class (WuW or unknown). The output probability is obtained by:

$$p_{\text{student},i} = \frac{e^{z_{\text{student},i}}}{\sum_j e^{z_{\text{student},j}}} \tag{8}$$

where $z_{\text{student},i}$ are the logits, and the j index refers to the summation over both classes. To continue with, $D_{\text{KL}}(p^{\tau}_{\text{teacher}} \,\|\, p^{\tau}_{\text{student}})$ is the KL divergence between two temperature-scaled probability distributions:

$$D_{\text{KL}}(p^{\tau}_{\text{teacher}} \,\|\, p^{\tau}_{\text{student}}) = \sum_i p^{\tau}_{\text{teacher},i} \log \frac{p^{\tau}_{\text{teacher},i}}{p^{\tau}_{\text{student},i}} \tag{9}$$

where $p^{\tau}_{\text{student},i}$ is the probability output from the student WuW model, obtained by applying a temperature-scaled softmax:

$$p^{\tau}_{\text{student},i} = \frac{e^{z_{\text{student},i}/\tau}}{\sum_j e^{z_{\text{student},j}/\tau}} \tag{10}$$

Similarly, p_{teacher} is also temperature-scaled:

$$p^{\tau}_{\text{teacher},i} = \frac{e^{z_{\text{teacher},i}/\tau}}{\sum_j e^{z_{\text{teacher},j}/\tau}} \tag{11}$$

where $z_{\text{teacher},i}$ are the logits from the teacher model and τ is the temperature parameter. Higher τ produces softer targets, encouraging the student to match relative class confidences rather than only hard decisions.

3. Datasets

We utilize a proprietary in-domain corpus, OK Aura (Section 3.1), and several publicly available out-of-domain resources for augmentation and robustness. Specifically, we incorporate Spanish Common Voice v7.1 (Mozilla Foundation, 2021), the M-AILabs Spanish corpus (Solak, 2019), real and simulated room impulse responses (RIRs) and noises from OpenSLR SLR28 (OpenSLR, 2016), and environmental noise recordings from DEMAND (Joachim Thiemann and Vincent, 2013). To further improve device robustness, we additionally use microphone impulse-response collections including MicIRP[1] and the Multi-Angle Multi-Distance Microphone IR dataset (Juan Carlos Franco Hernández, 2021).

Figure 2 summarizes how these resources are used across the experimental pipeline. OK Aura is used for training, validation, and testing. In contrast, the public corpora (Common Voice, M-AILabs, SLR28, DEMAND, MicIRP, and Multi-Angle Multi-Distance Microphone IR) are used primarily for training and validation to support augmentation and robustness. We restrict bias quantification and fairness evaluation to OK Aura because the out-of-domain datasets lack the necessary demographic metadata, provide insufficient granularity, or exhibit annotation mismatches relative to the WuW task.

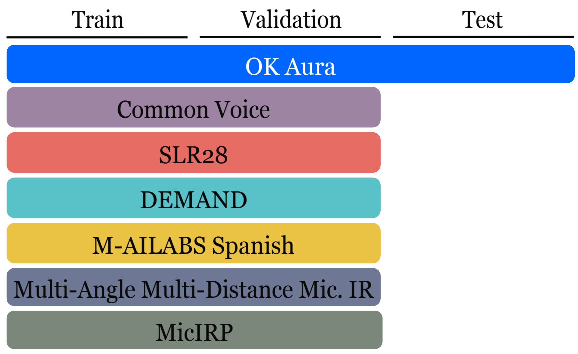


Figure 2: Dataset usage across train/validation/test splits. OK Aura is used in all phases (training, validation, and testing). Public resources are used primarily for training and validation, mainly for augmentation and robustness.

3.1. OK Aura Database

OK Aura contains approximately 5.8k audio samples (~4.5 hours) from 546 anonymized speakers. It includes both speech and non-speech material (e.g., background noise), and all speech content is in Spanish, comprising the wake-up phrase and additional utterances. The corpus provides speaker-level demographic annotations (sex, age, accent) and sample-level metadata including class labels (positive/negative), transcriptions for speech samples, and a sound-type tag ("speech" vs. "noise").

[1] https://micirp.blogspot.com/

To provide a concrete sense of the speech data, the corpus covers various realistic usage scenarios and challenging negative samples. Positive instances range from the isolated WuW (*"OK Aura"*) to the WuW embedded within context sentences (e.g., *"Perfecto, voy a mirar qué dan hoy. OK Aura"* 'Perfect, I am going to see what is on today. OK Aura'). Additionally, negative samples include utterances with partial matches, such as containing only the word "Aura" (*"Hay un aura de paz y tranquilidad."* 'There is an aura of peace and tranquility.') or "OK" (*"OK, a ver qué ponen en la tele."* 'OK, let's see what is on TV.'). Furthermore, it includes distractors with words that sound similar to the target phrase, such as *"Hola Laura,"* ('Hello Laura,'), *"Prefiero el hockey al baloncesto,"* ('I prefer hockey to basketball,'), or a combination of both: *"Porque Laura, ¿qué te pareció la película?"* ('Because Laura, what did you think of the movie?').

Furthermore, recordings span a wide range of acoustic environments as they were recorded in different spaces, from quiet rooms to natural background noise scenarios, and recorded across different devices. The dataset also includes temporal speech-event annotations (start/end times and total duration), obtained with the alignment procedure described by López and Luque (2022). A portion of OK Aura was released publicly as part of the Albayzin 2024 Wake-up Word Detection Challenge (Guillermo Cámbara and Segura, 2024) (López and Luque, 2024).

3.1.1. Demographic groups

For bias assessment, we consider three demographic attributes available in OK Aura: sex, age, and Spanish accent variety. We perform univariate analyses, evaluating each attribute independently. Sex is treated as a binary variable (Female/Male); age is grouped into 0–20, 21–30, 31–40, 41–50, and 51+; and accent labels cover the full annotation set: Unknown, Central Southern Spain, Southern Spain, Caribbean, Northern Spain, Northwestern Spain, Chilean, Eastern and Balearic Spain, Non-Native, Rio Plata, Canary Islands, Central America, Andean Pacific, Mexico, and Philippines.

3.1.2. Train and validation splits

We next analyze demographic distributions in the OK Aura training and validation splits to characterize representation imbalances. Table 1 reports the sex distribution, indicating a higher proportion of Male than Female samples. The age distribution has an average speaker age of 37 years, with most samples concentrated between 20 and 50 years old, and comparatively few samples from speakers under 20 or over 51 (Figure 3). Finally, accent labels are highly skewed, with Central Southern Spain dominating the training/validation data (Figure 4).

Sex	# Samples	Percentage
Female	2131	41.74%
Male	2974	58.26%

Table 1: Number of samples by Sex in the OK Aura Database (training and validation).

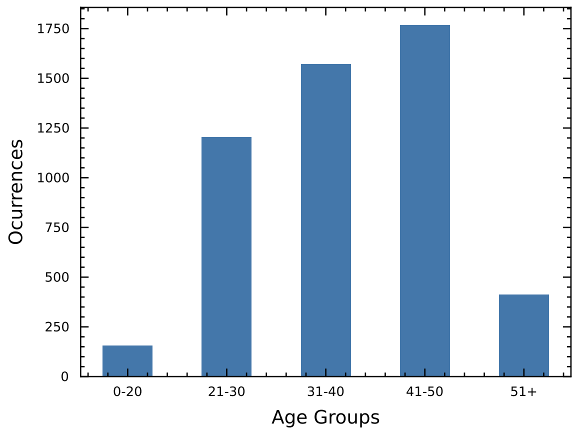


Figure 3: Age distribution in the OK Aura Database (training and validation).

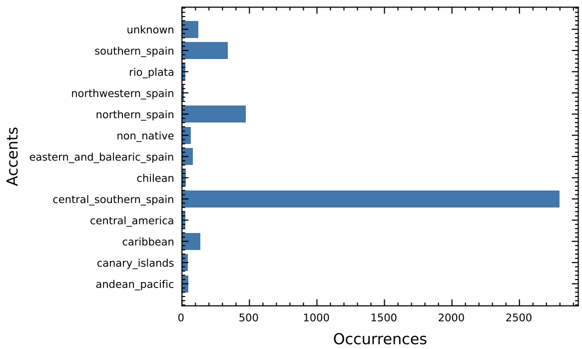


Figure 4: Accent distribution in the OK Aura Database (training and validation).

3.1.3. Test split

The OK Aura test split contains 575 samples of 47 unique speaker. Sex remains imbalanced (Table 2), and the age distribution is uneven, with a strong overrepresentation of middle-aged adults (41–50) and no samples from speakers aged 0–20 (Table 3). Accent coverage in the test set is also limited (Table 4), with several regional varieties either absent or severely underrepresented.

To ensure reliable subgroup estimates, we exclude demographic groups with fewer than 20 test samples from bias quantification and fairness reporting. Furthermore, because some of the retained groups still feature a limited number of unique speakers (e.g., only 20 female speakers), subsequent fairness metrics should be interpreted as indicative trends rather than absolute guarantees of generalizability.

4. Experimental Setup

We first describe the WuW model, which is designed for on-device inference (Section 4.1) and the training procedure (Section 4.2). In the same section, we also detail how data augmentation and knowledge distillation are integrated into training. Finally, we define the metrics used to quantify data imbalance and predictive disparities (Section 4.3).

4.1. WuW detection model

We adopt a Gated Recurrent Unit (GRU)-based classifier as a practical trade-off between real-time efficiency and WuW detection accuracy. The model consists of a single GRU layer with 200 hidden units followed by a fully connected classification layer that discriminates between *WuW* and *unknown*. We refer to this architecture as `device-sgru`; it counts 145.6k parameters, and one inference takes ~25 ms on a Pixel XL device (López et al., 2023), making it suitable for real-time inference on-device.

The input features are 13 MFCCs extracted with a 100 ms analysis window and a 50 ms hop, yielding 29 frames for a 1.5 s audio window. We replace the zeroth MFCC coefficient with the log-energy to better capture overall signal intensity (López et al., 2023).

For clarification, the `w2v-BERT2-kws` model described in Section 2.2 is just used to distill knowledge from it, it is not intended to be executed or deployed on device.

4.2. Training

All hyperparameters were set based on our previous research (López et al., 2023), prioritizing preservation of the strong baseline detection performance. Specifically, models are trained for up to 700 epochs by minimizing cross-entropy loss with a batch size of 128. We use Adam with an initial learning rate (LR) of 0.001 and reduce the LR by a factor of 10 when validation performance plateaus; training stops after four successive LR reductions without improvement.

Sex	# Samples	# Speakers
Female	254 (44.88%)	20
Male	321 (55.12%)	27

Table 2: Number of samples by Sex in the OK Aura Database (test).

Age Group	# Samples	# Speakers
0-20	0	0
21-30	135	11
31-40	138	11
41-50	295	24
51+	7	1

Table 3: Number of samples by Age Group in the OK Aura Database (test).

Accent	# Samples	# Speakers
central southern	313	26
eastern & balearic	15	1
non native	49	4
northern	90	7
southern	84	7
unknown	12	2

Table 4: Number of samples and speakers by regional Spanish accent in the OK Aura Database (test).

To improve robustness under diverse acoustic conditions, we apply additive noise and reverberation (RIR convolution) during validation. Because such augmentation introduces additional variance in the validation loss, we select the final checkpoint as the one minimizing the mean of the three lowest validation-loss values across epochs, which stabilizes model selection under stochastic validation augmentation.

This procedure is used to train both the primary `device-sgru` model and the SSL-based teacher `w2v-BERT2-kws` model used for KD. The `device-sgru` model is trained from scratch with uniformly initialized weights; we refer to this configuration as `baseline`. For `w2v-BERT2-kws`, the w2v-BERT 2.0 pre-trained encoder is kept frozen, and only the task-specific layers are trained from scratch (uniform initialization).

We then integrate two demographics-unaware training strategies for bias mitigation: data augmentation and KD. During training, augmentations are applied with probability $p = 0.2$ (i.e., 20% of training samples), aiming to preserve strong baseline characteristics while injecting robustness-inducing perturbations.

The following augmentation configuration is used:

- **FilterAugment:** number of frequency bands uniformly sampled in $[3, 9]$, minimum bandwidth 187 Hz, gain sampled in ± 6 dB.
- **FreqMixStyle:** $\alpha = 0.4$ for the Beta distribution; mixing is restricted to pairs of samples with the same label.

Attribute	Advantaged Group	Disadvantaged Group	Disparate Impact
Sex	Male	Female	0.7170
Age	41-50	21-30	0.6804
Accent	central_southern	northern	0.1692

Table 5: Disparate Impact by attribute in train and validation splits of OK Aura database.

- **Frequency masking:** $W_F = 30$ and $\nu = 128$ mel channels.

For KD, we initialize the student `device-sgru` with the weights of the pre-trained `baseline` to accelerate convergence. We then optimize the distillation objective in Eq. 6. During distillation we switch to Stochastic Gradient Descent (SGD) (momentum 0.9, weight decay 10^{-4}), as it can yield flatter minima and improved generalization. The initial LR is 0.0001 with an on-plateau scheduler, and we set $\delta = 0.2$ and $\tau = 2$.

4.3. Evaluation and metrics for bias quantification

Predictive disparities across demographic groups are often linked to data imbalance, where underrepresented cohorts tend to suffer degraded performance (Barocas and Selbst, 2016). We therefore relate demographic imbalance in the OK Aura training/validation data to disparities observed in model predictions on the test split. Concretely, we quantify imbalance in the input data (Section 4.3.1) and quantify predictive differences (Section 4.3.2).

4.3.1. Bias in data

To quantify demographic imbalance in the dataset, we use Disparate Impact (DI), a commonly used ratio-based metric in algorithmic fairness. Let G denote the set of demographic groups, with $a, d \in G$ representing an advantaged and disadvantaged group, and let $Y \in \{0, 1\}$ denote the binary label (1 for WuW presence and 0 otherwise). DI is defined as:

$$DI = \frac{P(Y = 1 \mid G = d)}{P(Y = 1 \mid G = a)}. \tag{12}$$

For multi-valued attributes (e.g., accent, age groups), we report the maximum ratio (or equivalently the most imbalanced pair) across all group pairs.

4.3.2. Bias in predictions

We follow the pairwise group comparison protocol described by Singh et al. (2023) to assess predictive disparities. We evaluate WuW detection on fixed 1.5 s windows and use a fixed decision threshold of 0.5. We report performance using the F1-score:

$$\mathrm{F1} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}, \tag{13}$$

where

$$\mathrm{Precision} = \frac{TP}{TP + FP} \tag{14}$$

$$\mathrm{Recall} = \frac{TP}{TP + FN} \tag{15}$$

Predictive Disparity (PD). We define PD as the maximum absolute difference in F1-score across demographic groups:

$$PD = \max_{i,j \in G} |\mathrm{F1}(g_i) - \mathrm{F1}(g_j)|, \tag{16}$$

where $g_i, g_j \in G$ denote group identities. Larger values indicate stronger performance gaps and potential fairness concerns.

Relative reduction of predictive disparity (RRPD). To compare mitigation strategies, we report the relative reduction in PD with respect to the baseline:

$$RRPD = 100 \times \frac{PD_{\mathrm{baseline}} - PD_{\mathrm{technique}}}{PD_{\mathrm{baseline}}}. \tag{17}$$

Here, PD_{baseline} denotes disparity for the baseline model and $PD_{\mathrm{technique}}$ for the model trained with a given mitigation technique.

5. Results and Discussion

We report (i) demographic imbalance in the OK Aura training/validation splits (data bias; Section 5.1.1), (ii) predictive disparities of the `baseline` WuW detector on the OK Aura test split (prediction bias; Section 5.1.2), and (iii) the impact of demographics-agnostic training strategies for mitigation (Section 5.2). Following our evaluation protocol, demographic groups with fewer than 20 test samples are excluded from bias quantification to ensure stable subgroup estimates.

5.1. Bias quantification

5.1.1. Bias in data

Table 5 reports DI for the OK Aura training/validation splits, revealing systematic representation imbalances across all examined attributes. Male speakers are overrepresented relative to female speakers ($DI = 0.717$), the 41–50 age group dominates compared to 21–30 ($DI = 0.6804$), and accent imbalance is most severe: Central Southern Spain is disproportionately represented relative

Group	F1-score	Support
Male	**0.9863**	296
Female	0.9825	204
21–30	**0.9956**	115
31–40	0.9828	118
41–50	0.9827	265
southern_spain	0.9818	84
central_southern_spain	**0.9873**	278
northern_spain	0.9781	70
non_native	0.9870	39
PD (sex)	0.0038	
PD (age)	0.0129	
PD (accent)	0.0092	

Table 6: Performance across demographic groups with predictive disparity (PD) for baseline model.

to Northern Spain ($DI = 0.1692$). These skewed distributions are likely to affect generalization and may translate into unequal predictive performance across groups.

5.1.2. Bias in predictions

Table 6 presents subgroup F1-scores and PD for the `baseline` model. Sex-related disparity is small but measurable ($PD = 0.0038$), with slightly higher F1 for male speakers (0.9863 vs. 0.9825). Age exhibits the largest performance gap ($PD = 0.0129$): the 21–30 group performs best (0.9956), while the 41–50 group performs worst (0.9827), highlighting older adults as a key cohort for mitigation. Accent disparities are also evident ($PD = 0.0092$), where Central Southern Spain achieves the highest F1 (0.9873) and Northern Spain is lower (0.9781), suggesting that accent variability remains a relevant source of error.

Overall, the baseline results indicate that demographic imbalance in the training data co-occurs with predictive disparities, motivating mitigation methods that increase robustness without requiring demographic labels.

5.2. Bias mitigation and analysis

First, the high-capacity SSL-based classifier `w2v-BERT2-kws` exhibits substantially lower disparities than `baseline` (Table 7), indicating that large-scale SSL pretraining can reduce, but not eliminate, demographic performance gaps. This motivates its use as a teacher model for KD.

Table 8 reports the relative reduction of predictive disparity achieved by each technique with respect to `baseline`. We observe that augmentation and KD show attribute-dependent behavior. Specifically, DIR only improves disparity for sex (67.35% RRPD), suggesting that device-specific impulse responses may not capture the heterogeneous acoustic variations typically associated with demographic attributes. FilterAugment yields the largest reduction in sex disparity (88.26% RRPD) and also improves age fairness (30.14% RRPD), but it increases accent disparity. This suggests that while smooth frequency-energy perturbations help reduce reliance on certain demographic-specific spectral cues, they do not universally benefit all attributes. On the other hand, FreqMixStyle improves age and accent fairness but degrades sex fairness (negative RRPD), indicating that frequency-wise statistics mixing affects demographic attributes differently and may not generalize across all cues simultaneously. We hypothesize that FreqMixStyle and FilterAugment, underperform on some speaker demographics as they can reshape frequency statistics too aggressively. They may destroy critical formant/prosodic cues and yielding mixed fairness gains at higher error cost.

In contrast, Frequency Masking provides the most consistent gains across all attributes, achieving a strong reduction in age disparity (83.65%) while also narrowing the gaps for sex and accent. Furthermore, Table 9 shows that these gains are achieved while maintaining competitive subgroup F1-scores, making Frequency Masking a suitable fairness-oriented augmentation. Suppressing specific frequency bands appears to force the model to distribute evidence across multiple regions of the spectrum rather than overfitting to a single demographic-correlated band (e.g., the F0 or low-formant region). This distributed attention both improves overall robustness and reduces reliance on demographic-specific cues.

Finally, KD reduces sex and age disparity but does not consistently reduce accent disparity. One plausible explanation is that accent invariance is constrained by limited accent diversity in the labeled in-domain data used during distillation, which may limit the teacher's ability to provide accent-neutral soft targets. Finally, combining KD with Frequency Masking does not improve over the best single-technique settings and can degrade results for some attributes, suggesting an interaction between stochastic spectral corruption and logit matching that may be difficult for a small student architecture to optimize jointly.

6. Conclusion and Future Work

This work shows that demographic-agnostic training can mitigate bias in Wake-up Word detection without requiring demographic labels during training. We studied two complementary families of methods: (i) data augmentation that perturbs or removes frequency information, and (ii) knowledge

Classifier	Sex RRPD (%)	Age RRPD (%)	Accent RRPD (%)
w2v-BERT2-kws	79.64	**85.35**	41.05

Table 7: w2v-BERT2-kws Relative Reduction of Predictive Disparity for demographic attributes in comparison to baseline model. It demonstrates reduced PD across sex, age, and accent categories.

Experimental Setting	Sex RRPD (%)	Age RRPD (%)	Accent RRPD (%)
DIR	67.35	0.00	-20.13
FreqMixStyle	-21.42	34.12	**40.48**
FilterAugment	**88.26**	30.14	-40.19
FreqMasking	39.94	**83.65**	**40.48**
KD	67.35	15.10	-20.13
KD + FreqMasking	21.24	15.10	-40.19

Table 8: Relative Reduction of Predictive Disparity (RRPD) across sex, age, and accent for different training techniques. Higher is better (negative values indicate increased disparity).

distillation (KD) from a large pre-trained speech model. Across speaker groups defined by sex, age, and accent, these approaches reduce performance disparities while maintaining competitive overall accuracy.

Our results highlight that augmentation design is critical for effective mitigation. In particular, frequency-energy perturbations and statistics mixing exhibited attribute-dependent behavior, sometimes degrading fairness for specific groups. Contrary, Frequency Masking emerged as the most robust single technique. It consistently achieved a relative reduction in predictive disparity (RRPD) of 39.94% (sex), 83.65% (age), and 40.48% (accent). By suppressing specific frequency bands, Frequency Masking prevents the model from overfitting to demographic-correlated acoustic cues (e.g., fundamental frequency) and forces it to distribute evidence across the broader spectrum. Additionally, KD achieved high RRPD, especially for sex and age, but showed limited impact on accent. This suggests that accent-invariant transfer remains constrained by the limited accent diversity available in the in-domain distillation data.

Group	F1-score	Support
Male	0.9828	296
Female	**0.9851**	204
21–30	**0.9880**	115
31–40	0.9828	118
41–50	0.9847	265
southern_spain	0.9818	84
central_southern_spain	0.9835	278
northern_spain	0.9781	70
non_native	**0.9870**	39
PD (sex)	0.0023	
PD (age)	0.0052	
PD (accent)	0.0089	

Table 9: Predictive Disparity using FreqMasking technique to train device-sgru model.

Future work will extend this analysis to intersectional fairness settings (e.g., older females with specific regional accents) and broaden demographic coverage by curating more balanced data across attributes. Particular emphasis will be placed on underrepresented age and accent groups.

Limitations. (i) Our analysis is univariate and does not capture intersectional effects (e.g., older female speakers with regional accents). (ii) The training/validation procedure includes out-of-domain audio sources, which can introduce distribution mismatch and metadata inconsistencies. (iii) Several demographic groups are underrepresented in the test split and were excluded from fairness reporting; furthermore, the limited number of speakers within the retained groups implies that our specific conclusions regarding them should be interpreted with caution. (iv) F1 does not decompose disparities into false accepts and false rejects, which carry asymmetric costs in WuW detection. (v) Conclusions at a fixed threshold of 0.5 may not generalize across operating points. Future work should adopt multi-objective evaluation frameworks that jointly optimize overall accuracy, per-group F1, and cross-attribute fairness, rather than treating each in isolation.

Ethics Statement This research addresses fairness in speech systems, which has positive ethical implications for reducing discrimination. The OK Aura dataset involves anonymized speakers with informed consent. Our demographics-agnostic approach specifically avoids requiring sensitive demographic labels during deployment, protecting user privacy. However, we acknowledge that bias mitigation techniques may have unintended effects on other demographic groups not examined in this study, and that univariate analysis may miss intersectional discrimination patterns.

7. Acknowledgments

This project has been partially funded by the Spanish Project 6G-RIEMANN (Grant Agreement No. 2022/0005420) and by the European Union's Horizon 2020 RIA ELOQUENCE project (Grant Agreement No. 101135916). Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or European Commission-EU. Neither the European Union nor the granting authority can be held responsible for them.

8. Bibliographical References

Giuseppe Attanasio, Beatrice Savoldi, Dennis Fucci, and Dirk Hovy. 2024. Twists, humps, and pebbles: Multilingual speech recognition models exhibit gender performance gaps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21318–21340, Miami, Florida, USA. Association for Computational Linguistics.

Andrew Bailey and Mark D Plumbley. 2021. Gender bias in depression detection using audio features. In *IEEE 29th European Signal Processing Conference (EUSIPCO)*, pages 596–600.

Solon Barocas and Andrew D. Selbst. 2016. Big data's disparate impact. *California Law Review*, 104(3):671–732.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, et al. 2023. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*.

Junyi Chai, Zhihao Wang, Jiaxin Chen, Hao He, Dawn Song, and Xia Li. 2022. Fairness without demographics through knowledge distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35.

Zhengyang Chen, Sanyuan Chen, Yu Wu, Yao Qian, Chengyi Wang, Shujie Liu, Yanmin Qian, and Michael Zeng. 2022. Large-scale self-supervised speech representation learning for automatic speaker verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6147–6151.

Anna Seo Gyeong Choi and Hoon Choi. 2025. Fairness of automatic speech recognition: Looking through a philosophical lens. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(1):605–614.

Pranav Dheram, Murugesan Ramakrishnan, Anirudh Raju, I-Fan Chen, Brian King, Katherine Powell, Melissa Saboowala, Karan Shetty, and Andreas Stolcke. 2022. Toward fairness in speech recognition: Discovery and mitigation of performance disparities. *INTERSPEECH*.

Hend ElGhazaly, Bahman Mirheidari, Heidi Christensen, and Nafise Sadat Moosavi. 2025. Fairness in automatic speech recognition isn't a one-size-fits-all. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 19169–19178, Suzhou, China.

Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. 2021. Quantifying bias in automatic speech recognition. *arXiv preprint arXiv:2103.15122*.

Marcio Fuckner, Sophie Horsman, Pascal Wiggers, and Iskaj Janssen. 2023. Uncovering bias in asr systems: Evaluating wav2vec2 and whisper for dutch speakers. In *IEEE International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 146–151.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Jiangyu Han, Federico Landini, Johan Rohdin, Anna Silnova, Mireia Diez, and Lukáš Burget. 2025. Leveraging self-supervised learning for speaker diarization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

James D Harnsberger, Rahul Shrivastav, William S Brown Jr, Howard Rothman, and Harry Hollien. 2008. Speaking rate and fundamental frequency as speech cues to perceived age. *Journal of voice*, 22(1):58–69.

Camille Harris, Chijioke Mgbahurike, Neha Kumar, and Diyi Yang. 2024. Modeling gender and dialect bias in automatic speech recognition. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 15166–15184.

Wiebke Hutiri, Aaron Yi Ding, Fahim Kawsar, and Akhil Mathur. 2023. Tiny, always-on, and fragile: Bias propagation through design choices in on-device machine learning workflows. *ACM Transactions on Software Engineering and Methodology*, 32(6):1–37.

Veton Këpuska and Jason Breitfeller. 2006. Wake-up-word speech recognition application for first responder communication enhancement. In *Sensors, and Command, Control, Communications,* *and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense V*, volume 6201, page 62011E. International Society for Optics and Photonics, SPIE.

Byeonggeun Kim, Seunghan Yang, Jangho Kim, Hyunsin Park, Juntae Lee, and Simyung Chang. 2022. Domain generalization with relaxed instance frequency-wise normalization for multi-device acoustic scene classification. *arXiv preprint arXiv:2206.12513*.

Gwantae Kim, David K Han, and Hanseok Ko. 2021. Specmix: A mixed sample data augmentation method for training withtime-frequency domain features. *arXiv preprint arXiv:2108.03020*.

Alkis Koudounas, Flavio Giobergia, Eliana Pastor, and Elena Baralis. 2024. A contrastive learning approach to mitigate bias in speech models. *arXiv preprint arXiv:2406.14686*.

Veton Këpuska. 2011. Wake-up-word speech recognition. In Ivo Ipsic, editor, *Speech Technologies*, chapter 12. IntechOpen, London.

Beltrán Labrador, Pai Zhu, Guanlong Zhao, Angelo Scorza Scarpati, Quan Wang, Alicia Lozano-Diez, and Ignacio Lopez-Moreno. 2025. Personalizing keyword spotting with speaker information. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Yi-Cheng Lin, Tzu-Quan Lin, Hsi-Che Lin, Andy T Liu, and Hung-yi Lee. 2024. On the social bias of speech self-supervised models. In *Proceedings of INTERSPEECH*, pages 4638–4642.

Fernando López and Jordi Luque. 2024. Albayzin evaluation 2024: Wake-up word detection challenge.

Iván López-Espejo, Zheng-Hua Tan, John HL Hansen, and Jesper Jensen. 2021. Deep spoken keyword spotting: An overview. *IEEE Access*, 10:4169–4199.

Fernando López and Jordi Luque. 2022. Iterative pseudo-forced alignment by acoustic CTC loss for self-supervised ASR domain adaptation. In *Proceedings of IberSPEECH*, pages 46–50.

Fernando López, Jordi Luque, Carlos Segura, and Pablo Gómez. 2023. Robust wake-up word detection by two-stage multi-resolution ensembles. *arXiv preprint arXiv:2310.11379*.

Joshua L Martin and Kelly Elizabeth Wright. 2023. Bias in automatic speech recognition: The case of african american language. *Applied Linguistics*, 44(4):613–630.

Yen Meng, Yi-Hui Chou, Andy T. Liu, and Hung-yi Lee. 2022. Don't speak too fast: The impact of data bias on self-supervised speech models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3258–3262.

Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D. Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, Tara N. Sainath, and Shinji Watanabe. 2022. Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1179–1210.

Jacob Mørk, Holger Severin Bovbjerg, Gergely Kiss, and Zheng-Hua Tan. 2024. Noise-robust keyword spotting through self-supervised pretraining. *arXiv preprint arXiv:2403.18560*.

Tobias Morocutti, Florian Schmid, Khaled Koutini, and Gerhard Widmer. 2023. Device-robust acoustic scene classification via impulse response augmentation. In *IEEE 31st European Signal Processing Conference (EUSIPCO)*, pages 176–180.

Dena Mujtaba, Nihar Mahapatra, Megan Arney, J Yaruss, Hope Gerlach-Houck, Caryn Herring, and Jia Bin. 2024. Lost in transcription: Identifying and quantifying the accuracy biases of automatic speech recognition systems against disfluent speech. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4795–4809, Mexico City, Mexico.

Hyeonuk Nam, Seong-Hu Kim, and Yong-Hwa Park. 2022. Filteraugment: An acoustic environmental data augmentation method. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Ankita Pasad, Bowen Shi, and Karen Livescu. 2023. Comparative layer-wise analysis of self-supervised speech models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Eliana Pastor, Luca De Alfaro, and Elena Baralis. 2021. Looking for trouble: Analyzing classifier behavior via pattern divergence. In *Proceedings of the International Conference on Management of Data*, pages 1400–1412.

Junyi Peng, Ladislav Mošner, Lin Zhang, Oldřich Plchot, Themos Stafylakis, Lukáš Burget, and Jan Černockỳ. 2025. Ca-mhfa: A context-aware multi-head factorized attentive pooling for ssl-based speaker verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Marina Piat, Dominique Fohr, and Irina Illina. 2008. Foreign accent identification based on prosodic parameters. In *INTERSPEECH*, pages 759–762.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning (ICML)*, volume 202, pages 28492–28518. PMLR.

Daniel Roncel Díaz, Federico Costa, and Javier Hernando. 2024. On the use of audio to improve dialogue policies. In *IberSPEECH*, pages 151–155.

Harvineet Singh, Fan Xia, Mi-Ok Kim, Romain Pirracchio, Rumi Chunara, and Jean Feng. 2023. A brief tutorial on sample size calculations for fairness audits. *arXiv preprint arXiv:2312.04745*.

Isaac Slaughter, Craig Greenberg, Reva Schwartz, and Aylin Caliskan. 2023. Pre-trained speech processing models contain human-like biases that propagate to speech emotion recognition. pages 8967–8989.

Loes Vandenberghe et al. 2023. Exploring data augmentation in bias mitigation against non-native-accented speech. *arXiv preprint arXiv:2312.15499*.

Houri K Vorperian, Raymond D Kent, Yen Lee, and Daniel M Bolt. 2019. Corner vowels in males and females ages 4 to 20 years: Fundamental and f1–f4 formant frequencies. *The Journal of the Acoustical Society of America*, 146(5):3255–3274.

Xin Wang, Héctor Delgado, Hemlata Tak, Jee-weon Jung, Hye-jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi Kinnunen, et al. 2024. Asvspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale. *arXiv preprint arXiv:2408.08739*.

Mingdong Yu, Xiaofeng Jin, Bangxian Wan, and Guirong Wang. 2023. A few-shot speech keyword spotting method based on self-supervise learning. In *16th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–5.

Maryam Zolnoori, Sasha Vergez, Zidu Xu, Elyas Esmaeili, Ali Zolnour, Krystal Anne Briggs, Jihye Kim Scroggins, Seyed Farid Hosseini Ebrahimabad, James M Noble, Maxim Topaz, et al. 2024. Decoding disparities: evaluating automatic speech recognition system performance in transcribing black and white patient verbal communication with nurses in home healthcare. *JAMIA open*, 7(4):ooae130.

9. Language Resource References

Guillermo Cámbara, Jordi Luque, David Bonet, Fernando López, Mireia Farrús, Pablo Gómez and Carlos Segura. 2024. *Okey Aura Wake-up Word Dataset*. Zenodo, 1.1.0.

Maliha Jahan, Yinglun Sun, Priyam Mazumdar, Zsuzsanna Fagyal, Thomas Thebaud, Jesus Villalba, Mark Hasegawa-Johnson, Najim Dehak, and Laureano Moro Velazquez. 2025. Faist: A benchmark dataset for fairness in speech technology. In *Proceedings of Interspeech*, pages 1343–1347.

Joachim Thiemann, Nobutaka Ito and Emmanuel Vincent. 2013. *DEMAND: Diverse Environments Multi-Channel Acoustic Noise Database*. The Journal of the Acoustical Society of America.

Juan Carlos Franco Hernández, Tim Brookes, Enzo De Sena. 2021. *Multi-Angle, Multi-Distance Microphone Impulse Response Dataset*. Zenodo, 1.0.0.

Mozilla Foundation. 2021. *Common Voice Corpus (Spanish), version 7.1*. Mozilla Common Voice.

OpenSLR. 2016. *Room Impulse Response and Noise Database (SLR28)*. OpenSLR.

Imdat Solak. 2019. *The M-AILABS Speech Dataset*. M-AILABS.

Irina-Elena Veliche, Zhuangqun Huang, Vineeth Ayyat Kochaniyan, Fuchun Peng, Ozlem Kalinli, and Michael L Seltzer. 2024. Towards measuring fairness in speech recognition: Fair-speech dataset. *arXiv preprint arXiv:2408.12734*.

Scalable Expansion of Multilingual Speech LLMs for ASR: A Continual Learning Approach

Lorenzo Concina, Marco Matassoni, Alessio Brutti
Center for Augmented Intelligence, Fondazione Bruno Kessler, Trento, Italy
{lconcina, matasso, brutti}@fbk.eu

Abstract

Speech Large Language Models have recently enabled the processing of spoken language by coupling powerful language models (LLMs) with pre-trained speech encoders. However, their multilingual scalability remains limited, particularly for low-resource and unseen languages, while naïve fine-tuning often triggers catastrophic forgetting of previously learned languages. This work investigates how Continual Learning (CL) can be used to sustainably expand multilingual Speech LLMs. We first demonstrate that multilingual projectors can be efficiently bootstrapped to new languages, even with extremely small datasets, but at the cost of severe degradation on the original supported languages. To address this, we adopt rehearsal-based CL strategies and show that interleaving even small amounts of replay data effectively stabilizes multilingual performance. Through extensive ablations, we quantify the minimum rehearsal budget required to prevent forgetting and identify fragile languages that require more targeted reinforcement. We further evaluate sequential acquisition of four linguistically diverse languages (Ukrainian, Japanese, Thai, and Vietnamese), revealing the trade-offs between buffer size and long-term stability. Finally, based on these empirical observations, we propose a Fragility-Based Sampling heuristic as a pathway to allocate rehearsal data more efficiently by tiering languages according to their stability thresholds. Our findings provide a practical roadmap for scalable, resource-efficient multilingual expansion of Speech LLMs, enabling inclusive ASR systems that can grow over time without sacrificing prior knowledge.

Keywords: continual learning, data replay, speech recognition, LLM, low-resource languages, multilinguality, fragility

1. Introduction

Large Language Models (LLMs) have transformed natural language processing, and their integration with audio modalities has birthed the era of Speech Large Language Models (Speech LLMs). By bridging pre-trained speech encoders with high-capacity language decoders, these models move beyond simple transcription to achieve a more context-aware understanding of spoken communication. Traditionally, speech understanding has relied on cascaded architectures that sequentially link ASR, language modeling, and TTS modules. However, this pipeline approach suffers from several flaws (1). First, it is prone to error propagation, where inaccuracies in transcription directly degrade the quality of the subsequent text generation. Second, these systems lack a shared context; once speech is converted to text, paralinguistic features, such as pitch and tone are lost. Consequently, the language model cannot take advantage of the rich information contained in the original audio signal. Furthermore, the multi-stage process introduces significant processing latency. Speech Language Models (SLMs) (2) seek to resolve these problems by processing audio directly, utilizing the inherent capabilities of LLMs to match or exceed the performance of specialized, task-specific models in a unified framework. Despite recent advances, there is still a gap in the multilingual scalability of these systems. Most state-of-the-art models are optimized primarily for high-resource languages such as English, often struggling with the phonetic and structural nuances of low-resource or underrepresented languages (3). In the context of Automatic Speech Recognition (ASR), extending Speech LLMs to a global scenario is not just a matter of data volume; it requires overcoming challenges such as catastrophic forgetting (4), where the acquisition of a new language degrades performance on previously learned ones.

Building on these challenges, the scope of this work focuses on the sustainable expansion of Speech LLM architectures through Continual Learning (CL). Rather than resorting to the computationally expensive and data-intensive process of retraining models from scratch to accommodate new linguistic domains, we investigate methods to incrementally integrate additional languages into an existing system. Using as a starting point the MEUSLI [1] projector (a multilingual interface connecting a frozen Whisper encoder to a 1.7B EuroLLM based on SLAM-ASR framework (5)), we examine how to efficiently bootstrap support for unseen and low-resource languages. We specifically explore rehearsal-based (6) (Data Replay) strategies to mitigate the catastrophic forgetting inherent in the SLAM-ASR framework. By analyzing the minimum rehearsal data required and proposing

[1] https://huggingface.co/SpeechTek/MEUSLI_projector_v2

Proceedings of SPEAKABLE @ LREC 2026, pages 59–65
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

a Fragility-Based Sampling heuristic, we provide a pathway for ASR Speech-LLM systems that is both inclusive of underrepresented languages and resource-efficient.

2. The Speech-LLM Paradigm for ASR

The emergence of Speech Large Language Models (SLMs) marks a shift toward unified, multimodal architectures that bridge the gap between raw acoustic signals and high-level linguistic reasoning. Unlike traditional cascaded systems, which reduce speech to plain text before processing, SLMs aim to maintain a continuous flow of information, preserving paralinguistic cues such as tone, pitch, and rhythm. These models typically leverage the vast pre-trained knowledge of Large Language Models (LLMs) to perform tasks like Automatic Speech Recognition (ASR) or Spoken Question Answering directly from audio input.

2.1. The SLAM-ASR Architecture

This study leverages the SLAM-ASR framework[2], which provides an efficient, modular approach to multimodal integration. As shown in Figure 1, the architecture is composed of three main components:

- **Speech Encoder**: Transforms raw audio into high-dimensional acoustic representations.
- **Projector**: A lightweight module that maps acoustic representations into the same embedding space as the LLM's token embeddings. This interface can range from simple linear layers to more complex neural architectures. In this work we leverage the MEUSLI linear projector.
- **Language Model**: A pre-trained LLM that processes the projected speech embeddings as if they were text tokens.

A significant advantage of this paradigm is the efficiency, during training in fact, both the encoder and the LLM are typically kept frozen, with the possibility of eventually use Low-Rank Adaptation (LoRA) (7) to further enhance LLM performance.

2.2. Baseline Model: The MEUSLI Projector

To evaluate the ability to extend a Speech-LLM based model to new languages by applying Continual Learning techniques, we utilize the MEUSLI (Multilingual EU Speech LInear projector) model as our experimental baseline. MEUSLI is an open-science initiative designed to connect a frozen **Whisper-large-v3-turbo** (8) encoder with the multilingual **EuroLLM 1.7B-Instruct** (9) backend.

The model was initially trained on 7,622 hours of open-source data from Common Voice, FLEURS, and VoxPopuli, covering 28 European languages. The projector itself is a linear layer with approximately 17.31M trainable parameters, supplemented by a LoRA configuration in the LLM adding 1.38M tunable parameters.

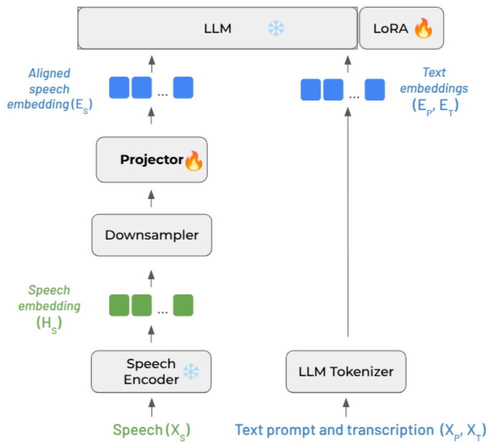


Figure 1: The proposed Speech LLM architecture used in MEUSLI (Multilingual EU Speech LInear projector) training pipeline (5; 3).

2.3. Multilingual Capabilities and Constraints

The current iteration of the projector natively supports 28 European languages, ranging from high-resource (Spanish, French, German) to medium and low-resource ones (Breton, Maltese, Welsh). To enhance performance, Low-Rank Adaptation (LoRA) is applied to the LLM, introducing 1.38M tunable parameters. Despite its broad coverage, the model faces some limitations:

- Language Gaps [1]: Performance remains uneven, with very low-resource languages (e.g., Irish or Breton) exhibiting significantly higher Word Error Rates (WER) compared to their high-resource counterparts.
- The model is only focused on European languages, but it does not cover them all.
- The model has been trained on open source data: Common Voice 17 (10), Fleurs (11) and VoxPopuli (12), therefore it is not suitable for other specific domains or accents.

3. Bootstrapping New Languages

To explore the model's capacity for expansion, we investigate the process of bootstrapping—fine-

[2]https://github.com/X-LANCE/SLAM-LLM

tuning the pre-trained MEUSLI projector on languages entirely absent from its initial training set. This stage is critical for assessing whether a multilingual foundation model can effectively transfer knowledge to low-resource scenarios with minimal train data.

We conducted two primary bootstrapping experiments on languages that utilize different scripts or linguistic structures than those natively supported. For these two experiments, we used the training recipe of SLAM-ASR while loading the MEUSLI projector as starting checkpoint:

- Ukrainian: Fine-tuning on 30 hours of Common Voice data enabled the model to achieve a 16.3% Word Error Rate (WER), a notable improvement over the monolingual training of the same exact pipeline that only reaches 20.4% WER.

- Albanian: Even with an extremely limited dataset of only 46 minutes of Common Voice data, the model successfully bootstrapped to a 75.6% WER. In contrast, a monolingual projector trained from scratch on this same data failed to converge, reaching 389% WER.

These two experiments reported in Table 1 show promising results on how to extend this model to new languages. These results confirm that multilingual pre-training provides a superior initialization point for new language acquisition.

Table 1: WER for the bootstrapping experiment on Ukrainian and Albanian. The training data comes from Common Voice 17. "Mono" refers to a monolingual projector trained on the CV data, "→" indicates MEUSLI fine-tuned on CV data.

Language	Training	Mono	MEUSLI→
Ukrainian	30h	20.4%	**16.3%**
Albanian	46min	389%	**75.6%**

3.1. The Emergence of Catastrophic Forgetting

While the model successfully acquires the new target language, this specialized fine-tuning induces a severe side effect known as catastrophic forgetting. This phenomenon occurs when the weights of the projector are adjusted to the distribution of the new task or in this case language, causing a sudden and near-total loss of performance on the original 28 supported European languages. Table 2 shows the side effects on some of the 28 supported languages after fine-tuning on Ukrainian. Specifically, for almost all these test languages in both test sets (Common Voice 17 and Fleurs), the model is no longer able to transcribe properly. The output transcription is instead full of hallucinations with Ukrainian characters.

Table 2: Comparison of WER showing the catastrophic forgetting effect after fine-tuning on Ukrainian. High WER values in the Fleurs and CV17 columns indicate a collapse of original multilingual capabilities compared to the base model performances reported on the third and fourth columns. "→" indicates MEUSLI fine-tuned on CV data

Language	MEUSLI→ FL	MEUSLI→ CV	MEUSLI CV	MEUSLI FL
Spanish	90.34	–	5.22	4.09
German	15.50	19.87	7.11	7.79
French	68.90	–	11.24	7.83
Portuguese	109.6	123.66	9.39	4.86
English	9.02	22.72	12.94	6.34
Polish	100.6	–	14.19	8.68
Czech	108	–	11.16	11.32
Italian	39.28	60.18	6.01	3.32
Danish	40.02	95.16	18.81	14.65
Latvian	68.38	–	27.12	17.23

4. Continual Learning for Multilingual ASR

To learn how to transcribe a new language and at the same time mitigate the catastrophic forgetting identified in Section 3.1, we adopt a **Continual Learning (CL)** framework based on **Data Rehearsal** (or Data Replay) and inspired by the framework proposed in CL-MASR (13). The core objective is to integrate new languages sequentially, without degrading the performance of the original 28 European languages already supported by the MEUSLI model.

Within this Continual Learning framework, we formulate the expansion of the Speech LLM as a **Task Incremental Learning (TIL)** problem. In this paradigm, the model is trained on a sequence of distinct tasks $T_1, T_2, \ldots, T_N$, where each task T_i represents the acquisition of a new, previously unsupported language (e.g., $T_1 =$ Ukrainian, $T_2 =$ Japanese). The objective is to minimize the error on the new task T_i while strictly bounding the performance degradation on all previously learned tasks $T_{<i}$, including the original 28 base languages.

To formally evaluate the global stability of the model across this sequence, we adopt the **Average Word Error Rate (AWER)** metric as in (13). Let $WER_{i,j}$ be the Word Error Rate evaluated on task j after the model has finished training on task i. The AWER after training on task i is defined as the arithmetic mean of the errors across all active tasks up to that point:

$$AWER_i = \frac{1}{i} \sum_{j=1}^{i} WER_{i,j} \tag{1}$$

This metric allows us to monitor the evolution of performances.

4.1. Establishing a Continual Learning Baseline: The Case of Ukrainian

We first validate the effectiveness of data rehearsal by repeating the Ukrainian experiment as the initial step in our TIL sequence ($T_1 =$ Ukrainian). In this setup, we fine-tune the projector on the new task T_1 using the Common Voice (CV) training set, while simultaneously replaying a small subset of samples from the original 28 base languages. Initial results using a rehearsal budget of 1,000 samples per language—selected deterministically as the first 1,000 utterances from each training set—demonstrated that it is possible to acquire the new task (reaching 17.56% WER on T_1) while keeping performance on the base tasks T_0 stable. This result shown in Table 3 confirms that interleaving even a minimal amount of data from previous tasks acts as a powerful regularizer, anchoring the projector's weights and preventing the collapse of the shared multilingual embedding space.

Table 3: Comparison of Baseline (Meusli) vs. Sequential Sampling (First 1000) for Ukrainian acquisition. Metrics reported in WER (%).

Exp	ES	DE	FR	IT	EN	DA	PL	CS	UK	Avg
Meusli	4.09	7.79	7.83	3.32	6.34	14.65	8.68	11.32	106.9	**18.99**
first 1k	4.82	9.54	7.96	5.94	5.20	18.67	10.95	15.21	17.56	**10.65**

4.2. Ablation Study: Rehearsal Buffer Sensitivity

To optimize computational efficiency and storage, we investigated the *minimum* rehearsal budget required to maintain stability. We tested buffer sizes ranging from 1000 down to a single sample per language. As shown in Figure 2 , the model demonstrates remarkable resilience:

- **High Stability**: As small as 5 samples per language are sufficient to preserve the performance of the original base tasks (T_0) after the acquisition of the first new task (T_1).
- **Random vs. Sequential Sampling**: We found that **Random Sampling** outperforms the "First N" sequential strategy used in the first Ukrainian CL experiment showed in section 4.1 when using a rehearsal buffer of 1000 or 500 samples per language. Randomization ensures a broader coverage of the acoustic and linguistic distribution of each language, which is vital for preventing drift in the projector and prevent overfitting on small amount of training data.
- **The Breakdown Point**: At only 1 sample per language, we observe the re-emergence of catastrophic forgetting in specific "fragile" languages such as Polish, Czech, Hungarian, and Estonian. In these cases, the model begins to hallucinate Ukrainian tokens when presented with base language audio, indicating that a single data point is insufficient for reinforcement.

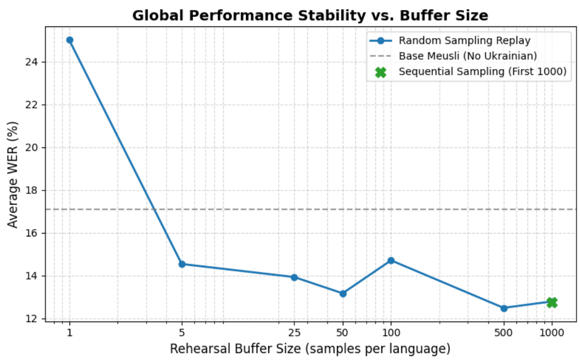


Figure 2: Ablation study of rehearsal buffer size on the global Average WER. The plot illustrates the stability plateau down to 5 samples per language and the collapse when the buffer is reduced to a single sample.

4.3. Sequential Language Acquisition

Building on the insights from the Ukrainian baseline, we extended the CL process to a multi-stage sequential integration of four new languages: Ukrainian (UK) $\rightarrow$ Japanese (JA) $\rightarrow$ Thai (TH) $\rightarrow$ Vietnamese (VI). This sequence was designed to test the model's ability to incorporate diverse scripts and distinct phonologies in a truly incremental fashion and with a challenging test using low resource languages very different from the ones natively supported by the base model. For this set of experiments, we employed the Ukrainian checkpoint obtained in section 4.1 with Common Voice data and then, for Japanese, Thai, and Vietnamese we opted to use the training set of the INTERSPEECH 2025 MLC-SLM challenge (14) to further diversify the train data domain. Figure 3 shows this pipeline. We evaluated two versions of this sequential loop to further probe the limits of memory retention introduced in section 4.2: one using a conservative buffer of 500 samples per language and a more aggressive setup with only 25 samples. A comparative analysis of these two experiments, shown in figure 4, reveals a significant trade-off between buffer size and long-term stability:

- **High-Fidelity Retention (500 Samples)**: With a 500-sample buffer, the model maintains strong retention of the original 28 European languages. For instance, high-resource languages such as Spanish (ES) remain stable, moving from a 4.09% baseline to 9.32% after four integration loops. However, we observe a steady performance degradation in

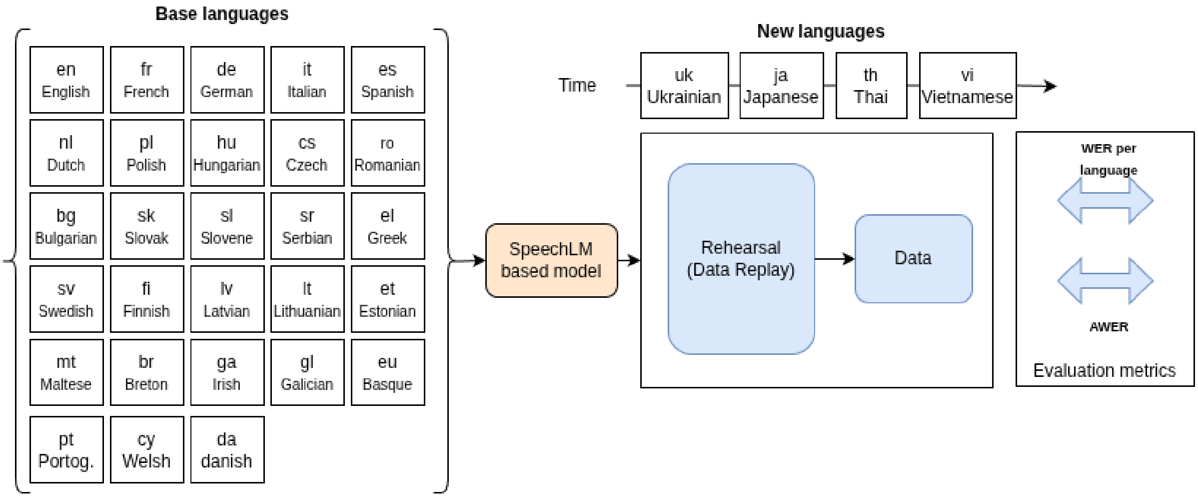


Figure 3: The proposed Continual Learning framework utilizing Data Rehearsal to mitigate catastrophic forgetting during the integration of new languages.

the newly acquired languages as the sequence progresses. While Ukrainian (UK) is initially learned at 17.10% WER, it drifts to 26.57% by the final loop. Similarly, Japanese (JA) worsens from an initial 27.28% to 40.45% following the addition of Thai and Vietnamese. This suggests that even a larger buffer provides only partial stability for newly integrated representations.

- **Accumulated Degradation (25 Samples):** In the aggressive 25-sample setup, we observe a "catastrophic drift" as the sequence progresses. The model reaches a breaking point during the Thai loop, where the **Avg WER** surges to **65.98%**. By the final stage, the AWER reaches 51.67%, more than double the error of the 500-sample experiment.

The divergence in stability is most evident when examining the **AWER** across all supported languages as shown in figure 4. In the 500-sample loop, the average error remains controlled, increasing from 15.74% at the base to 29.46% after the final loop. Conversely, the 25-sample loop experiences a "catastrophic drift," with the AWER nearly tripling to 56.06%. This global metric underscores that insufficient rehearsal does not just affect specific languages, but gradually destabilizes the entire multimodal alignment. These results, illustrated in our sequential evaluation logs, demonstrate that while Speech LLMs possess a high degree of "multilingual core" stability, a minimal rehearsal budget of 25 samples is insufficient to anchor representations over multiple sequential steps.

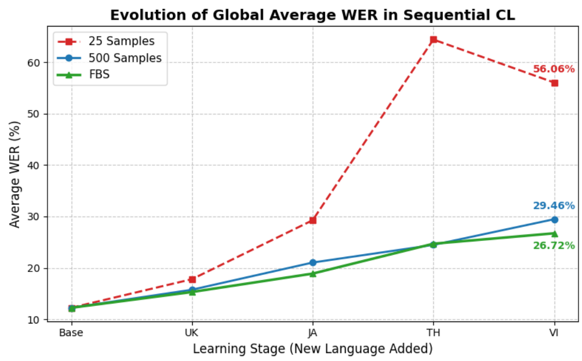


Figure 4: Evolution of Average WER across sequential learning loops (UK $\rightarrow$ JA $\rightarrow$ TH $\rightarrow$ VI). The 500-sample buffer (blue) maintains global stability, while the 25-sample buffer (red) exhibits a sharp catastrophic drift after the first loop, reaching a final Average WER of 51.67%. Finally the FBS based CL show the lowest AWER.

4.4. Fragility-Based Sampling

The varying levels of degradation observed in the uniform-buffer experiments indicate that not all languages require the same amount of rehearsal to prevent catastrophic forgetting. Motivated by this, we introduce a model agnostic **Fragility-Based Sampling (FBS)** heuristic. Instead of a uniform budget, FBS distributes the rehearsal samples based on the inherent stability (WER) of each language, keeping the overall system memory footprint constant.

For this experiment, we maintained the same total buffer capacity as the 500-sample uniform setup (28 languages $\times$ $500 = 14,000$ samples), since it is the buffer size that showed the best AWER. We categorized the base languages into three tiers and

Table 4: Sequential Continual Learning results: comparison between uniform rehearsal buffers (500 and 25 samples per language) and the Fragility-Based Sampling (FBS) strategy. Language codes: ES (Spanish), DE (German), FR (French), PT (Portuguese), EN (English), IT (Italian), PL (Polish), CS (Czech), DA (Danish), HU (Hungarian), BG (Bulgarian), RO (Romanian), ET (Estonian), LT (Lithuanian), EL (Greek), SL (Slovenian), FI (Finnish), UK (Ukrainian), JA (Japanese), TH (Thai), and VT (Vietnamese). The **AWER** column tracks the arithmetic mean across all active languages, illustrating the impact of different rehearsal strategies on global stability.

Loop Stage	**Buffer**	**ES**	**DE**	**FR**	**PT**	**EN**	**IT**	**PL**	**CS**	**DA**	**HU**	**BG**	**RO**	**ET**	**LT**	**EL**	**SL**	**FI**	**UK**	**JA**	**TH**	**VT**	**Avg WER**
Meusli (Base)	–	4.09	7.79	7.83	4.86	6.34	3.32	8.68	11.32	14.65	16.87	15.20	9.65	19.93	24.30	18.35	19.41	15.29	–	–	–	–	**12.23**
+ Ukrainian	25	4.95	10.96	7.60	6.38	8.97	6.67	10.98	17.23	22.22	25.68	18.62	15.62	30.87	32.20	33.34	27.33	23.35	17.16	–	–	–	**17.79**
	500	4.56	8.54	7.86	7.09	7.20	6.64	11.21	14.21	19.97	22.22	17.47	18.50	24.75	30.80	23.33	21.90	19.97	17.10	–	–	–	**15.74**
	FBS	4.78	9.01	7.92	7.31	7.18	6.80	11.26	14.94	18.38	22.66	16.78	14.35	25.03	28.90	22.84	22.30	18.30	16.62	–	–	–	**15.30**
+ Japanese	25	15.91	13.23	15.27	14.99	15.67	11.53	18.69	25.56	28.53	43.03	30.26	29.85	56.14	57.52	41.33	43.11	36.92	28.48	29.90	–	–	**29.26**
	500	7.06	10.58	10.51	9.65	8.40	7.67	13.54	18.86	22.83	34.98	22.49	20.78	32.79	39.43	29.78	34.41	24.36	24.17	27.28	–	–	**21.03**
	FBS	6.68	9.36	10.04	9.22	8.97	7.44	14.04	17.62	19.73	26.72	21.32	16.07	29.23	36.96	27.31	27.97	23.22	22.89	23.87	–	–	**18.88**
+ Thai	25	34.42	36.10	44.56	25.71	49.28	34.53	40.86	58.61	63.48	122.11	67.50	60.84	98.33	92.43	92.75	86.67	42.20	40.57	278.0	19.45	–	**64.42**
	500	8.23	11.73	11.49	11.20	10.93	8.89	16.57	23.41	25.34	41.67	26.83	26.95	34.69	41.27	38.23	35.39	30.48	28.15	38.62	19.21	–	**24.46**
	FBS	8.25	12.80	13.19	10.24	12.52	8.72	16.68	23.04	25.51	36.16	33.58	26.11	37.18	44.91	37.85	37.71	29.70	24.53	35.35	19.86	–	**24.69**
+ Vietnamese	25	23.47	32.43	34.72	22.75	23.81	30.71	40.04	58.64	52.25	90.36	55.09	55.20	82.19	80.51	74.73	67.59	64.85	24.99	142.44	106.45	13.95	**56.06**
(Final Loop)	500	9.32	16.73	12.73	11.60	11.16	11.59	18.53	26.05	27.22	42.49	27.88	27.46	42.50	48.93	37.12	38.18	34.59	26.57	40.45	91.62	15.98	**29.46**
	FBS	9.44	19.04	13.42	13.59	11.78	14.14	18.55	23.60	28.14	38.36	30.49	22.37	39.54	46.97	36.61	35.18	31.51	26.45	47.47	39.44	15.01	**26.72**

allocated samples inversely proportional to their baseline robustness:

- **Tier A (Robust):** 10 high-performing languages (e.g., ES, DE, FR) allocated 250 samples each ($2,500$ total).
- **Tier B (Intermediate):** 8 stable languages (e.g., PL, CS, UK) allocated the standard 500 samples each ($4,000$ total).
- **Tier C (Fragile):** 10 vulnerable languages (e.g., ET, LT, EL) allocated high-density reinforcement of 750 samples each ($7,500$ total).

As shown in Table 4, this reallocation strategy effectively balances the multilingual embedding space. By the final sequential loop (VT), fragile Tier C languages show improved retention compared to the uniform 500-sample setup; for instance, Estonian (ET) degrades to 39.54% (compared to 42.50% previously). Crucially, halving the buffer size for Tier A languages does not trigger catastrophic drift—Spanish (ES) finishes at a stable 9.44%, nearly identical to the 9.32% achieved with double the rehearsal data. Also the AWER, as shown in figure 4, results to be lower with respect to the uniform 500-sample setup. FBS achieves a significantly fairer and more equitable performance distribution across high- and low-resource languages.

5. Conclusions and Future Work

In this paper, we investigated the sustainable expansion of multilingual Speech LLMs through Continual Learning. Using the MEUSLI projector as a foundation, we demonstrated that while the SLAM-ASR paradigm is highly efficient for bootstrapping unseen languages like Ukrainian and Albanian, it is inherently vulnerable to catastrophic forgetting. Without intervention, fine-tuning on a single new language causes a near-total collapse of the model's original multilingual knowledge, often resulting in linguistic hallucinations. Our experiments with Data Rehearsal provide a clear roadmap for mitigating this collapse.

We established that stability can be maintained with surprisingly low rehearsal budgets—down to 5 samples per language—provided that random sampling is used rather than sequential selection. However, we also identified a "fragility threshold": when the budget is reduced to a single sample, performance on specific languages like Polish, Czech, and Estonian degrades significantly. Furthermore, our sequential integration loops (UK $\rightarrow$ JA $\rightarrow$ TH $\rightarrow$ VI) revealed that while a conservative buffer of 500 samples ensures high-fidelity retention across multiple stages, a minimal buffer of 25 samples leads to an accumulated "catastrophic drift," where the global Average WER nearly triples by the final stage. To address these non-uniform stability trade-offs, we introduced and evaluated a **Fragility-Based Sampling (FBS)** strategy. By categorizing languages into tiers, we optimized the rehearsal budget to distribute samples based on observed robustness rather than applying a uniform allocation. Our results demonstrate that this performance-aware reallocation significantly improves the retention of vulnerable linguistic representations without destabilizing robust languages. Crucially, FBS maintains a constant total memory budget while achieving a lower global Average WER compared to the uniform-buffer baseline, ensuring a much more equitable performance distribution across high- and low-resource languages. Future work will focus on automating this tiering process through dynamic sensitivity analysis during the training phase. Moreover, we will run experiments to better understand how the language diversity and order of these languages impact the final WER. By adaptively prioritizing vulnerable linguistic representations on the fly, we move closer to universal, resource-efficient Speech-LLM systems that can seamlessly grow their linguistic capabilities without sacrificing prior knowledge.

6. Acknowledgements

This paper was partially funded by the European Union's Horizon 2020 project ELOQUENCE (grant 101070558).

7. Bibliographical References

[1] C. Wenqian, Y. Dianzhi, J. Xiaoqi, M. Ziqiao, Z. Guangyan, W. Qichao, G. Yiwen, and K. Irwin, "Recent advances in speech language models: A survey," *arXiv:2410.03751*, 2024.

[2] P. Jing, W. Yucheng, F. Yangui, X. Yu, L. Xu, Z. Xizhuo, and Y. Kai, "A survey on speech large language models," *arXiv:2410.18908*, 2024.

[3] S. Fong, M. Matassoni, and A. Brutti, "Speech LLMs in Low-Resource Scenarios: Data Volume Requirements and the Impact of Pretraining on High-Resource Languages," in *Interspeech 2025*, 2025, pp. 2003–2007.

[4] I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville, and Y. Bengio, "An empirical investigation of catastrophic forgetting in gradient-based neural networks," *arXiv:1312.6211*, 2015.

[5] Z. Ma, G. Yang, Y. Yang, Z. Gao, J. Wang, Z. Du, F. Yu, Q. Chen, S. Zheng, S. Zhang *et al.*, "An embarrassingly simple approach for LLM with strong ASR capacity," *arXiv preprint arXiv:2402.08846*, 2024.

[6] D. Rolnick, A. Ahuja, J. Schwarz, T. Lillicrap, and G. Wayne, "Experience replay for continual learning," in *Advances in Neural Information Processing Systems*, 2019.

[7] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv:2106.09685*, 2021.

[8] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv:2212.04356*, 2022.

[9] P. H. Martins, P. Fernandes, J. Alves, N. M. Guerreiro, R. Rei, D. M. Alves, J. Pombal, A. Farajian, M. Faysse, M. Klimaszewski, P. Colombo, B. Haddow, J. G. C. de Souza, A. Birch, and A. F. T. Martins, "Eurollm: Multilingual language models for europe," *arXiv:2409.16235*, 2024.

[10] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *LREC 2020*, 2020, pp. 4211–4215.

[11] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "Fleurs: Few-shot learning evaluation of universal representations of speech," *arXiv preprint arXiv:2205.12446*, 2022. [Online]. Available: https://arxiv.org/abs/2205.12446

[12] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *ACL*, 2021.

[13] L. D. Libera, P. Mousavi, S. Zaiem, C. Subakan, and M. Ravanelli, "Cl-masr: A continual learning benchmark for multilingual asr," *arXiv:2310.16931*, 2023.

[14] B. Mu, P. Guo, Z. Sun, S. Wang, H. Liu, M. Shao, L. Xie, E. S. Chng, L. Xiao, Q. Feng, and D. Wang, "Summary on the multilingual conversational speech language model challenge: Datasets, tasks, baselines, and methods," *arXiv preprint arXiv:2509.13785*, 2025.

Responsible Benchmarking of Fairness for Automatic Speech Recognition

Felix Herron[1,2], Ange Richard[‡(2,3)],
François Portet[‡, (2)], Alexandre Allauzen[(1)], Solange Rossato[‡, (2)]
MILES Team, LAMSADE, Université Paris Dauphine-PSL (1)
GETALP Team, LIG, Université Grenoble Alpes (2)
PACTE, Université Grenoble-Alpes (3)
felix.herron@univ-grenoble-alpes.fr

Abstract

Many studies have shown automatic speech processing (ASR) systems have unequal performance across speaker groups (SG's). However, the manner in which such studies arrive at this conclusion is inconsistent. To pave the way for more reliable results in future studies, we lay out best practices for benchmarking ASR fairness based on literature from machine learning fairness, social sciences, and speech science. We then perform a case study on the Fair-speech benchmark, applying aforementioned best practices, and discuss how failing to do so can result in erroneous conclusions. On the whole, we advocate for as fine-grained an analysis as possible, taking into account as many variables as are available, in order to eschew dataset-level bias.

Keywords: Fairness, benchmarking, statistics

1. Introduction

In recent years, automatic speech recognition (ASR) software has grown increasingly performant (Nayeem et al., 2025), which has led to a complementary increase in prevalence of ASR use among diverse populations (Yang et al., 2024; Wald et al., 2024; Dino et al., 2025). It is therefore increasingly imperative to ensure that existing ASR systems perform equally regardless of the identity of the speaker. There is ample research demonstrating that certain speaker groups (SG's), such as children and non-native speakers, are treated worse than others by ASR systems. However, the methodology for identifying such SG bias in ASR systems is inconsistent across studies and sometimes marred by lack of precise analysis of the multifaceted identities of individual speakers.

Indeed, applying fairness benchmarks without sufficient oversight can lead researchers to stumble into erroneous conclusions which are incongruous with real world biases, a common blunder in fairness research (Selbst et al., 2019). This paper discusses the importance of defining SG-level fairness as intentionally and precisely as possible to avoid such blunders. We start by diagnosing our observed lack of consistency across ASR fairness benchmarking studies, and suggest this is due in part to a lack of clarity about how SG-level fairness should be defined and measured. We then list several best practices to avoid accidentally measuring bias stemming from fairness corpora rather than real-world bias. We formally define fairness in ASR, as motivated by broader fairness literature in machine learning (ML). Finally, we perform a case study on a common fairness corpus, Fair-speech, and apply many of the best practices to our analysis. We highlight some pitfalls that could entrap unaware users of the corpus. We finish by suggesting how future work, particularly in dataset creation, can help facilitate fairness benchmarking in ASR.

‡ contributed towards formulating framework of speaker group intersectionality and multivariate SG's.

2. Motivation

This study is motivated by a lack of consistency both within single and among studies examining fairness in SOTA ASR systems. We were puzzled to find that some studies report that women experience *significantly worse* treatment (Hutiri and Ding, 2022; Garnerin et al., 2021; Tatman and Kasten, 2017; ElGhazaly et al., 2025), *significantly better* treatment (Veliche et al., 2024; Feng et al., 2024, 2021; Abushariah and Sawalha, 2013), or *both*, depending on subgroup studied (Attanasio et al., 2024; Tatman, 2017; Kulkarni et al., 2024). Likewise, while most studies find that non-native speakers are *less well understood* by ASR systems (Ghorbani and Hansen, 2018; Zhang et al., 2022; Sekkat et al., 2024), some *find the opposite* (Veliche et al., 2024). Studies seem to all agree that *children receive worse performance* by ASR systems; however, whether older adults are better understood than younger *varies broadly by publication* (Aman et al., 2013; Feng et al., 2021; Kulkarni et al., 2024; Sekkat et al., 2024).

It is possible that this effect is due in some part to

Proceedings of SPEAKABLE @ LREC 2026, pages 66–78
11 May 2026.

the different ASR models being evaluated by each of these studies. Fixing a dataset and two SG's g_1,g_2 (e.g. native vs non-native speakers), some studies find that different ASR models perform better on g_1 while others on g_2 (Attanasio et al., 2024). However, most studies in the literature find that different ASR systems tend to be biased against the same SG's (Feng et al., 2024, 2021; Fuckner et al., 2023). If we assume that all ASR systems reflect the same SG-level biases, we would hope that all studies into ASR fairness would arrive at similar conclusions. That they don't is therefore likely due to methodological variance which allows researchers analyzing the same data to arrive at different conclusions.

3. Best practices in reducing transmission of dataset bias

It is important to remember that any conclusions reached by fairness studies are *estimates* of real life bias, influenced by data on which the experiments were performed. With this in mind, researchers should try and limit dataset-level bias as much as possible so that their estimates are as close as possible to a real world simulation. In this section, we will highlight several key notions to attenuate filtration of biases due to dataset construction into fairness results. For each, we cite examples from the literature. It is important to note that these best practices remain general - we see them as akin to tools in a belt, whereby the user still must know how to use each in the manner most beneficial to their use case. The subsequent section describes a case study on how these tools can be put to use - however, each setting is different.

3.1. Ensure equal distribution of recording quality

Background noise (and recording quality in general) have been shown to impact ASR performance (Rodrigues et al., 2019). It is possible that some SG's will have different levels of background noise, thus potentially biasing results of a fairness study. Some benchmarks circumvent this by recording all of their inputs in the same conditions (Sekkat et al., 2024; Veliche et al., 2024), though benchmarks compiled from diverse sources cannot (Meyer et al., 2020). That said, it can be useful to have variably noisy recordings in the dataset, as real life ASR often occurs in noisy environments using old recording equipment. Furthermore, some SG's are more likely experience such potential impediments to ASR transcription, such as the "low" socio-economic status SG in the Fair-speech dataset (Veliche et al., 2024).

3.2. Verify text complexity

If different SG's use different vocabulary/grammar (i.e. text) in the real world, it is important that these attributes be taken into account during bias testing. A corpus where all speakers speak comparable texts cannot be considered to faithfully estimate bias if this is not the case in the real world; likewise, if a certain SG in a corpus is comprised of recordings of complex texts, while another SG speaks easy texts, this likewise has the potential to engender unfaithful bias estimations. On the other hand, if one is trying to capture *merely* the bias due to acoustic features of a speaker's voice, then controlling for text (complexity) is beneficial. Once again, this is a decision that researchers must consider intentionally in the context of their individual study.

For example, Koenecke et al. (2020) proposes calculating each text's perplexity to assess whether each SG speaks similarly complex sentences. They also measure the "dialect density" of text to determine the extent to which it contains grammar typical of African American Vernacular English.

3.3. Understand intra-SG speaker diversity

When we measure ASR performance w.r.t. a SG, it is tempting to treat SG labels as precise, immutable defining characteristics of speakers. However, speakers within a given SG can be diverse. It is therefore essential to understand how each SG defined, what it means to belong to multiple SG's at once, and SG's are balanced throughout the dataset.

3.3.1 Intersectionality Traditionally, fairness in ML has focused on comparing between outcomes for single groups from the same demographic variables (DV's), such as between different races or sexes (Bellamy et al., 2018; Friedler et al., 2018). However, recently the application of **intersectionality** in ML fairness research has gained traction as a technique to more precisely gauge bias (Foulds et al., 2019; Wang et al., 2022). These studies argue that measuring fairness w.r.t. a single DV in isolation is insufficient; to best understand fairness, one must look at as fine-grained SG's as possible, comprised of as many DV's as possible.

Intersectionality has its roots in the social sciences where Crenshaw (1989) defines it as discrimination faced by members of multiple marginalized classes at the *intersection* of several groups, for example Black. Wang et al. (2022) emphasizes that treating heterogeneous groups as a monolith (e.g. all people from Pacific islands) can hide unfair treatment experienced by subgroups thereof (e.g. specific islands). Foulds et al. (2019) emphasizes the importance of prioritizing protected classes which are underrepresented in fairness benchmarks, as their discrimination can be more easily ignored than a large underprivileged group.

Several existing benchmarks for fairness in ASR already consider the intersectionality of SG's (without necessarily using that exact term) for both aforementioned motivations. For example, Feng et al. (2021) and Feng et al. (2024) examine the WER gap between Dutch spoken in Flanders vs in the Netherlands, intersecting with regional dialects, native-ness of speakers, age, and speech format (read vs. human-machine interaction). One finding is that the gender-based WER gap is least significant among children, while the age-based WER gap is most significant for women, as well as that the regional gap is strongest among children and teens, and weakest among older adults. This is a crucial insight that would be ignored if the authors had only compared between genders, ages, or regions.

3.3.2 Conditional statistical parity

The metrics used to measure fairness in ASR correspond to **statistical parity** as introduced by Verma and Rubin (2018) in their landmark fairness taxonomy paper. A more rigorous version of this is **conditional statistical parity**, which requires that all secondary attributes about the setting be the same, such as background noise or text complexity. Furthermore, we can condition on/take into account secondary DV's in our calculations. This is important both in order to uncover intersectional biases, as well as to equilibrate potential unbalance in other DV's. Failure to do this might end up spuriously measuring the random side-effects of subgroup imbalance captured during dataset construction.

As a toy example, let us imagine we are measuring the fairness of performance of an ASR system on men vs women, on some benchmark B. By chance, 1 in 20 men in B suffer from Parkinson's disease (diminishing their ASR comprehensibility (Moro-Velázquez et al., 2019)), but only 1 in 100 women in B suffer from Parkinson's. If we are either unaware of this, or ignore it, then we might conclude that men experience worse ASR performance than women; however, our observation would at least in part be due to the confounding influence of a higher prevalence of Parkinson's in the male population, rather than due to their masculinity.

Koenecke et al. (2020) responsibly attempt to avoid contamination of their race DV by either age and gender by retaining the same proportion of both gender and age groups for both White and Black speakers. However, they don't consider the intersection of age and gender, an oversight of intersectionality. They then delve into the geography of both racial groups where, crucially, they find that race is not sufficient to explain the WER gap; Black Americans from Rochester had comparable WER to White Americans.

Sekkat et al. (2024) control for the confounding effects of other DV's, noting that this causes some univariate effects w.r.t. age and gender to vanish. Likewise, Tatman (2017) finds a greater difference in by-gender performance in certain dialect groups.

On the other hand, Veliche et al. (2024) finds that men have twice as high a WER as women, which they explain by citing previous work showing men tend to have worse ASR performance. However, while they note that the men in their dataset are far more likely to be African-American than their women, they don't perform an intersectional analysis to interrogate what is likely at least partially responsible for that effect.

Another example of failure to take SG diversity into account is the "Asian" dialect category in the Sonos dataset (Sekkat et al., 2024), which is comprised of speakers from Southern as well as Eastern Asian countries. The authors acknowledge the extraordinary diversity of this category and the incumbent challenges this causes in interpreting results based on it.

3.3.3 Beware (un)known confounding factors

In the previous examples, researchers were able (or failed) to avoid jumping to conclusions more reflective of biases in their dataset than biases in their ASR systems. Or they were able to uncover intersectional bias specific to multidimensional SG's. However, these effects can only be explicitly controlled when potentially confounding metadata are available in the dataset. For example, the Sonos dataset has ethnicity tags for only a small portion of its speakers, where they show it to have a statistically significant relationship with ASR error rate (Sekkat et al., 2024). However, they cannot control for equal ethnicity distribution over the rest of the dataset, and therefore cannot control this bias.

Indeed, there may be many other confounding variables related to speaker identity which are not included in the dataset. It is by definition impossible to directly control for these; however, authors could estimate the extent to which their datasets are free from such effects by using phonetic priors. For example, studies have shown that human's voices don't change very much during middle age (Rojas et al., 2020). The reliability of fairness experiments can therefore be benchmarked by performance variance across middle-aged age groups: if there is significant performance difference between any two middle-aged SG's, that is an indication of methodological error, likely due to lack of balancing (Sekkat et al., 2024; Veliche et al., 2024).

3.4. Define SG-level performance based on speaker-level performance

When measuring SG-level bias, it is imperative to calculate error for each SG as a function of error for each speaker adhering to said SG. This is based on two observations: first, utterances from the same speaker are not independent, and thus we cannot perform a statistical test that assumes independence of samples. Second, this avoids bias due to imbalance in representation for each speaker. For example, if speaking time is not equally distributed across speakers in SG (e.g. D'_{SG} contains many more utterances/words for some speaker S_1 than another speaker S_2), then calculating SG-level error as a function utterances will engender bias towards Speaker S_1, and will not be a faithful representation of the SG overall.

Not all studies adhere to this principle, however. Sekkat et al. (2024) explicitly argues for the simplicity of measuring fairness based on individual utterances. Furthermore, Feng et al. (2021) and Feng et al. (2024) base some of their conclusions on a small numbers of speakers (see i.e. Table 1 in Feng et al. (2024)). They claim statistical significance, likely based on the number of overall samples or hours of recorded speech, rather than the diversity of speakers per SG.

3.4.1 The challenge of speaker paucity

If we define SG-level performance as a function of individual speakers, the statistical significance of our results will depend on the number of speakers in each SG. This leads us to a set of contradictory incentives: the more precisely we define SG's as the intersections of multiple DV's (as encouraged in the previous section), the more precise are our conclusions into SG-level fairness. However, the more precisely we define SG's as the intersections of multiple DV's, the fewer speakers will be included in each class, thus reducing the significance of tests we perform on them. This is logical: if a SG contains too few speakers, we risk measuring bias due to the unique nature of those several individuals, rather than due to the SG they belong to.

We can derive the number n of speakers per SG necessary for statistical (with confidence α and power β) significance in, for example, a one-sided two-sample Z-test (e.g. comparing the mean error of two SG's given fixed population variance) with test-statistic Z:

$$
\begin{aligned}
&Z := \frac{\hat{\delta}}{\sigma\sqrt{2/n}} \\
&Z > z_\alpha + z_\beta \implies \frac{\hat{\delta}^2}{\sigma^2 * (2/n)} > (z_\alpha + z_\beta)^2 \\
&\implies n > 2 * \frac{(z_\alpha + z_\beta)^2 \cdot \sigma^2}{\hat{\delta}^2} \qquad (1)
\end{aligned}
$$

where $\hat{\delta}$ is the difference in estimated WER for two SG's, σ the variance between speakers, and $z_{\alpha,\beta}$ the quantiles defined by the significance and power respectively. For example, given typical values of 95% confidence with 0.8 power, taking $\hat{\delta} = 0.1$ and $\sigma = 0.15$ (reasonable estimates based on our analyses in Section 5), we would need $n \approx 35$ speakers per SG. This could be a serious hindrance for corpora with few speakers, and/or multivariate SG's defined by many DV's.

This bound cannot be shrunk simply by increasing the number of utterances per speaker. $\hat{\sigma} := \sigma + \epsilon$, where $\epsilon > 0$ varies inversely to the number of words for each speaker - the more words available per speaker, the smaller ϵ and thus the less noisy $\hat{\sigma}$, which results in smaller n. However, $\hat{\sigma}$ can never be lower than σ, which means that an increase in the number of words per speaker has a limited effect on improving our measured error bounds.

3.5. (Sometimes,) aggregate SG's

Just because metadata are available in a corpus doesn't mean they are useful in fairness analysis!

3.5.1 Too few speakers per SG

In this case, we will be unable to draw statistically significant conclusions based on performance over this SG. Therefore, it could make sense to create an "other" category which groups semantically unusual SG's together. The upside of such aggregation is that it allows "other" to be represented by sufficient speakers so as to be statistically significant, whereas as initially constituted, those SG's would be statistically meaningless. The downside is that this marginalizes the individual identities of those underrepresented SG's, which Wang et al. (2022) warns against. However, we are limited by the data available, and there is interpretable value in creating a class to compare with the mode SG's.

3.5.2 Superfluous level of precision in metadata

We stand to gain nothing by measuring fairness w.r.t subgroups defined by attributes likely independent of ASR fairness, for example different middle-age subgroups (Rojas et al., 2020; Hustad et al., 2021). Instead, this can lead to two harms: 1) we risk accidentally creating SG's which are unbalanced w.r.t. other underlying SG's, thereby potentially corrupting our analysis. 2) it reduces the number of speakers in each multivariate SG, thereby lowering the statistical significance of results. Thus, we might consider aggregating all middle-aged speakers into the same SG.

3.6. Outlier speaker removal

A final source of bias potentially contaminating our SG-level fairness results are outlier speakers. If we want to want to measure the general behavior of a certain SG, which potentially contains few speakers (due to dataset constraints), and one of the speakers is understood much worse than the rest, that is potentially due to some individual speaker-level characteristics which are not relevant to our analysis. It can therefore make sense to exclude extrema values from SG-level mean calculation, for example. This is less of a problem for datasets with very high numbers of speakers; however, that is unfortunately rarely the case in practice.

4. Quantifying fairness in ASR

We describe the two metrics most often used in the literature to quantify bias in ASR systems.

4.1. Relative SG-level error/WER gap

Most studies into SG-level ASR fairness measure the relative error rate for each SG. For a dataset D, they calculate the word error rate (WER) for each utterance, a measurement of the number of substitutions, deletions, and insertions necessary to correct the automatic transcription by an ASR model M over some $D' \subseteq D$. Some studies then calculate the average WER for each SG as the average WER for all utterances $D'_{SG} \subseteq D'$:

$$\text{WER avg.}^*(D'_{SG}) := M^*(D'_{SG}) := \frac{1}{|\{u \in D'_{SG}\}|} \sum_{u \in D'_{SG}} \text{WER}(u;M) \tag{2}$$

However, as mentioned in Section 3.4, Metric 2 falls immediately into a hazard of imprecise measurement by failing to first average by speaker. A more prudent approach is:

$$\text{WER avg.}(D'_{SG}) := M(D'_{SG}) := \frac{1}{|\{S \in SG\}|} \sum_{S \in SG} \frac{1}{|\{u \in D'_S\}|} \sum_{u \in D'_S} \text{WER}(u;M) \tag{3}$$

Then, one can measure bias against a particular SG SG_i in terms of the relative performance between SG_i and some the rest of the subset $D' \subseteq D$ (typically, studies use $D' = D$) (Veliche et al., 2024; Feng et al., 2024):

$$\text{Err}_{rel}(SG;D',M) := 100 \times \frac{M(D'_{SG}) - M(D')}{M(D')} \tag{4}$$

Some studies also calculate the unfairness w.r.t. a DV as the difference between the best and worst relative error for two constituent SG's, often denoted the WER gap (ElGhazaly et al., 2025; Attanasio et al., 2024; Kim et al., 2025):

$$\text{Unfairness} := \text{WER gap}(DV;D',M) := \max_{SG_i \in DV} \{\text{Err}_{rel}(SG_i;D',M)\} - \min_{SG_j \in DV} \{\text{Err}_{rel}(SG_j;D',M)\} \tag{5}$$

Then, one can perform a statistical test, such as a 1-sample (or 2-sample, if comparing between pairs of SG's rather than with respect to the dataset on average) t-test, to determine whether the relative error and unfairness are statistically significant for SG's and DV's respectively. An ASR model is deemed fair w.r.t. a SG if it delivers statistically negligible relative error, and fair w.r.t. a DV if it delivers a statistically negligible WER gap.

4.1.1 Isolated effect of single DV's As mentioned in Section 2, we recognize that individual SG's potentially contain heterogeneous subgroups. Thus we seek to isolate the effect of individual DV's on ASR performance. For any $DV^i \in [DV^1,..,DV^k]$ (e.g. gender), one of many DV's included in metadata, we define a subset $D'_{cond,i} \subseteq D$ by fixing a SG_{DV_j} for every other $DV_j \in [DV^1,..,DV^k] \backslash \{DV_i\}$ (e.g. only children, only non-native speakers, only African Americans, etc.):

$$D'_{cond,i} = \bigcap_{j \in [1..k] \backslash \{i\}} D_{SG_{DVj}} \tag{6}$$

and calculate $\text{Err}_{rel}(SG;D'_{cond,i},M)$ for all $SG \in DV^i$ as well as WER gap$(DV^j;D'_{cond,i},M)$. We can then aggregate mean performances for each permutation of other the other DV^j and perform a statistical test, such as 1-sample t-test, to determine, w.r.t. a $SG \in DV^i$, whether the means of the relative error rates were statistically significantly different from zero, or w.r.t. DV^i overall, whether the unfairness levels were statistically significantly greater than zero.

4.1.2 Worst treated multivariate SG's We define *multivariate SG's* as the intersection of every DV available in the dataset. For example, if the dataset is annotated with gender, age, and native language, an multivariate SG might be "female children who speak native English". For each multivariate SG, we calculate the relative error w.r.t. the dataset overall (Eq. 4); then observe which multivariate SG's, if any, have the lowest/highest relative performance. This will uncover SG's whose marginalization is compounded by their intersectionality, as proposed by Wang et al. (2022).

5. Case study on Fair-speech

Unfortunately, many of the most prominent corpora for evaluating ASR systems do not permit bias evaluation due to a lack of sufficient recorded demographic metadata (Ma et al., 2024; Panayotov et al., 2015; Linguistic Data Consortium, 2013) or unreliable labeling thereof (Ardila et al., 2020; Wang et al., 2024). However, there are several corpora specifically designed for bias/fairness evaluation of ASR systems whose multitudinous metadata categories permit finer-grained ASR evaluation. We proceed to analyze the Fair-speech corpus (Veliche et al., 2024) and discuss how to implement some of the best practices from the previous section.

We replicate each experiment using three different near-SOTA ASR models: Whisper-medium (`Whisper`), wav2vec2-large-960h-lv60 (`Wav2vec 2.0`), and wav2vec2-large-xlsr-53-english (`XLS-R-En`). `Whisper` was trained end-to-end for ASR on 680k hours of YouTube transcripts (Radford et al., 2022); `Wav2vec 2.0` was pretrained on 60k hours of LibriLight (Baevski et al., 2020); `XLS-R-En` was pretrained on a multilingual corpus comprising 53 languages (Babu et al., 2021). `Wav2vec 2.0` was finetuned on 960h of LibriSpeech (Panayotov et al., 2015); `XLS-R-En` finetuned on the English split of CommonVoice (Ardila et al., 2020). We can thus test whether our results are specific to one architecture/training set, or general across ASR systems.

5.1. Dataset description

The Fair-speech Dataset (Fair-speech) comprises 593 speakers over 56 hours (Veliche et al., 2024). Fair-speech is comprised of recordings of paid speakers speaking (not reading) smart speaker commands. Speakers self-report metadata including: gender, age, ethnicity, first language, and socioeconomic background. See Fig. 1.

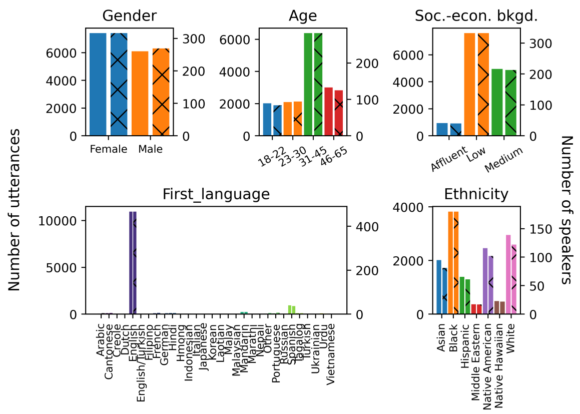


Fig. 1: Overall distribution of demographic labels in the Fair-speech corpus. Note that most speakers are native English speakers, with a small (illegible) minority of Spanish native speakers.

Fair-Speech represents many first languages; however, given a limited number of speakers per language, that might limit the statistical significance of results pertaining to speakers of sparsely represented languages. Furthermore, the number of age categories is likewise too precise - we likely stand to gain little by differentiating between different classes of middle-aged adults. Fair-speech lacks children and old adults, the two age-related SG's with phonetic motivation for divergent ASR error rates (Hustad et al., 2021; Rojas et al., 2020).

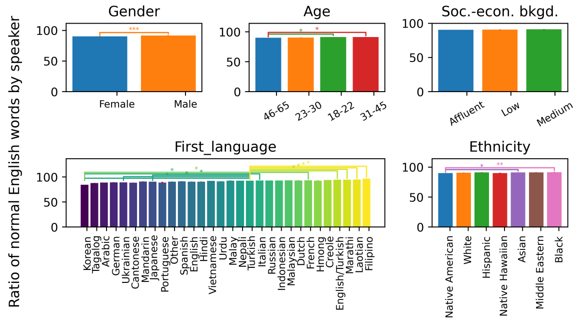


Fig. 2: Ratio of non-English words per sentence, averaged by speaker.

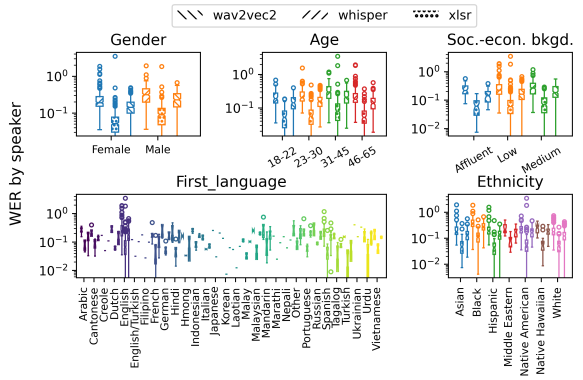


Fig. 3: Variance of WER for each ASR model, averaged by speaker.

5.2. Filtering out outlier speakers and utterances

First, we consider filtering out outlier speakers and utterances based on WER for each ASR model. Figure 3 shows the variance of WER for each univariate SG and model. Note that some speakers have an average WER of over 1 - this is likely due to speaker-specific anomalies and not an accurate reflection of the ASR system overall. We proceed by filtering out all speakers (and utterances per speaker) with a z-score of >3 in each analysis that

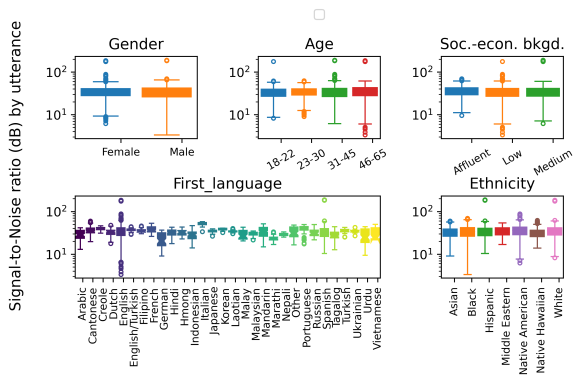


Fig. 4: Signal-to-noise ratio for each recording.

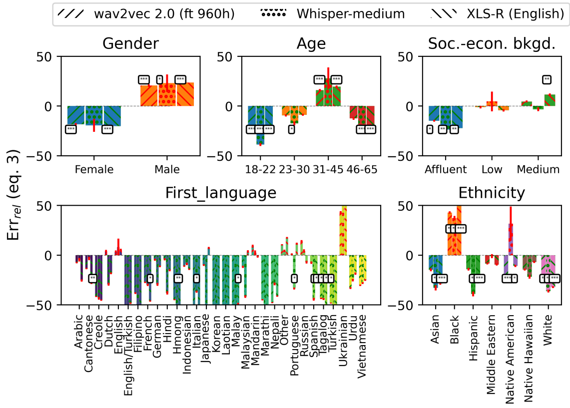


Fig. 5: Fair-speech when measuring relative WER gap between over SG's belonging to a single DV at once. Values < 0 indicate below average WER, i.e. above average performance. * denotes statistically significantly greater/less than 0 according to a 1-sample t-test - * implies $p < 0.05$, ** implies $p < 0.01$, *** implies $p < 0.001$.

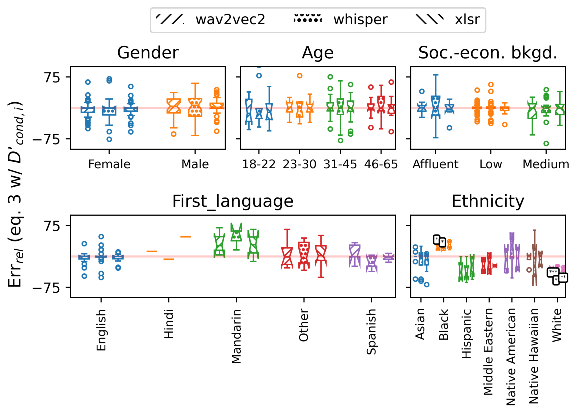


Fig. 6: Fair-speech when measuring relative WER gap between intersectional SG's differing only on one specific DV. Each datapoint is a statistically significant difference between SG's differing by only one DV. Values < 0 indicate below average WER (above average performance). * denotes statistically significantly greater/less than 0 in the aggregate according to a 1-sample t-test. * implies $p < 0.05$, ** implies $p < 0.01$, *** implies $p < 0.001$.

we conduct.

5.3. Recording quality and text complexity

Figure 2 shows the average text complexity for every speaker, measured by number of words in the transcript that are not standard English (we use NLTK English dictionary (Loper and Bird, 2002)). Overall, there is little variance, particularly among SG's with high representation; that said, the most

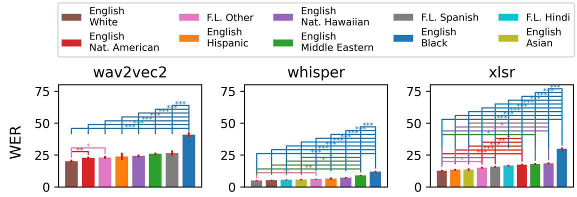


Fig. 7: Overall WER for Fair-speech when conditioning on first language and ethnicity. * implies denotes significantly higher WER on a one-sided, two-sample two-sample t-test (* implies $p < 0.05$, ** implies $p < 0.01$, *** implies $p < 0.001$).

disadvantaged SG's, as we will see in later analyses, are not necessarily those with the lowest ratio of standard vocabulary.

Figure 4 shows the signal-to-noise ratio of each utterance, broken down by SG. Note that most recordings have are 10 dB (a reasonable threshold for ASR performance (Bouchakour and Debyeche, 2022)). For our experiments, we will remove all recordings at $< 10dB$.

5.4. Calculating relative WER for each SG

We begin our case study by analyzing the results from Fair-speech. First, in Figure 5, we present relative error rate as the dataset was constructed, without conditioning or manipulation. We measure statistically significant performance discrepancies w.r.t. each of the five DV's recorded in Fair-speech. Our results correspond to what was initially published in (Veliche et al., 2024). Several odd results stand out, which raise some red flags about our experimental setup:

1. 31-45 year-old's have higher WER than all other age groups. This is likely evidence of poor subgroup balancing, as there is no logical reason for different age groups of middle-aged adults to have variant performance.

2. Men have vastly higher WER than women. Veliche et al. (2024) attempt to explain the gender discrepancy by citing previous work showing men tend to have worse ASR performance; however, 100% worse is much higher than peer studies (Sekkat et al., 2024; ElGhazaly et al., 2025; Attanasio et al., 2024).

3. Most first languages have statistically insignificant relative WER. This is due to those SG's not being represented by enough speakers in Fair-speech.

4. Native English speakers have negligibly higher worse-than-average WER, while several non-native speakers have statistically significantly

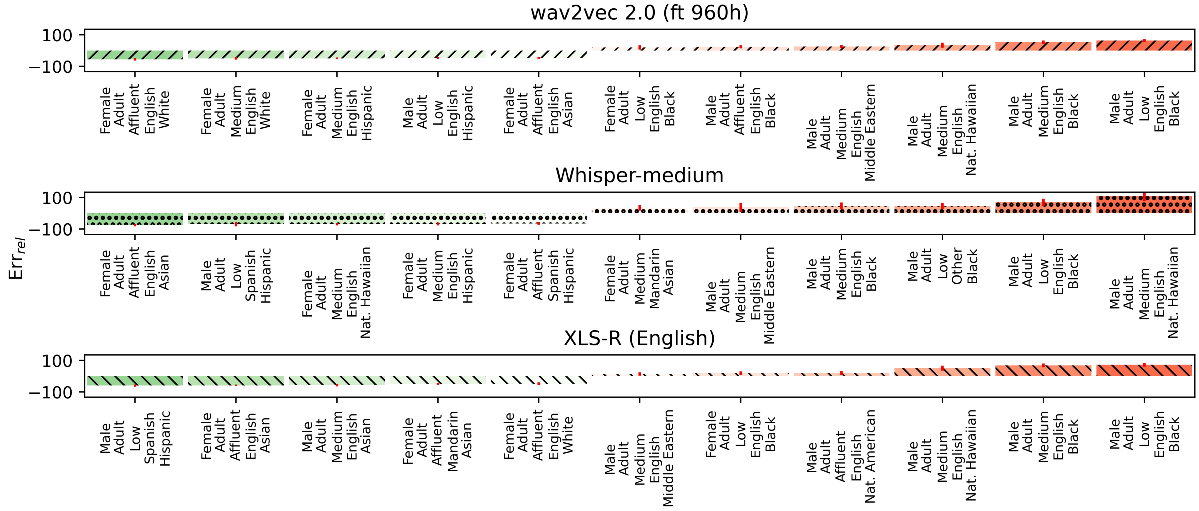


Fig. 8: Relative error of intersectional SG's in Fair-speech with least, greatest WER (conditional on sufficient speakers w.r.t. Eq. 1).

better-than-average WER. This stands in contrast to most peer studies (Feng et al., 2024; Fuckner et al., 2023; Sekkat et al., 2024; Ghorbani and Hansen, 2018). This is potentially an artifact of disregarding intersectionality of multivariate SG's.

5.4.1 Analysis of multivariate SG's can help

Issue 1 is a warning that our "age" DV is probably improperly balanced w.r.t. other DV's. We reiterate that there is no good reason to analyze separate age groups of middle-aged adults; however, to investigate the hypothesis of imbalance in other DV's, we propose comparing only multivariate SG's where all DV's are the same except for age (see Section 5.4). Figure 6 shows that in doing so, the divergent performance w.r.t. age vanishes. That said, we note that the distribution for each class contains many outliers - this effect would likely diminish if we raised our threshold for number of speakers per multivariate SG. On the whole, this experiment supports our balancing hypothesis and supports the technique of multivariate comparison to avoid DV imbalance.

Issue 2 is similar to issue 1, and is similarly alleviated when considering only multivariate SG's differing only by gender (see Figure 6). As before, we note that the variance for both men and women is rather high - for some multivariate SG's, women have much higher WER than men, and vice versa. This further motivates in-depth analysis of individual multivariate SG's to uncover potential intersectional SG's with compounded ASR error.

We note that the ethnicity DV is the only one to deliver statistically significantly different results - for Black and White speakers. Based on this experiment, we conclude that being Black is associated with inferior ASR performance across every permutation of other DV's, while the opposite is true for being White. This is a stronger conclusion than what we were able to draw from Fig. 5.

5.4.2 Conditioning on all DV's is no silver bullet

Conditioning on all DV's allowed us to draw strong conclusions about Black and White speakers; furthermore, we can reasonably rule out gender, age, or socio-economic background having broad impacts on ASR performance. However, we cannot draw statistically significant conclusions about many first languages (issue 3) even after aggregating uncommon languages into "other", due to persistent lack of representation across multivariate SG's. Thus, equipped with our conclusions about insignificance of gender, age, and socio-economic background, we propose removing conditions on those three DV's, considering only difference in first language and ethnicity. Furthermore, for first languages which we retained after aggregation, we note that apart from English, there is one primary ethnic group that covers nearly all speakers of that language (all native Mandarin speakers are of "Asian" ethnicity, almost all native Spanish speakers are "Hispanic", etc). Thus, we move to condition on ethnicity *only* for native English speakers.

Figure 7 allows us to get a better sense of bias against multivariate SG's defined by first language and ethnicity. We find once again that Black native English speakers are less well understood than every other class in every model. One added insight is that not all native English speakers are treated the same - White speakers are the best native English speakers understood by every model, though not statistically significantly so for all ethnicities. This provides a clearer picture than the previous experiment which showed that White speakers were statistically always better understood than the mean.

It also provides insight into issue 4, that non-native speakers are often better understood than native English speakers. When we condition on ethnicity, we find that it is only Black native English speakers that are worse understood, and that there is little statistical consensus regarding the relationship between being a native speaker and ASR error.

5.4.3 Intersectional multivariate SG's with compounded ASR error Finally, we can analyze the intersectional SG's which have the least and greatest WER, which we show in Figure 8. One surprising finding is that the worst performing group overall for `Whisper` is a Native Hawaiian group, which overall experienced much better treatment than Black speakers (Figure 7). One downside of this analysis, however, is that given the large difference in performance in extrema groups from the mean, fewer speakers are necessary to draw statistically significant conclusions based on SG (e.g. Eq. 1). Thus, such fine-grained analysis has heightened risk of drawing erroneous conclusions.

6. Discussion and outlook

The primary takeaway from this work is an exhortation to future studies on fairness in ASR to be as fastidious as possible designing their experiments. We underscore the importance of an intimate understanding of the datasets on which one is evaluating before designing experiments, tailoring experiments to that data, and being transparent about limitations that can be drawn therefrom.

We also encourage authors to clearly delineate the questions which they seek to answer. On the one hand, we can estimate how individual DV's affect performance of ASR systems, like in RQ 2 - this is primarily useful for understanding ASR systems from a computational level and could help steer future work towards disaffected SG's in particular. On the other hand, we can estimate which SG's defined by the intersection of multiple DV's are treated the absolute best and worst by ASR systems. This gives us greater sociological insight into the ramifications of unfair ASR - if a speaker belongs to such a class, they risk acute discrimination.

Furthermore, we encourage humility on behalf of researchers into ASR fairness in the face of statistical uncertainty. As we describe in this study, current fairness benchmarks suffer from lack of speaker diversity. Using overly broad SG's reduces the narrative power of their analysis, while using overly precise SG's silos speakers into groups without enough speakers, rendering conclusions statistically insignificant. With these two goalposts in mind, given the constraints of current ASR fairness benchmarks, **the conclusions we can draw from this type analysis remains limited**. Indeed reporting a result as statistically insignificant or sociologically broad isn't a problem (as long as this is duly noted); rather, it is a reflection of the reality of limitations of current corpora. This leads to the obvious recommendation to collect more data with high-quality annotations. However, this is both expensive and ethically fraught - Meyer et al. (2020) explicitly avoids children, while some countries forbid the analysis of ethnic minorities (fre, 1978).

It is important to note that the DV's included in the three benchmarks which we studied are themselves sources of potential bias. Had the benchmark designers decided to record different metadata, our results would reflect that, both in our ability to observe both the SG's which trigger unfairness, as well as the intersectional SG's which maximize unfairness. Future work in defining ASR fairness datasets should be in consultation with sociologists and phoneticians to determine which DV's, and SG's therein, are a) important in the larger context of social discrimination and b) likely to contribute to disparate ASR performance.

We encourage future study into linguistic or phonological mechanisms which are the actually underlying causal drivers of SG-level unfairness, beyond SG labels. Future work might eschew some of the pitfalls of relying on unbalanced and heterogeneous SG's advertised in this study by focusing on more rigorously defined measurements, such as dialect density measure (Koenecke et al., 2020) or speed of speech (Meng et al., 2022). Alternatively, unsupervised feature discovery has also been shown to uncover proxies for disadvantaged SG's in ASR without having to explicitly label them (Dheram et al., 2022; Alonzo-Canul et al., 2025). In other areas of ML, where fairness depends on sensitive attributes being evenly distributed in decision functions, there is work in automatically selecting attributes to include in fairness analysis (Pelegrina et al., 2023). This approach takes SG intersectionality into account by testing which combinations of attributes are related to an outcome variable.

One potential avenue for more precise fairness analysis is in conditional synthetic voice generation, such as (Sadok et al., 2025). This allows for two utterances from two different speakers to be *exactly* the same, apart from some parameters that are meant to mimic particularly DV's. For example, Masson and Carson-berndsen (2023) show that artificial generation simulates the patterns of non-native speakers well in ASR systems. This would address one major shortcoming of any current ASR fairness study, which is unable to truly isolate individual speaker characteristics while selecting from a small population of speakers.

7. Ethics statement

Collecting recordings of minority groups, particularly children, requires care to avoid revealing their identities. Fair-speech avoids children altogether. Furthermore, our work is meant to increase fairness in ASR; however, by focusing on a small number of datasets, we potentially overlook SG's which face ASR discrimination, thereby reinforcing it. In this case we didn't consider patients with adverse health conditions, for example, a group which has been studied extensively in ASR fairness research and no less deserving of attention than those highlighted in our study (Moro-Velázquez et al., 2019). By avoiding analyzing SG's for which benchmarks don't provide enough data, we are reinforcing the discrimination likely behind this unbalance in the first place.

8. Bibliography

1978. Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés.

Mohammad Abushariah and Majdi Sawalha. 2013. The effects of speakers' gender, age, and region on overall performance of Arabic automatic speech recognition systems using the phonetically rich and balanced Modern Standard Arabic speech corpus. In *Proceedings of the 2nd Workshop of Arabic Corpus Linguistics*, Lancaster, UK.

Laura Alonzo-Canul, Benjamin Lecouteux, and François Portet. 2025. Vers l'apprentissage de modèles auto-supervisés de reconnaissance automatique de la parole plus équitables sans a priori démographique.

Frederic Aman, Michel Vacher, Solange Rossato, and Francois Portet. 2013. Speech recognition of aged voice in the AAL context: Detection of distress sentences. In *2013 7th Conference on Speech Technology and Human - Computer Dialogue (SpeD)*, pages 1–8, Cluj-Napoca, Romania. IEEE.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus.

Giuseppe Attanasio, Beatrice Savoldi, Dennis Fucci, and Dirk Hovy. 2024. Twists, Humps, and Pebbles: Multilingual Speech Recognition Models Exhibit Gender Performance Gaps.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations.

Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias.

Lallouani Bouchakour and Mohamed Debyeche. 2022. Noise-robust speech recognition in mobile network based on convolution neural networks. *International Journal of Speech Technology*, 25(1):269–277.

Kimberle Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist policies. *University of Chicago Legal Forum*, 1:139–167.

Pranav Dheram, Murugesan Ramakrishnan, Anirudh Raju, I.-Fan Chen, Brian King, Katherine Powell, Melissa Saboowala, Karan Shetty, and Andreas Stolcke. 2022. Toward Fairness in Speech Recognition: Discovery and mitigation of performance disparities. In *Interspeech 2022*, pages 1268–1272.

Alex DiChristofano, Henry Shuster, Shefali Chandra, and Neal Patwari. 2023. Global Performance Disparities Between English-Language Accents in Automatic Speech Recognition.

Michael Joseph Dino, Carla Leinbach, Gerald Dino, Ladda Thiamwong, Chloe Margalaux Villafuerte, Mona Shattell, Justin Pimentel, Maybelle Anne Zamora, Anbel Bautista, John Paul Vitug, Joyline Chepkorir, and Nerceilyn Marave. 2025. Smart Speakers for Health and Well-Being of Older Adults: A Mixed-Methods Review. *Healthcare*, 13(21):2772.

Hend ElGhazaly, Bahman Mirheidari, Nafise Sadat Moosavi, and Heidi Christensen. 2025. Exploring Gender Disparities in Automatic Speech Recognition Technology.

Siyuan Feng, Bence Mark Halpern, Olya Kudina, and Odette Scharenborg. 2024. Towards inclusive automatic speech recognition. *Computer Speech & Language*, 84:101567.

Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. 2021. Quantifying Bias in Automatic Speech Recognition.

James Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2019. An Intersectional Definition of Fairness.

Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. 2018. A comparative study of fairness-enhancing interventions in machine learning.

Marcio Fuckner, Sophie Horsman, Pascal Wiggers, and Iskaj Janssen. 2023. Uncovering Bias in ASR Systems: Evaluating Wav2vec2 and Whisper for Dutch speakers. In *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 146–151, Bucharest, Romania. IEEE.

Mahault Garnerin, Solange Rossato, and Laurent Besacier. 2021. Investigating the Impact of Gender Representation in ASR Training Data: A Case Study on Librispeech. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 86–92, Online. Association for Computational Linguistics.

Shahram Ghorbani and John H. L. Hansen. 2018. Leveraging native language information for improved accented speech recognition. In *Interspeech 2018*, pages 2449–2453.

Katherine C. Hustad, Tristan J. Mahr, Phoebe Natzke, and Paul J. Rathouz. 2021. Speech Development Between 30 and 119 Months in Typical Children I: Intelligibility Growth Curves for Single-Word and Multiword Productions. *Journal of Speech, Language, and Hearing Research*, 64(10):3707–3719.

Wiebke Toussaint Hutiri and Aaron Ding. 2022. Bias in Automated Speaker Recognition. In *2022 ACM Conference on Fairness Accountability and Transparency*, pages 230–247.

Jongsuk Kim, Jaemyung Yu, Minchan Kwon, and Junmo Kim. 2025. FairASR: Fair Audio Contrastive Learning for Automatic Speech Recognition.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Ajinkya Kulkarni, Atharva Kulkarni, Miguel Couceiro, and Isabel Trancoso. 2024. Unveiling Biases while Embracing Sustainability: Assessing the Dual Challenges of Automatic Speech Recognition Systems. In *Interspeech 2024*, pages 4628–4632.

Linguistic Data Consortium. 2013. CABank English CallHome Corpus.

Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit.

Min Ma, Yuma Koizumi, Shigeki Karita, Heiga Zen, Jason Riesa, Haruko Ishikawa, and Michiel Bacchiani. 2024. FLEURS-R: A Restored Multilingual Speech Corpus for Generation Tasks.

Margot Masson and Julie Carson-berndsen. 2023. Investigating Phoneme Similarity with Artificially Accented Speech. In *Proceedings of the 20th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 49–57, Toronto, Canada. Association for Computational Linguistics.

Yen Meng, Yi-Hui Chou, Andy T. Liu, and Hung-yi Lee. 2022. Don't speak too fast: The impact of data bias on self-supervised speech models.

Josh Meyer, Lindy Rauchenstein, Joshua D. Eisenberg, and Nicholas Howell. 2020. Artie Bias Corpus: An Open Dataset for Detecting Demographic Bias in Speech Applications. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6462–6468, Marseille, France. European Language Resources Association.

Laureano Moro-Velázquez, Jaejin Cho, Shinji Watanabe, Mark Hasegawa-Johnson, Odette Scharenborg, Heejin Kim, and Najim Dehak. 2019. Study of the Performance of Automatic Speech Recognition Systems in Speakers with Parkinson's Disease. pages 3875–3879.

Md Nayeem, Md Shamse Tabrej, Kabbojit Jit Deb, Shaonti Goswami, and Md Azizul Hakim. 2025. Automatic Speech Recognition in the Modern Era: Architectures, Training, and Evaluation.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Guilherme Dean Pelegrina, Miguel Couceiro, and Leonardo Tomazeli Duarte. 2023. A statistical approach to detect sensitive features in a group fairness setting.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision.

Ana Rodrigues, Rita Santos, Jorge Abreu, Pedro Beça, Pedro Almeida, and Sílvia Fernandes. 2019. Analyzing the performance of ASR systems: The effects of noise, distance to the device, age and gender. In *Proceedings of the XX International Conference on Human Computer Interaction*, Interacción '19, pages 1–8, New York, NY, USA. Association for Computing Machinery.

Sandra Rojas, Elaina Kefalianos, and Adam Vogel. 2020. How Does Our Voice Change as We Age? A Systematic Review and Meta-Analysis of Acoustic and Perceptual Voice Data From Healthy Adults Over 50 Years of Age. *Journal of Speech, Language, and Hearing Research*, 63(2):533–551.

Samir Sadok, Simon Leglaive, Laurent Girin, Gaël Richard, and Xavier Alameda-Pineda. 2025. AnCoGen: Analysis, Control and Generation of Speech with a Masked Autoencoder.

Chloe Sekkat, Fanny Leroy, Salima Mdhaffar, Blake Perry Smith, Yannick Estève, Joseph Dureau, and Alice Coucke. 2024. Sonos Voice Control Bias Assessment Dataset: A Methodology for Demographic Bias Assessment in Voice Assistants. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15056–15075, Torino, Italia. ELRA and ICCL.

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 59–68, New York, NY, USA. Association for Computing Machinery.

Rachael Tatman. 2017. Gender and Dialect Bias in YouTube's Automatic Captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

Rachael Tatman and Conner Kasten. 2017. Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions. In *Interspeech 2017*, pages 934–938. ISCA.

Irina-Elena Veliche, Zhuangqun Huang, Vineeth Ayyat Kochaniyan, Fuchun Peng, Ozlem Kalinli, and Michael L. Seltzer. 2024. Towards measuring fairness in speech recognition: Fair-Speech dataset.

Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, pages 1–7, New York, NY, USA. Association for Computing Machinery.

Rebecca Wald, Jessica Taylor Piotrowski, Johanna M.F. Van Oosten, and Theo Araujo. 2024. Who are the (Non-)Adopters of Smart Speakers? A Cross-Sectional Survey Study of Dutch Families. *Tijdschrift voor Communicatiewetenschap*, 52(1):4–28.

Angelina Wang, Vikram V. Ramaswamy, and Olga Russakovsky. 2022. Towards Intersectionality in Machine Learning: Including More Identities, Handling Underrepresentation, and Performing Evaluation. In *2022 ACM Conference on Fairness Accountability and Transparency*, pages 336–349.

Wenbin Wang, Yang Song, and Sanjay Jha. 2024. GLOBE: A High-quality English Corpus with Global Accents for Zero-shot Speaker Adaptive Text-to-Speech. In *Interspeech 2024*, pages 1365–1369. ISCA.

Rumei Yang, Shiying Gao, and Yun Jiang. 2024. Digital divide as a determinant of health in the U.S. older adults: Prevalence, trends, and risk factors. *BMC Geriatrics*, 24(1):1027.

Yuanyuan Zhang, Yixuan Zhang, Bence Halpern, Tanvina Patel, and Odette Scharenborg. 2022. Mitigating bias against non-native accents. In *Proc. Interspeech 2022*, pages 3168–3172.

9. Language Resource References

Veliche, Irina-Elena and Huang, Zhuangqun and Kochaniyan, Vineeth Ayyat and Peng, Fuchun and Kalinli, Ozlem and Seltzer, Michael L. 2024. *Towards Measuring Fairness in Speech Recognition: Fair-Speech Dataset*. arXiv, arXiv:2408.12734. PID https://ai.meta.com/datasets/speech-fairness-dataset/.

A. Regression with constituent DV's

Another technique which has been used in the literature to estimate the effect of individual DV's on ASR performance is fitting a regression to predict ASR system's error on each speaker based on their constituent DV's (Sekkat et al., 2024; Tatman, 2017; DiChristofano et al., 2023). As in the previous case, it is imperative to fit this regression based on mean speaker performance rather than overall utterances in order to avoid biasing it towards individual speakers. The simplest form, regarding only univariate models and assuming categorical SG's, takes the following form:

$$\text{WER avg.}(D) = \sum_{DV^i} \sum_{SG \in DV^i} \alpha_{SG} \cdot \mathbb{1}_{(S=SG)} \tag{7}$$

where we expand DV^i to include the everything-SG (to simulate a bias term α_0). Sekkat et al. (2024) then goes on to define a multivariate model, which takes the intersection of SG's into account:

$$\text{WER avg.}(D) = \left(\sum_{DV^j} \cdots \sum_{DV^k} \sum_{SG_j \in DV^j} \cdots \sum_{SG_k \in DV^k} \right) \alpha_{SG_j,\ldots,SG_k} \cdot \mathbb{1}_{(S=SG_j,\ldots,SG_k)} \tag{8}$$

which is the semantic equivalent of Section 5.4. However, we consider relative SG-level WER to be a more intuitive measure; therefore, we focus on this for the remainder of the study.

Addressing Accent Disparities in Automatic Speech Recognition: A Comparative Study of Single and Two-Step Adaptation

Mykhailo Danilevskyi, Fernando Perez-Tellez, Jelena Vasic
Technological University Dublin
Blessington Rd, Dublin, D24 FKT9, Ireland
D22126578@mytudublin.ie, {Fernando.PerezTellez,Jelena.Vasic}@TUDublin.ie

Abstract

Automatic speech recognition (ASR) systems often exhibit uneven performance across accents, raising concerns about fairness and bias. This study investigates the impact of model fine-tuning strategies on ASR performance and accent-related disparities. We conduct a controlled empirical evaluation of two adaptation approaches—single-step and two-step fine-tuning—using pretrained Whisper (small) and Wav2Vec2-XLSR-53 models on African-accented English speech from the AfriSpeech-200 dataset, covering Yoruba, Igbo, Swahili, and Hausa accents. Both fine-tuning strategies substantially reduced mean word error rate (WER) for all models. However, these improvements did not translate into consistent reductions in accent-related performance gaps. When analysed separately across general and clinical subsets, WER gaps often increased due to uneven gains across accents. Although two-step fine-tuning provided modest improvements over single-step adaptation, its impact on reducing disparities remained limited. These findings indicate that fine-tuning primarily optimises performance without effectively addressing systematic bias across speaker groups, even when models are specialised for individual accents. This highlights the limitations of per-accent specialisation as a practical bias mitigation strategy.

Keywords: automatic speech recognition, accent disparities, fairness in AI; bias in speech recognition, fine-tuning

1. Introduction

Automatic Speech Recognition (ASR) performance has improved significantly in recent years, largely driven by self-supervised acoustic models such as Wav2Vec2 (Baevski et al., 2020; Conneau et al., 2020) and large-scale multilingual transformer models such as Whisper (Radford et al., 2022). The current level of ASR performance has led to its widespread acceptance and deployment across industries, accelerating the adoption of transcription technologies in healthcare, customer service, education, and other domains.

Despite strong performance on general English speech, ASR systems remain sensitive to non-native accents and domain-specific terminology, resulting in persistent performance disparities across speaker groups (Veliche et al., 2024; Liu et al., 2021). These disparities raise important ethical concerns related to fairness, equal access, and equitable distribution of technological benefits.

Unequal ASR performance across accents can result in systematically higher word error rates (WER) for speakers of non-dominant language varieties. Such discrepancies risk marginalising already underrepresented groups and may undermine trust in AI-driven systems. As ASR increasingly replaces or supplements human transcription, performance gaps can have tangible social and professional implications. Therefore, addressing accent-based disparities is not only a technical challenge but also a matter of linguistic equity. Reflecting this growing concern, research on accent bias in ASR continues to expand, with numerous studies published each year (Prinos et al., 2024).

The contributions of this study are:

- Analysed accent-related model bias by comparing WER and WER gaps under single-step and two-step fine-tuning strategies.
- Investigated whether additional fine-tuning and accent-specific models mitigate performance disparities across accent pairs.
- Provided a systematic comparison of how Whisper and Wav2Vec2-XLSR-53 models exhibit disparities across accents.

The remainder of this paper is organised as follows. Section 2 reviews related work. Section 3 presents the methodology, including the dataset, evaluation metrics, models, and fine-tuning procedures. Section 4 presents the experimental results. Section 5 discusses the findings, and Section 6 concludes the paper.

2. Related work

Prior studies have proposed a range of strategies for mitigating accent-related bias in ASR. These can be broadly classified into accent-agnostic and accent-specific approaches. Accent-agnostic methods typically rely on large-scale training with

Proceedings of SPEAKABLE @ LREC 2026, pages 79–85
11 May 2026.

adversarial bias mitigation to encourage accent-invariant representations. In contrast, accent-specific approaches incorporate explicit accent labels or leverage architectures such as mixture-of-experts (MoE), in which specialised submodules handle specific accents (Bagat et al., 2025; Lee et al., 2026). However, MoE-based solutions require accurate routing between experts, which can be overly complex in scenarios with a limited number of accents and ASR system users.

In these constrained settings, simpler adaptation strategies may prove to be more practical. Following the ASR model adaptation presented in (Meyer et al., 2020; Pillai et al., 2024), we explore a two-step fine-tuning procedure to improve ASR performance in accent-aware contexts without requiring routing or explicit accent classification at inference time. In this approach, the pretrained model is first adapted to the entire dataset and then further fine-tuned on each target accent individually. In (Meyer et al., 2020), this strategy was applied with the DeepSpeech model (Hannun et al., 2014), fine-tuning first on the full Common Voice dataset and then on each demographic group, yielding five models. The authors (Pillai et al., 2024) used a similar two-step strategy to adapt Whisper from Tamil to Malasar, leveraging the linguistic similarities between resource-rich and low-resource languages.

While recent work has primarily focused on improving model performance for specific accents or languages, this paper instead examined accent-related bias and investigated whether fine-tuning on accent-specific data can reduce WER disparities. In contrast to (Torgbi et al., 2025), which evaluated Whisper on spontaneous telephone speech from native English speakers across two Scottish accents, our work analysed both Whisper and Wav2Vec2-XLSR-53 models on non-native English speakers reading scripted text, providing a different linguistic and acoustic setting. Furthermore, we investigated whether additional fine-tuning on a specific accent improves performance for that accent while also reducing performance disparities across accents. Although (Özyilmaz et al., 2025) adopted a similar fine-tuning strategy, its primary focus remained on improving performance for Arabic dialects rather than measuring bias across groups. In addition to the original Afrispeech-200 paper (Olatunji et al., 2023), which introduced the dataset and evaluated several state-of-the-art models, including Whisper and Wav2Vec2-XLSR-53, we extended this work by examining the impact of further model fine-tuning on both performance and bias. Specifically, we assessed overall model bias and evaluated whether observed differences between accents were statistically significant.

In this paper, we adopted a two-step fine-tuning approach on the Afrispeech dataset, first fine-tuning pretrained Whisper and Wav2Vec2-XSLR-53 models on the full dataset, followed by accent-specific fine-tuning. We compared this with single-step fine-tuning on the full dataset to evaluate overall performance and accent disparities.

3. Methodology

3.1. Data

Our experiments were conducted on the Afrispeech-200 dataset (Olatunji et al., 2023), which contains audio recordings, transcripts, and accent labels covering 120 African accents from 13 countries and 2,463 unique speakers, covering both general and clinical domains. The data were originally partitioned into training, development, and test splits with no speaker overlap across the splits[1]. Its composition is summarised in Table 1, with the proportion of clinical data indicated in brackets.

Table 1: Speaker counts and total duration (minutes) per accent across train, development, and test splits. Clinical-domain data proportions are indicated in brackets.

Accent	Speakers			Duration (min)		
	Train	Dev	Test	Train	Dev	Test
Yoruba	454 (51%)	57 (53%)	172 (47%)	2527 (62%)	55 (53%)	107 (36%)
Igbo	246 (64%)	34 (50%)	94 (56%)	1457 (70%)	35 (50%)	55 (53%)
Swahili	46 (59%)	12 (58%)	61 (44%)	805 (71%)	47 (53%)	80 (44%)
Hausa	176 (80%)	19 (74%)	53 (74%)	1125 (85%)	15 (66%)	38 (78%)
Others	544 (51%)	125 (60%)	370 (60%)	4459 (55%)	371 (55%)	841 (55%)
Total	**1466** (57%)	**247** (58%)	**750** (56%)	**10373** (63%)	**523** (54%)	**1121** (56%)

For performance and bias analysis, we selected the four most represented accents in the dataset: Yoruba, Hausa, Swahili, and Igbo, which we refer to as the target accents. Model fine-tuning was performed using the training and development splits of the dataset, whereas evaluation was conducted on the test split of the target accents.

The duration of speech per speaker varied across splits and accents. On average, speakers in the training split contributed 7.1 min of audio, compared to 2.1 and 1.5 min in the development and test splits, respectively. For the target accents, the average audio duration in the training split was approximately 6 min per speaker; however, Swahili

[1] https://huggingface.co/datasets/intronhealth/afrispeech-200

exhibited substantially longer recordings, with an average of 17.5 min per speaker. The data were unevenly distributed across the general and clinical domains, with Hausa exhibiting the highest proportion of clinical audio recordings (~80%) and Yoruba the lowest (36% in the test split).

All audio files were resampled from 44.1 kHz to 16 kHz to be consistent with the model requirements. To ensure correct comparison across models, both reference transcripts and model predictions were normalised using the Hugging Face Whisper basic normaliser, followed by punctuation removal.

3.2. Evaluation Metrics

The transcription accuracy in all experiments was measured using the word error rate (WER) metric.

To estimate the overall model bias in the ASR system, we used the relative *WERgap* metric, defined as the difference between the maximum and minimum WER across groups (e.g. accents), normalised by the maximum WER:

$$\text{WER}_{\text{gap}} = \frac{\text{WER}_{\text{max}} - \text{WER}_{\text{min}}}{\text{WER}_{\text{max}}} \quad (1)$$

For pairwise disparity evaluation between accents, we employed the methodology proposed by Liu et al. (2021) using the following test statistic:

$$\theta_{i,j} = \frac{\text{WER}_i}{\text{WER}_j} - 1 \quad (2)$$

where $\theta_{i,j}$ represents the bias between the two groups, and WER_i and WER_j are the word error rates for groups i and j, respectively. Statistical significance was assessed using the bootstrap method (Marriott et al., 1995; Bisani and Ney, 2004) with 95% percentile confidence intervals (CI). If the CI $[\text{ln}, \text{un}]$ does not include 0, the WER difference between the two groups is considered statistically significant. The bootstrap method was also used to estimate whether the difference in WER between fine-tuning methods is statistically significant.

3.3. Models

The experiments were conducted using the pretrained Whisper small model with 244M parameters (Radford et al., 2022) and the Wav2Vec2-XLSR-53 model with 317M parameters (Conneau et al., 2020). Whisper is a transformer-based encoder–decoder architecture, commonly described as a sequence-to-sequence model. It was trained on approximately 680,000h of labelled speech data annotated through large-scale weak supervision. Wav2Vec2-XLSR-53 model builds on the Wav2Vec 2.0 architecture (Baevski et al., 2020) by extending it to a cross-lingual setting. Wav2Vec2-XLSR-53 is an encoder-only architecture that predicts output tokens using connectionist temporal classification (CTC) (Graves et al., 2006).

For this study, we used pretrained models available on Hugging Face: wav2vec2-large-xlsr-53-english[2] and whisper-small multilingual model [3]. Both models are end-to-end architectures and were pretrained on multilingual data, which have been shown to be beneficial for improving ASR performance on accented speech (Vu et al., 2014).

3.4. Experimental method

We evaluated two ASR model adaptation approaches using Whisper (small) and Wav2Vec2-XLSR-53: (i) single-step fine-tuning on the full dataset and (ii) a two-step fine-tuning consisting of accent adaptation on the entire dataset followed by accent-specific fine-tuning yielding specialised models for each accent (Figure1). The two-step approach separates general accent adaptation and accent-specific specialisation. This allows the model to better capture acoustic and linguistic characteristics of each accent and may help reduce performance disparities across accents.

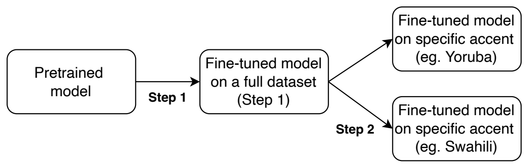


Figure 1: Two-step fine-tuning

For each architecture, we fine-tuned a pretrained baseline model on the entire dataset to implement the single-step approach, selecting the best model based on the lowest word error rate (WER). This approach yields a single model specialised for all accents in the dataset. For the two-step approach, we first fine-tuned the model on the entire dataset, selecting the best checkpoint based on minimum validation loss. In the second step, we continued the adaptation separately for each target accent, selecting the best accent-specific model using WER as the evaluation criterion. This strategy results in one specialised model per accent.

Whisper models were trained for 10 epochs using a linear learning rate scheduler with a 0.025 warm-up ratio, weight decay of 0.03 and effective batch size of 32 for both adaptation strategies. The learning rate was set to 1e-5 for adaptation on the entire dataset and reduced to 1e-7 for accent-specific fine-tuning. Wav2Vec2-XLSR-53 models were trained for 20 epochs with a linear scheduler and a 0.1

[2]https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english

[3]https://huggingface.co/openai/whisper-small

warm-up ratio for both strategies, a learning rate of 1e-5 and effective batch size of 32. All models were fine-tuned with the encoder unfrozen to allow maximal adaptation to the acoustic and linguistic characteristics of each accent. Early stopping was applied to terminate training once performance plateaued, preventing overfitting. These hyperparameters were selected based on prior literature (Olatunji et al., 2023; Baevski et al., 2020; Bagat et al., 2025) and preliminary experiments, providing stable and reasonable performance within our computational constraints.

4. Results

Table 2 reports WER results for the pretrained baselines and their single-step (FT1) and two-step fine-tuned (FT2) variants of Whisper (small) and Wav2Vec2-XLSR-53 across four African accents: Yoruba, Igbo, Swahili, and Hausa. We assess whether additional fine-tuning and accent-specific models reduce WER and cross-accent differences, analysing (i) within-model WER changes and (ii) pairwise variation across accents.

The pretrained Whisper (small) model achieved a mean WER of 0.446 across the four target accents, with an absolute WER gap of 0.420. Both fine-tuning strategies led to substantial reductions in mean WER, reaching 0.189 and 0.183 for FT1 and FT2, respectively. Across all configurations, Swahili consistently exhibited the lowest overall WER. In contrast, the highest WER was observed for Hausa in the clinical subset, while the lowest within Hausa occurred in the general subset. The overall WER gaps remained relatively stable across all model variants (0.404–0.425). However, when analysed separately by domain, the gaps increased after fine-tuning. In the general subset, the gaps were primarily driven by differences between Yoruba and Hausa, whereas in the clinical subset they were driven by Hausa and Swahili. Although fine-tuning increased the WER gaps in both subsets compared to the baseline, FT2 resulted in a lower gaps than FT1. Overall, FT2 achieved both lower WER and smaller WER gaps than FT1 across domains.

Wav2Vec2-XLSR-53 baseline model achieved a mean WER of 0.631 across the target accents with an absolute WER gap of 0.384. Both fine-tuning strategies substantially reduced the mean WER, to 0.270 (FT1) and 0.259 (FT2). Swahili consistently exhibited the lowest WER across all model variants, while Hausa showed the highest WER overall, except in the general subset where its WER was lower than that of Swahili. FT2 achieved lower WER than FT1 across all accents and domains, with statistically significant differences observed for Yoruba and Swahili. The overall WER gaps ranged from 0.318 to 0.384 and was primarily driven by differences between Hausa and Swahili. FT1 reduced the absolute WER gap by approximately 0.06 (from 0.384 to 0.318), whereas FT2 reduced it to 0.346. This reduction was primarily driven by improvements in the clinical domain, whereas the gap increased in the general domain due to lower WER for Hausa. Overall, FT2 achieved lower WER across accents and a smaller WER gap in the general subset.

Table 2: WER comparison across accents (statistically significant difference between FT2 and FT1 within accents and models are in bold).

Accent	Base	FT1	FT2 Step 1	FT2 Step 2
Whisper (small)				
Yoruba	0.523	0.226	0.231	0.223
general	0.496	0.229	0.235	0.221
clinical	0.562	0.221	0.226	0.227
Igbo	0.472	0.174	0.198	0.178
general	0.389	0.137	0.137	0.138
clinical	0.541	0.205	0.249	0.211
Swahili	0.303	0.141	0.133	**0.132**
general	0.289	0.138	0.132	0.131
clinical	0.323	0.147	0.134	**0.135**
Hausa	0.520	0.239	0.224	0.219
general	0.351	0.089	0.097	0.096
clinical	0.582	0.293	0.269	0.263
WER mean	0.446	0.189	0.191	0.183
general	0.399	0.169	0.169	0.164
clinical	0.496	0.212	0.216	0.205
WER gap	0.420	0.407	0.425	0.408
general	0.417	0.611	0.587	0.566
clinical	0.445	0.498	0.502	0.487
Wav2Vec2-XLSR-53				
Yoruba	0.676	0.298	0.296	**0.281**
general	0.563	0.280	0.274	**0.260**
clinical	0.835	0.323	0.327	0.309
Igbo	0.640	0.260	0.250	0.256
general	0.535	0.215	0.209	0.212
clinical	0.728	0.298	0.285	0.293
Swahili	0.501	0.228	0.226	**0.216**
general	0.433	0.204	0.205	**0.193**
clinical	0.590	0.261	0.253	**0.246**
Hausa	0.813	0.335	0.333	0.331
general	0.426	0.175	0.183	0.181
clinical	0.954	0.392	0.386	0.385
WER mean	0.631	0.270	0.267	0.259
general	0.507	0.234	0.231	0.221
clinical	0.765	0.312	0.309	0.302
WER gap	0.384	0.318	0.321	0.346
general	0.243	0.375	0.332	0.304
clinical	0.382	0.334	0.345	0.361

The results indicate that fine-tuning primarily improved overall recognition accuracy, but it exacerbated performance disparities due to uneven gains across accents and data domains. Although

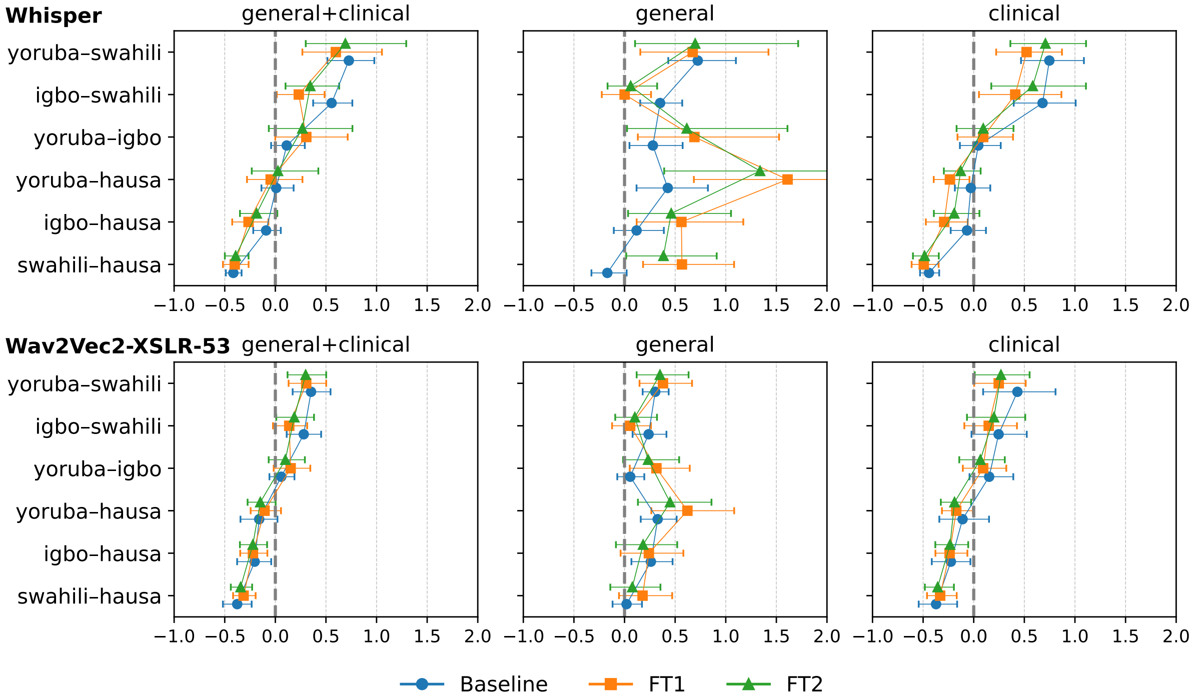


Figure 2: Pairwise WER comparison of relative difference using test statistic θ (Eq. 2). The x-axis was limited to [-1, 2] to enhonce results visibility, truncating the upper confidence intervals for the Yoruba-Hausa pair, which reach 2.98 and 2.96 for FT1 and FT2, respectively.

Whisper achieved higher overall performance than Wav2Vec2-XSLR-53, WER gaps remained larger across all model configurations, and both fine-tuning approaches even amplified these gaps in the general and clinical data subsets. The performance of Wav2Vec2-XSLR-53 is less varying across accents and data domains, which reflects on further pairwise analysis.

Pairwise WER measurements between accents using the test statistic θ (Eq.2) presented on Figure 2. For the Whisper model, bias patterns differed across domains. In the combined (general + clinical) setting, all pairs involving Swahili exhibited statistically significant differences at the baseline; these were reduced by both FT1 and FT2. In the general domain subset, baseline differences were observed for all pairs involving Yoruba, as well as for Igbo–Swahili. Both fine-tuning approaches resolved only the Igbo–Swahili difference, while increasing gaps for pairs involving Hausa; this increase was slightly smaller for FT2. In the clinical domain, all three pairs involving Swahili showed statistically significant differences at the baseline. These were not resolved by either fine-tuning approach. FT1 increased gaps for all pairs involving Hausa to statistically significant levels, while both methods slightly reduced differences for pairs involving Swahili. Overall, improvements in Swahili-related disparities appear to come at the cost of increased disparities for Hausa, highlighting a trade-off in bias mitigation across accents.

For the Wav2Vec2-XLSR-53 model, bias patterns also varied across domains. In the combined (general + clinical) setting, four pairs showed significant differences at baseline, including three involving Swahili and the Igbo–Hausa pair. Both FT1 and FT2 produced similar outcomes, primarily reducing differences for pairs involving Swahili. In the general domain, four pairs involving Swahili and Hausa showed significant differences at baseline. FT1 resolved differences for pairs involving Igbo (with Swahili and Hausa), but introduced a significant difference for the Yoruba–Igbo pair. In contrast, FT2 resolved two differences, both involving Igbo, without introducing new effects; for most pairs, bias gaps moved closer to zero. In the clinical domain, three pairs showed significant differences at baseline, including two involving Hausa (paired with Igbo and Swahili) and one Yoruba–Swahili pair. Both fine-tuning approaches slightly reduced bias gaps; however, these differences remained for all three pairs.

In conclusion, Whisper exhibited larger absolute performance gaps between certain accents, particularly involving Swahili, yet fewer statistically significant pairwise disparities overall. In contrast, Wav2Vec2-XLSR-53 demonstrated a smaller overall WER gaps and more consistently distributed disparities across accent pairs. These results indicate that Whisper's bias is more concentrated around specific accent, whereas Wav2Vec2-XLSR-53 exhibits more distributed differences across groups.

5. Discussion

This study investigates single-step (FT1) and two-step (FT2) fine-tuning approaches, focusing on their impact on reducing accent-related performance disparities in ASR models. The results show that for Whisper and Wav2Vec2-XSLR-53 models both fine-tuning approaches significantly outperform baseline and FT2 demonstrated lower WER in most cases. In contrast, WER gaps increased after fine-tuning, though after FT2 they were lower than after FT1. The main driver of WER gaps was Swahili accent with the lowest WER. Notably, this occurs despite Swahili having the smallest number of speakers in the training split. However, Swahili recordings contain substantially longer speech segments per speaker (approximately 17 min per speaker), whereas the other accents average around 6 min per speaker. Hausa is another accent with lowest WER for general subset that drove high gaps in the general subset, but this can be driven by the limited test subset containing only 8 min of audio.

Whisper model achieved lower WER but exhibited larger WER gaps compared to Wav2Vec2-XLSR-53. Pairwise analysis highlights distinct bias patterns between the models, likely arising from differences in architecture and training data. These findings suggest that Whisper's sequence-to-sequence architecture, together with its autoregressive decoding, may contribute to accent-specific biases. By contrast, Wav2Vec2's cross-lingual pretraining and CTC-based design appears to yield more consistent performance across accents.

The study is constrained by the limited size and uneven distribution of accent-specific data, particularly for Hausa, where some subsets contained as little as 8min of recordings compared to 17–68min for other accents. This imbalance likely contributed to amplified gaps in general domains after fine-tuning which requires further investigation. Additionally, our analysis focused on only four African accents and two ASR models; conclusions may not generalize to other languages and accents. Finally, while the two-step fine-tuning approach improves WER, it may not fully address systematic bias arising from pretraining data composition or domain-specific acoustic variability. Future work should explore balanced data augmentation, multi-task objectives, or bias-aware loss functions to improve both accuracy and fairness. Further analysis of domain-specific acoustic characteristics may also help in designing more equitable ASR systems for clinical and general speech.

6. Conclusion

In this paper, we investigated the effectiveness of single-step and two-step fine-tuning strategies for reducing accent-related performance differences in automatic speech recognition models across four African accents using Whisper (small) and Wav2Vec2-XLSR-53. The results demonstrated that fine-tuning significantly improved overall WER for both models; however, its impact on reducing accent-dependent disparities remained limited. Performance gaps across accents persisted under all examined fine-tuning approaches. Although two-step fine-tuning yielded incremental improvements over single-step adaptation, these gains were modest and unevenly distributed across accents - Hausa consistently exhibited the highest WER, whereas Swahili achieved the lowest.

More broadly, this study demonstrates that optimisation of overall performance does not necessarily lead to improved parity across groups. Therefore, mitigating accent-related disparities may require reweighting strategies, adversarial invariance training, or fairness-regularised optimisation. However, a two-step fine-tuning approach that yields one model per speaker group may be beneficial in settings with a limited number of users, such as medical laboratories. In such constrained environments, more complex approaches, such as mixture-of-experts architectures or adversarial debiasing, may be impractical. Consequently, two-step fine-tuning can serve as a method for improving performance in accent-specific deployments, although its bias mitigation capabilities are limited.

7. References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. ArXiv:2006.11477 [cs].

Raphaël Bagat, Irina Illina, and Emmanuel Vincent. 2025. Mixture of LoRA Experts for Low-Resourced Multi-Accent Automatic Speech Recognition. In *Interspeech 2025*, pages 1143–1147. ArXiv:2505.20006 [cs].

M. Bisani and H. Ney. 2004. Bootstrap estimates for confidence intervals in ASR performance evaluation. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–409–12, Montreal, Que., Canada. IEEE.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael

Auli. 2020. Unsupervised Cross-lingual Representation Learning for Speech Recognition. ArXiv:2006.13979 [cs].

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, New York, NY, USA. Association for Computing Machinery.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. 2014. Deep speech: Scaling up end-to-end speech recognition.

Wonjun Lee, Hyounghun Kim, and Gary Geunbae Lee. 2026. Mixture-of-Experts with Intermediate CTC Supervision for Accented Speech Recognition. ArXiv:2602.01967 [cs].

Chunxi Liu, Michael Picheny, Leda Sarı, Pooja Chitkara, Alex Xiao, Xiaohui Zhang, Mark Chou, Andres Alvarado, Caner Hazirbas, and Yatharth Saraf. 2021. Towards Measuring Fairness in Speech Recognition: Casual Conversations Dataset Transcriptions. ArXiv:2111.09983 [eess].

Paul Marriott, B. Efron, and R. J. Tibshirani. 1995. An Introduction to the Bootstrap. In *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, volume 158, page 347. ISSN: 09641998 Issue: 2 Journal Abbreviation: Journal of the Royal Statistical Society. Series A (Statistics in Society).

Josh Meyer, Lindy Rauchenstein, Joshua D. Eisenberg, and Nicholas Howell. 2020. Artie Bias Corpus: An Open Dataset for Detecting Demographic Bias in Speech Applications. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6462–6468, Marseille, France. European Language Resources Association.

Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure F. P. Dossou, Joanne I. Osuchukwu, Salomey Osei, A. Tonja, Naome A. Etori, and Clinton Mbataku. 2023. AfriSpeech-200: Pan-African Accented Speech Dataset for Clinical and General Domain ASR. *Transactions of the Association for Computational Linguistics*, 11:1669–1685.

Leena G. Pillai, Kavya Manohar, Basil K. Raju, and Elizabeth Sherly. 2024. Multistage Fine-tuning Strategies for Automatic Speech Recognition in Low-resource Languages. ArXiv:2411.04573 [cs].

Kerri Prinos, Neal Patwari, and Cathleen A. Power. 2024. Speaking of accent: A content analysis of accent misconceptions in ASR research. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 1245–1254, New York, NY, USA. Association for Computing Machinery.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. ArXiv:2212.04356.

Melissa Torgbi, Andrew Clayman, Jordan J. Speight, and Harish Tayyar Madabushi. 2025. Adapting whisper for regional dialects: Enhancing public services for vulnerable populations in the united kingdom.

Irina-Elena Veliche, Zhuangqun Huang, Vineeth Ayyat Kochaniyan, Fuchun Peng, Ozlem Kalinli, and Michael L. Seltzer. 2024. Towards measuring fairness in speech recognition: Fair-Speech dataset. ArXiv:2408.12734 [cs].

Ngoc Thang Vu, Yuanfan Wang, Marten Klose, Zlatka Mihaylova, and Tanja Schultz. 2014. Improving ASR performance on non-native speech using multilingual and crosslingual information. In *15th Annual Conference of the International Speech Communication Association, INTERSPEECH 2014, Singapore, September 14-18, 2014*, pages 11–15. ISCA.

Ömer Özyilmaz, Matt Coler, and Matias Valdenegro. 2025. Overcoming data scarcity in multi-dialectal arabic asr via whisper fine-tuning. pages 1158–1162.

Investigating speaker pronunciation variability in speech embeddings: speaker and L1 effects on French as a Second Language

Maxime Fily, Martine Adda-Decker, Guillaume Wisniewski
INALCO, Université Sorbonne Nouvelle, Université Paris Cité
2 rue de Lille 75007 Paris, 4 rue des Irlandais 75005 Paris, 8 rue Albert Einstein 75013 Paris
maxime.fily@inalco.fr, martine.adda-decker@sorbonne-nouvelle.fr, guillaume.wisniewski@u-paris.fr

Abstract

Speech variation between native and non-native speakers of French is addressed with a low-resource method based on a frame-wise comparison of wav2vec2 acoustic embeddings, using fine-grained phonetic transcriptions by expert annotators as baseline. z-normalisation and t-normalisation are explored to assess what the embeddings contain in terms of phonetically analysable information. We explore non-supervised methods for solving basic speech-related research questions. Adapting Dynamic Time Warping to speech embeddings, we compare phonologically similar recordings of sentences read-aloud by native vs. non-native speakers of French. The question is whether XLSR-53 embeddings are more robust than MFCCs to inter-speaker vs. intra-speaker variability for different occurrences of the same words. Then we investigate whether native speaker productions are more or less stable than those of non-native speakers. Results suggest that the model allows phonetically meaningful correlative analyses. Working on the raw embeddings shows however that the representations are not speaker-independent, so with a view to address issues in relationship with L2 pronunciation variability, we show that t-normalisation brings us a way to separate fluency and accuracy effects in L2-speech. This shows that wav2vec2 encapsulates time-dependent phonetic information in the embeddings, including speaker accent which can not easily be disentangled from other speaker-specific characteristics.

Keywords: Cosine similarities, speech, L2-acquisition, Query-by-Example Spoken Term Detection, unsupervised method

1. Introduction

Despite the extensive instrumental and analytical resource available in phonetics, studies of pronunciation acquisition often still depend on impressionistic criteria, including native-likeness, accentedness, and intelligibility judgments (Saito et al., 2016). It is as if the immense variability inherent to foreign-accented speech[1] led researchers to resort only to perceptual criteria with native judges. It is argued in Machine-Learning ASR that a mere increase in the amount of training data “dilutes” speaker characteristics, thereby offering better robustness to variability, but recent findings (Zee et al., 2024) claim, on the contrary, that without attention to corpus balance, more training data aggravates biases because the proportion of data from dominant categories increases more than from *minorised* groups. Large Audio Models (LAM) therefore encapsulate more bias in the representations, especially when the data is scraped across the internet rather than collected with controlled metadata and explicit consent (some evidence in Névéol et al., 2022; Gebru, 2019; Buolamwini and Gebru, 2018). For these groups, mass scraping results in less performant ASR, so is the case for people with a foreign accent (Khandelwal et al., 2020).

[1] driven by social factors, region, personal history. (See Moyer, 2013).

This study examines whether phonetic information is encoded in one multilingual (`XLSR-53`) and one monolingual (`wav2vec2-FR-7K-large`) `wav2vec2` (Baevski et al., 2020) model by comparing distances between embeddings generated from occurrences of the same words. With this method, we then move on to measuring pronunciation variability on read-aloud speech through the lens of (i) inter-speaker vs. intra-speaker variability and (ii) L2-speech vs native speech. For this part, we focus the analyses on `XLSR-53` (Conneau et al., 2021) to measure the sensitivity of a multilingual model to speaker characteristics and native language.

The method uses a corpus of read sentences in native-French and L2-French. We start by assessing the correlation between the variability observed on `wav2vec2` (Baevski et al., 2020) neural model embeddings and the associated MFCCs (Mean-Frequency Cepstral Coefficients, a direct measurement on the audio signal). This paper also investigates the amount of speaker-specific information present within `wav2vec2` speech embeddings in comparison with the MFCCs. Based on a Query-by-Example approach using Dynamic Time Warping on speech embeddings, speaker-specific information/pronunciation is evaluated by measuring how robust the alignment is when speaker ID and foreign accents vary.

Our contribution consists in an unsupervised

Proceedings of SPEAKABLE @ LREC 2026, pages 86–97

method to measure pronunciation variation using `wav2vec2` embeddings. It is tested on `XLSR-53` multilingual model.

2. Existing methods

This section lays out the relevant literature on foreign accents in L2-speech, in terms of experimental methods and NLP frameworks.

2.1. Experimental methods in foreign accent evaluation

Pronunciation learning theories, despite their early development (e.g., Flege, 1981, 1995), have been challenging to prove empirically and therefore remain rarely tested in L2 acquisition frameworks (Kennedy and Trofimovich, 2017), even less so with specific, objective approaches. Pronunciation acquisition evaluation frameworks need non-judgmental feedback to help learners' derive a critical but constructive view of their own productions (Suzukida, 2021). Among recent reviews of L2-acquisition research, (Derwing, 2008) still deplore that at this point, many of the publications have not yet exploited the potential benefits of new technologies, in particular in "measuring progress", which is still mostly perception-based (although a few publications in experimental phonetics outline the role of direct measurements in L2-acquisition improvement, e.g., for retroflex consonants (Bliss et al., 2018) in English or for palatalized consonants in Russian (Lecocq, 2021)). In summary, a lot of the feedback to learners basically amounts to "did I do good?" without necessarily knowing what "doing good" means or which aspects of pronunciation need to improve. In a recent study, (Saito et al., 2016) devised a range of measurable variables to evaluate what participates to native-likeness in L2 speech. Their results identify speech rate among first order parameters for nativelikeness, but they focused quasi-exclusively on lexical variables (lemma, morphology, polysemy). They did not include phonetic accuracy of the output among the parameters although several studies show that when rating native-likeness in L2 speech, expert listeners scores correlate with the character error rate calculated between the narrow transcription of the segment and its native-like phonological transcription (Munro, 2008).

2.2. NLP methods in foreign accent evaluation

Large Language Models, which train on massive datasets, "exhibit and amplify stereotypical bias" (Ducel et al., 2025), as illustrated in Tatman (2017) where different genders and accents do not transcribe equally well using the (proprietary) YouTube ASR system. Efforts have been made to offer less biased ASR systems for the minorised languages (Havard et al., 2025), or to remove biases in Large Language Models via adversarial training. In low-resource scenario cases where resources are too scarce for such advanced debiasing methods, evaluating WER values in an ASR task can help identify the root causes for the biases before potentially offering solutions (Feng et al., 2021; Zhang et al., 2022).

Bartelds et al. (2022) estimated Foreign accent on non-native American English (Weinberger and Kunath, 2011) and concluded on a correlation between LAM embeddings and a nativelikeness assessment by humans. By comparing their results with how Levenshtein Distance (LD) or MFCCs correlate with nativelikeness judgment, they obtain results which are on average similar to LD or MFCCs. Their study is based on evaluating similarities between segments based on the Dynamic Time Warping (DTW) approach (Giorgino, 2009). It establishes a link between alignment performance and nativelikeness judgment, regardless of how the alignment is done: on LAM embeddings, MFCCs, or LD.

In this study, we are interested in avoiding nativelikeness judgments by instead correlating embedding alignments with two different metrics: WER and MFCC alignment cost. Query-by-Example Spoken Term Detection (QbE-STD), a long-standing word retrieval method which typically searches the minimal alignment cost between an audio sample and an audio lexicon and by doing so retrieves the correct dictionary entry. It can be applied to either audio signal (Le Ferrand et al., 2020) or speech embeddings (San et al., 2021). QbE-STD has been used in document retrieval on large audio archives (Ram et al., 2020; Hazen et al., 2009), or as an alternative to fine-tuning for the retrieval of low-resource language utterances when there is insufficient data (typically less than 2h. See Guillaume et al., 2022) for a proper fine-tuning (San et al., 2021).

Among recent studies using frame-wise, time-dependent embeddings, we cite the example of how QbE-STD has been done on low-resource languages such as Kunwinjku, an Australian Aboriginal language and Mboshi, a Bantu Congo Brazaville language (Le Ferrand et al., 2020): the goal was to enable speakers of these oral languages to retrieve words from a dictionary. Among the acoustic features tested (including but not limited to MFCCs and fine-tuned `w2v2-en` embeddings, MFCCs gave out the best results, but were quite predictably very speaker-dependent. z-normalising the features (Lobanov, 1971; Xie and Jaeger, 2020) improved the recall values but the approach was only applied to MFCC coefficients since z-normalising did not have any visible effect on the embeddings.

3. Research objective

Our study explores whether or not phonetic/pronunciation information is represented in two `wav2vec2` models (`XLSR-53` and `wav2vec2-FR-7K-large`) and if so, how they encode this information. We compare several occurrences of the same target words by measuring distances between them in the models' representations space. This approach consists in (i) verifying how well audio and embeddings are correlated after forced alignment at word level, and then, focusing on `XLSR-53` exclusively, (ii) assessing `XLSR-53` sensitivity to basic linguistic questions. We first verify whether different occurrences of a same speaker align more or less easily than those of different speakers. Then, the effect of speaker L1 (French or Russian) is assessed via cross comparisons of the alignment cost. Effects of embeddings z-normalisation[2] and t-normalisation are also assessed, always with a view to reducing the biases due to speaker ID or non-native speech in corpora.

By controlling the recording conditions for all the participants and by applying to speech embeddings an approach derived from experimental phonetics methods in an NLP framework, we propose to compare distance measurements in the neural representations space and in the audio space. We are interested in understanding to what extent it allows us to draw conclusions on basic/intuitive research questions: can we measure inter - vs. intra-speaker variability from the embeddings and is native speech more stable than L2 speech? These questions are already addressed widely in the literature but not yet with a frame-wise dual-space approach. Our contribution ultimately aims at establishing a correlation between distance measurements on two signals and their discrepancies in terms of Character Error Rate.

4. Material and method

4.1. Variables

We are interested in how native speakers of Russian acquire French pronunciation. Our goal is to compare the same target words in French for two cohorts of participants: one *French-native* and one *Russian-native/L2-French learner* at the time of the experiment. we propose to assess the impact of the variables listed in Table 1.

[2]which,incidentally, reduces embedding anisotropy (Ethayarajh, 2019; Timkey and van Schijndel, 2021).

<table>
<tr><td>Experiment</td><td>Spk</td><td colspan="2">L1</td></tr>
<tr><td rowspan="3">Modality</td><td>$=$spk</td><td colspan="2">N/A</td></tr>
<tr><td rowspan="2">$\neq$spk</td><td>$=$L1</td><td>Fr
Ru</td></tr>
<tr><td>$\neq$L1</td><td>Fr vs. Ru</td></tr>
</table>

Table 1: Modalities of the comparison experiments. spk = speaker ID ; L1 = speaker's mother tongue.

For the `spk` experiment, we want to show if occurrences by the same speaker ($=$spk) yield lower alignment cost than for different speakers ($\neq$spk), when calculated in the embeddings space. For the `L1` experiment, we verify embedding alignment performance between utterances of two native speakers ($=$L1), compared to the performance of the alignment between native-speech and L2-speech ($\neq$L1). We are interested in checking how these hypotheses verify and whether they are modulated by model choice.

4.2. Data

The corpus used in all our experiments was collected to investigate the degree to which Russian-native L2 learners of French rely on the phonemes of their native language when pronouncing French words, as well as how this usage correlates with their L2 proficiency level as defined by the CEFR (Common European Framework of Reference for Languages). It consists of recordings of 12 target words within a frame sentence. The detail of the order in which the stimuli are presented is given in Table 2. Frame sentence is: "dis `target_word`" trois fois and distractor sentence is "La roue sur la rue roule, la rue sous la roue reste".

Order Word	1	2	3	4	5	6
tsarine	13	15	32	50	57	78
j'en chie	1	18	28	49	54	72
sérieux	2	23	31	38	51	69
cache-cache	3	14	26	41	58	67
hier	4	21	30	43	63	76
divan	5	20	29	45	52	68
pour Gabriel	6	22	33	48	60	75
louche	7	17	27	40	62	65
tulle	9	16	34	44	53	77
juxtaposer	10	36	46	61	70	73
pas ceux	11	19	25	39	56	74
garage	12	35	47	55	66	71
distractor	8	24	37	42	59	64

Table 2: Speaker metadata for the Russian – French language interference experiment.

There are nineteen participants, among which

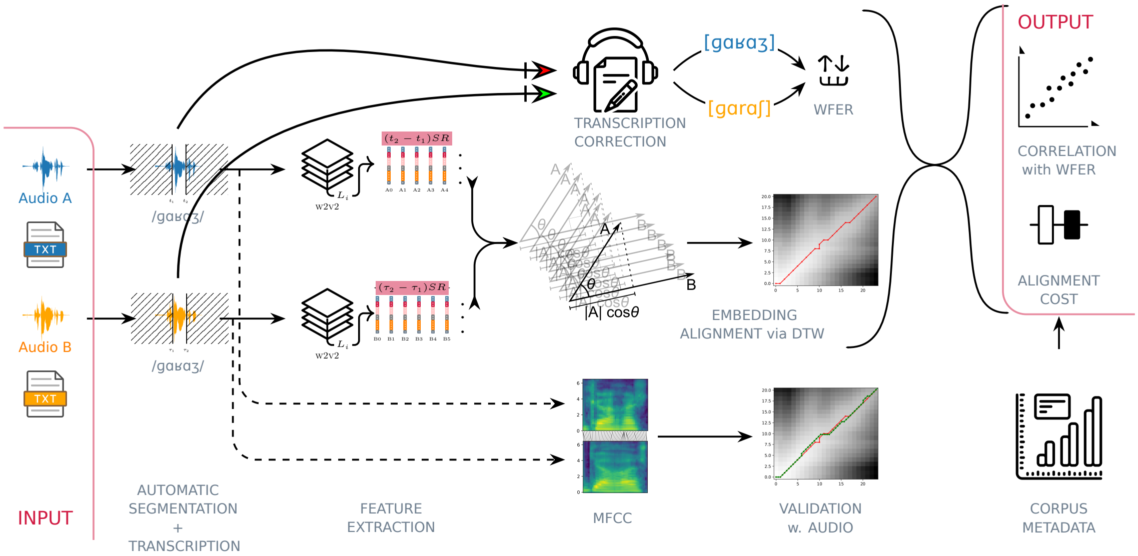


Figure 1: Experimental setup.

ten produce Russian-native accented French (L2 speech) and nine produce native French (native speech). Their age range from 18 to 41 y.o., L2 proficiency from A1 to C2. They were instructed to read out loud 78 randomized sentences, comprising 6 occurrences of the 12 target words and 6 occurrences of a distractor sentence. All recordings were made in the same room in April – May 2019 at the Moscow Lomonosov State University, and all speakers signed an informed consent upon participating to the experiment. Table 3 provides grouped statistics about the participants: natives (All-Fr), learners (All-Ru) including beginners (L2 A1-A2) and advanced (L2 C1-C2), etc. The data is made available in open-access (CC BY-NC-ND-SA 4.0) for verification and reuse in . The high number of occurrences for the same speaker, the shared recording conditions and the available metadata provide a unique framework for speaker-variability studies.

SubGrp	Nb	L1	Age	Gnd	L_{Fr}
All-Fr	9	Fr	23.4	2F;7M	L1
All-Ru	10	Ru	27.5	8F;2M	A1 to C2
L2 A1-A2	4	Ru	25.3	3F;1M	A1-A2
L2 C1-C2	4	Ru	28.5	3F;1M	C1-C2
Fr_ctrl	4	Fr	23.0	2F;2M	L1

Table 3: Average speaker metadata.

The corpus is aligned at word level using Montreal Forced Aligner (MFA. See McAuliffe et al., 2017) on the stimuli list provided to the participants. Alignments were checked manually and corrected by an experienced linguist. The corpus is 38 min 11 s long.

4.3. Experimental setup

The goal of the experiment is to compare different occurrences of a recorded word by measuring the difference in DTW alignment costs after `wav2vec2` feature extraction. These costs are then compared to the error rates determined by an experienced linguist, to verify whether or not the linguist's assessment is correlated with the alignment cost calculated automatically. An overview of the experimental method is presented in Figure 1. All similar pairs of recordings (i.e., same word, different occurrence or speaker, identified with the mention *Audio A* and *Audio B*) are processed using dynamic programming (Giorgino, 2009), either on the embeddings (our main study) or with the audio directly (dashed lines, for verification purposes).

As mentioned in section 4.2, the MFA-generated *phones* tier of each textgrid has been manually corrected to account for the surface realisations of the corresponding recording. These surface transcriptions provide a “gold standard” to our analysis, to which the alignment results can be compared. The phonetic transcriptions are in IPA, and for comparing the transcriptions of two occurrences, the panPhon Normalised Weighted Feature Edit Distance is used (Mortensen et al., 2016). Therefore, instead of the typical Character Error Rate, which accounts for differences without consideration for the phonetic distance between the phonemes replaced, the Weighted Feature Error Rate (WFER) has different multiplication factors according to the nature of the feature differences between reference phoneme and attested phoneme (e.g., 0.25 for a delta on a *velaric* feature, 1 for a *syllabic* feature). In this approach the feature differences between two characters and the weight factors are therefore

syl	son	cons	cont	delrel	lat	nas	strid	voi	sg	cg	ant	cor	distr	lab	hi	lo	back	round	tense	long	velaric
1	1	1	0.5	0.25	0.25	0.25	0.125	0.125	0.125	0.125	0.25	0.25	0.125	0.25	0.25	0.25	0.25	0.25	0.25	0.125	0.25

Table 4: Weights used in Panphon's `jt_weighted_feature_edit_distance`

vectors with one component per phonetic feature.

`jt_weighted_feature_edit_distance` is retained from PanPhon in order to be able to adjust the weight of insertions (i_i) and deletions (d_i) to 0.25 instead of 1. This is done because we consider that read-aloud speech does not favour insertion/deletion errors and they should not bear a weight substantially different from the replaced phonemes since in all likelihood an insertion/deletion will have coarticulatory characteristics in read speech. The weights are stored in a vector $\vec{w_i}$ with the components representing a feature, as illustrated in Table 4. Normalisation is performed by dividing the edit distance obtained by the longest word length. Let A and B denote the two words of length $len(A)$ and $len(B)$ to be compared via WFER :

$$WFER_{A,B} = \frac{\sum_{i=1}^{n_{sub}} \vec{w_i} \cdot \vec{s_i} + 0.25 \cdot (\sum_{i=1}^{n_{del}} d_i + \sum_{i=1}^{n_{ins}} i_i)}{max(len(A),len(B))}$$

with n_{sub}, n_{ins} and n_{del} the minimum number of substitutions, insertions, deletions necessary to turn A into B. Here, $WFER_{A,B}$ represents the phonetic deviation between expert transcription A and B.

Subsequently, the effect of time is analysed separately from the nature of the phonemes encountered by normalising time instead of normalising the embeddings: for a given target word, for all speakers, time is normalized to the mean target word duration value by applying – pre-DTW – a multiplying coefficient to compress or expand the time scale. By doing so, we hope to reduce the effect of time for recordings with large duration differences. The comparison to non time-normalised results shall allow us to determine separately the effect of accuracy and fluency factors in L2-speech.

5. Results

A preliminary study of layer effect is performed, based on the maximisation of the correlation between audio alignment and embeddings' alignment. The nature of the information contained in the embeddings is then addressed starting from section 5.2. The study continues by addressing successively speaker effect and L2 effect on phonologically similar recordings (Sections 5.3 and 5.4).

5.1. Layer effect

In the feature extraction stage, the choice of the layer is experimentally determined in Figure 2 by selecting the layer(s) which best correlate with audio.

A Pearson Correlation evaluation is calculated between audio and embedding alignment costs. Its evolution with layer depth is assessed, to determine the layer most suited to our task.

`XLSR-53` shows a high correlation plateau from layer 0 to 21, then a sudden degradation, such that the last three layers are completely non-correlated with the audio. While the decrease is brutal for `XLSR-53`, it is more progressive for `wav2vec2-FR-7K-large`, which exhibits a high audio-to-embedding correlation on the first layers. Correlation coefficient then gradually decreases with layer depth for `wav2vec2-FR-7K-large`.

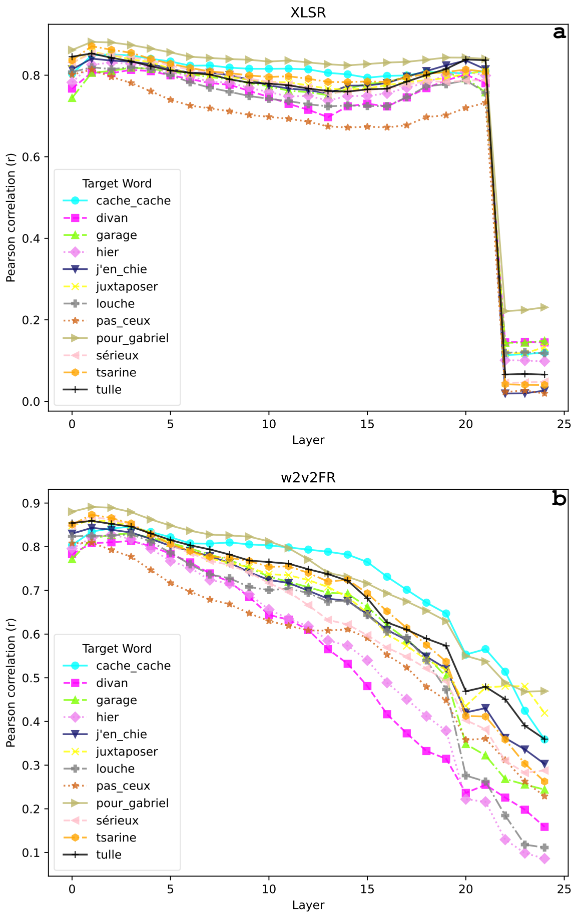
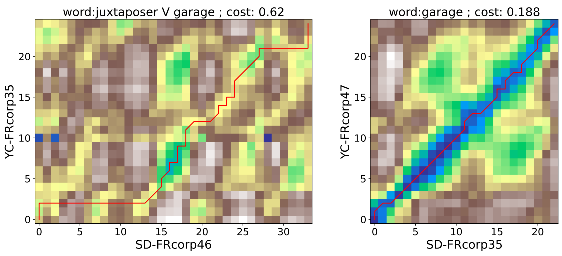


Figure 2: Embedding-based vs Audio-based alignments: Pearson Correlation per Layer (p<0.05) for `XLSR-53` (a) and `wav2vec2-FR-7K-large` (b).

This observation motivates the choice of `XLSR-53` layer 16 because audio and embeddings correlate well and the choice coincides with other studies (Pasad et al., 2021). The optimal layers differ for `wav2vec2-FR-7K-large` model, which suggests that the optimal layer(s) for outputting acoustic/phonetic features are model-specific.

5.2. Effect of the segment

First, the fact that the phonetic information on the segments is present, although self-evident for `wav2vec2` representations (Baevski et al., 2021), is illustrated in Figure 3 by comparing one cosine distance matrix between two different words and between two occurrence of the same word. It shows clearly that the DTW alignment cost on the embeddings of layer 16 is excessively high for words that are different (Figure 3a) when compared to different occurrences of the same word (Figure 3b). For the latter Figure, the alignment path is also more linear, which indicates a more successful alignment.



(a) **(b)**

Figure 3: two alignment matrices for two different speakers, (a) compares Fr: “garage” vs. Fr: “juxtaposer” while (b) compares two occurrences of Fr: “garage”.

5.3. Speaker effect

Speaker effect is addressed by checking if speaker information is as accessible in the embeddings as in the audio by comparing alignment costs for the `same_speaker` and `different_speaker` modalities. Figure 4 shows the variation of the alignment cost, calculated over all target words for these two modalities and expressed on a custom scale due to the differences in orders of magnitudes. We first give an account of these differences: (1) audio cost and embedding alignment costs just represent the same phenomenon in different units and (2) between z-norm and no-norm: the embedding alignment cost is known to increase after embedding z-normalisation.[3] As for (3) t-normalisation: the fact that the cost is increased after normalising time is counter-intuitive since one would have thought that normalising time would help the recordings match each other better. We show why the increase is homogeneous and perfectly normal. It is due to the way the normalised cost is defined: in our model, we defined the alignment cost for a (n, m) matrix as $\frac{S}{M}$ with S = number of steps to get from the bottom-left to the top-right of the cost matrix and $M = max(m, n)$. on time-normalised embeddings, $n = m$, which artificially maximizes the cost. The differences in numerical cost values mainly stem from how this variable is defined.

We are interested in the ratios between modalities: when shown on a proportional scale, audio cost (4a), z-normalised cost (4b) and non-normalised cost (4c), show relatively identical deltas, which suggests that z-normalisation does not add a significant advantage to the embeddings (z-norm is neither more nor less sensitive to speaker ID). The fact that the `same_speaker` – `different_speaker` alignment cost ratios are the same ratios as the audio is an indication that the acoustic information is present in layer 16, including speaker information.

To go further, Figure 4d shows that t-normalisation tends to reduce the gap between `same_speaker` and `different_speaker` modalities. By normalising time we hope to be able to focus on more abstract phonetic units. We therefore introduce a comparison to the PanPhon (Mortensen et al., 2016) Weighted Feature Error Rate (WFER) which represents in discrete, phonetically-informed terms, the differences at the segmental level between one occurrence of a given word (e.g., [gaʁaʃ]) and another occurrence (e.g., [gaʁaʒ]). Figure 5 shows the alignment cost variation with WFER, calculated over all target words. t-normalisation (Figure 5b) for the `same_speaker` – `different_speaker` configurations, tends to produce wider, thicker violin plot distributions. This results in an increased variance for t-normalised recordings, causing more overlap between the two modalities.

Figure 5a, which focuses on non-normalised embeddings and Figure 5b which treats t-normalised embeddings, are analysed:

- Figure 5a shows that the alignment cost is globally stable with WFER for the `same_speaker` modality and that it increases with WFER for the `different_speaker` modality.
- For Figure 5b, interestingly, we observe the opposite after time normalisation: alignment cost is globally stable with WFER for the `different_speaker` modality and increases with WFER for the `same_speaker` modality.

5.4. Foreign accent effect

L2-speech is analysed in light of the model’s performance aligning same occurrences of a word for native speakers or L2 speakers in Figure 6. First of all, the `fr vs. fr` modality exhibits the lowest alignment cost, which was expected.

[3] Same results as in (Le Ferrand et al., 2020).

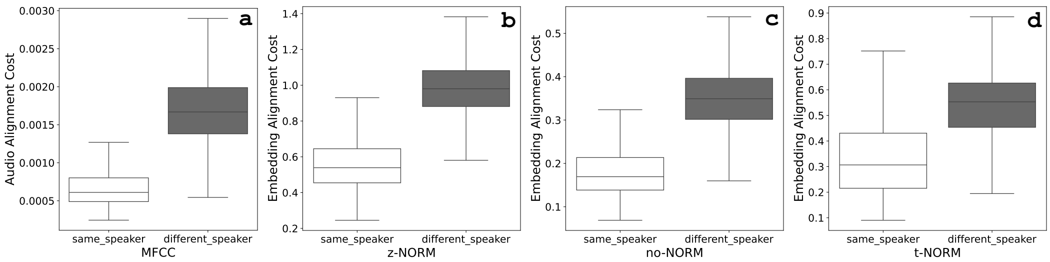


Figure 4: Differences in DTW alignment cost for the `same_speaker` vs. `different_speaker` modalities, for the audio (a), across several normalisation methods: z-normalisation (b) no-normalisation (c) and t-normalisation (d).

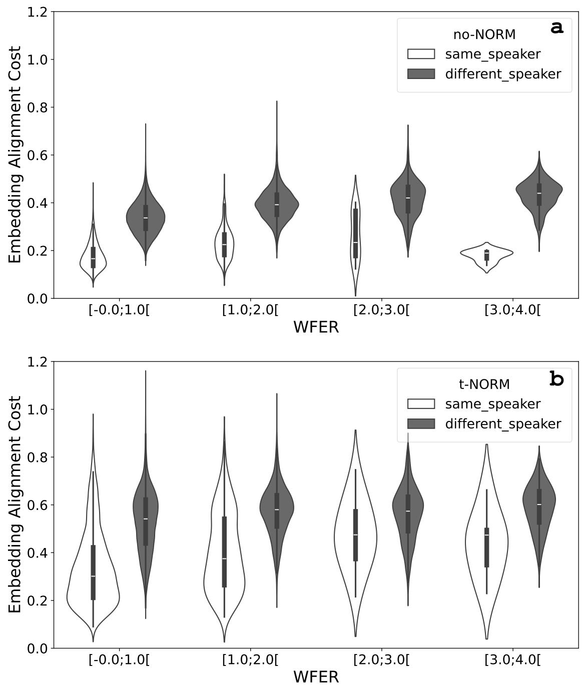


Figure 5: DTW alignment cost variation with WFER for the `same_speaker` vs. `different_speaker` modalities.

By contrast, Figure 6 also shows an odd result: the `ru vs. ru` modality exhibits a higher alignment cost than `fr vs. ru`. This is unexpected because speakers with the same L1, like natives, are expected to exhibit little variability compared to native speech vs. L2-speech (Xie and Jaeger, 2020). In other words, the `ru vs. ru` should have exhibited alignment costs closer to the `fr vs. fr` modality than to the `fr vs. ru` modality. This surprising result led us to split the *L2-French learners* into an advanced and a beginner sub-group. As evidenced in Figures 7a to 7d, this unexpectedly high alignment cost for the `ru vs. ru` modality seems to be rather due to advanced learners (7a, 7b) than to beginners (7c, 7e).

As for t-normalisation effect, overall results (Figure 6) show that the t-normalised setting reduces the discrepancies between groups: the difference in alignment cost between natives and non-natives is less marked after t-normalisation. this is still true after splitting natives into advanced and beginner sub-groups.

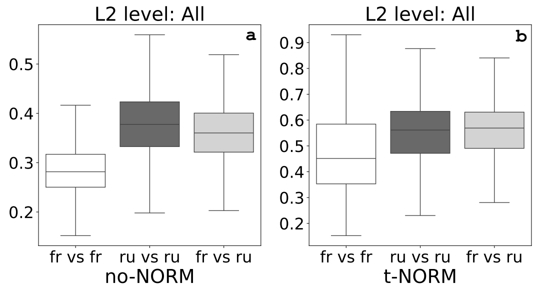


Figure 6: DTW alignment cost for speakers with a different L1: effect of time-normalisation on the whole group of speakers.

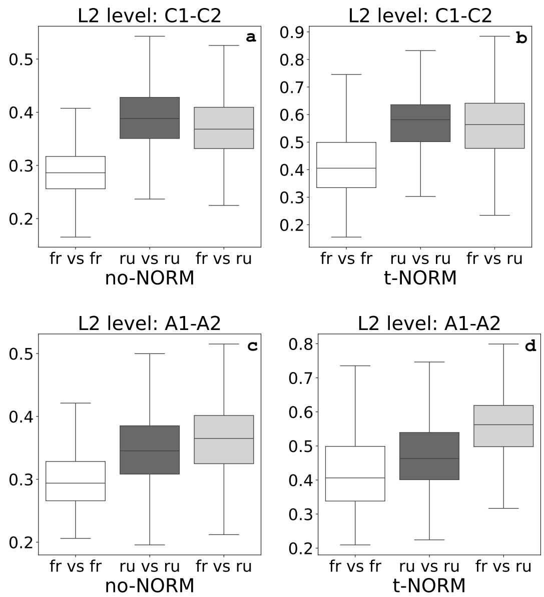


Figure 7: DTW alignment cost for speakers with a different L1: effect of time-normalisation and speakers sub-grouping in levels.

Figure 8 opportunely focuses on the beginner groups by breaking down the alignment costs shown in Figures 7c and 7d. Zooming in on the `ru vs. ru` modality, non-normalised embeddings have a constant alignment cost. Normalising time proportionally reduces the cost for low WFERs and increases the cost for high WFERs for the `ru vs. ru` modality. This means that differences in speech rate impact more strongly situations with low WFERs than with high WFERs.

Conversely, for the `fr vs. ru` modality, the alignment cost does not vary with WFER in both no-norm and t-norm conditions, suggesting that time is not a dominant parameter for this modality.

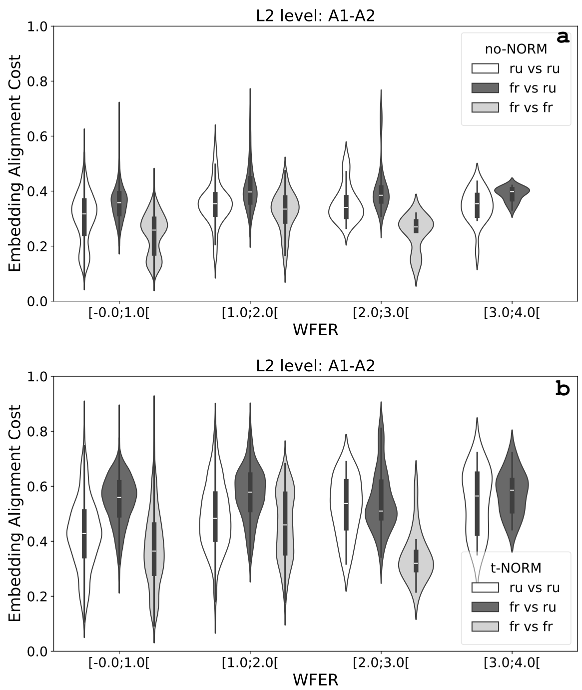


Figure 8: DTW alignment cost variation with WFER for native vs. L2 speech.

5.5. Language Resource References

5.5.1. Online corpus

Our corpus is made up of audio files associated with TextGrids (two per audio file) containing orthographic and phonetic transcriptions, in two versions: the silver version, directly output from MFA, and the gold version, provided by expert phoneticians. It is stored under an Ortolang repository (doi.org/10.82270/ru-fr_interference), is available in open-access with CC BY-NC-SA license (Dashkevich et al., 2026). As a corpus designed for experimental purposes on the acquisition of French (half of the participants) and Russian (the other half), with a single list of participants, a choice of target words phonetically balanced between French and Russian, this resource is valuable for language acquisition studies. Being only half-exploited since the other half (in Russian) did not fit in this submission, the corpus holds potential for future bidirectional studies in French and Russian acquisition. Ortolang is a CLARIN B-Centre (Pierrel et al., 2017).

5.5.2. Online NLP resources

The models used in this study are all available via the huggingface API:

- facebook/wav2vec2-large-xlsr-53
- bofenghuang/asr-wav2vec2-ctc-french

6. Discussion

Our results showed that z-normalization had a very homothetic impact on the embeddings, without any effect on the alignment cost as a function of speaker ID or L1. As a consequence, no z-normalisation was applied to the embeddings.

6.1. Speaker effect

Without time-normalisation, finding any relationship between alignment cost and WFER is a challenge since we were unable to correlate the alignment cost with the Weighted Feature Error Rate. This non-result could be meaningful, in the sense that it could mean that acoustic similarity is more driven by other parameters than the transcriptions, but it would be surprising, especially for a corpus collected in the lab. The literature or foreign accent, by mentioning the *fluency* vs. *accuracy* dichotomy, offered us an angle to approach the issue: comparing t-normalised with non-normalised embeddings.

The fact that t-normalisation reduced the gap between same_speaker alignment costs and different_speakers alignment costs indicates that the time factor, called *fluency factor* in acquisition studies, has for more variability across speakers than within speaker.

This phenomenon, however, is not evenly distributed within the WFER values: for non-normalised embeddings, the alignment costs are closer for low WFERs than for high WFERs, while for t-normalised embeddings, the alignment costs are closer for high WFERs than for low WFERs. It is probable that the effect of t-normalisation is proportionally higher for low WFER values whereas when WFER increases, the variation of the alignment cost is more and driven by segmental differences.

6.2. Foreign accent effect

We first report on a discrepancy between advanced learners and beginners: the beginners seemed more stable in their productions than the advanced learners. Given the very little time alloted to pronunciation acquisition, progress is uneven among learners and may probably diverge with L2-level. It is also more neglected in classes than syntax or semantics which can be done on paper contrary to pronunciation learning.

Given the alignment cost evolution with WFER for the different modalities, `ru vs. ru` and `fr vs. fr` modality are close to each other for low WFERs after t-normalisation. This means that more of the variation within the `ru vs. ru` modality is explained by a fluency factor than for high WFER values. For high WFER values, however, t-normalizing the data does not significantly improve one modality or another.

We see, to conclude, that the speakers' performance can sometimes be better explained decomposed into a *fluency factor*, materialised in the input by delta-to-mean duration value and an *accuracy factor*, materialised by the Feature Error Rate calculated between two utterances. These two effects are not systematically reflected in the sub-groups, which suggests that all sub-groups are not affected in the same way by the fluency factor.

7. Conclusion

This paper had two goals: investigating whether or not the embeddings encompass information other than abstract linguistic content and devising an approach to address basic linguistic problems using the embeddings in a non-supervised manner.

The first main result is that embeddings contain speaker-specific information because cosine distances between same words are smaller for the same speaker than for different speakers. The delta between these two modalities is the same between audio alignment (MFCCs), normalised or non-normalised embeddings. We showed that normalising time before performing DTW tends to bring alignment cost values closer to each other in proportion, which means that speech rate is one important distinguishing factor between speakers in read speech tasks.

The possibility of t-normalising was explored in order to obtain more speaker-independent representations. The goal is only partially reached because we were only able to reduce the discrepancies between `same_speaker` and `different_speaker` modalities, but speakers remain distinct in the embeddings.

One improvement of the approach would consist in being able to link local cost extrema on the alignment curve with the relative position of the error in the WFER calculation. Such a tool, if developed for didactic purposes, would provide learners with a detailed account of what differs in their productions and they would be able to compare to one or several chosen speakers, which is less normative. Using the method for didactic purpose allows creating targeted pronunciation exercises using live-acquired audio data, which provides interactive feedback to learners.

In terms of potential applications outside didactics, fields like endangered languages documentation can use these methods for interdialectal variability studies. Also, sociophonetics, or even pathological speech studies where privacy is a concern can benefit from these developments without any privacy issue since models are not re-trained with this approach.

8. Acknowledgements

We wish to thank the three anonymous reviewers whose very insightful feedback helped improve this paper.

We are also very grateful for the access to the language acquisition corpora, graciously granted by Daria Dashkevich and Ekaterina Biteeva (LPP). This research could not happen without access to this first-hand laboratory data.

This research was partially funded by the following projects, ranked in alphabetical order :

- The DeepTypo project, supported by the *Agence Nationale de la Recherche* (ANR-23-CE38-0003-01),
- The DiagnoSTIC project, supported by the *Agence d'Innovation de Défense* (grant n° 2022 65 007),
- The DIPVAR project, supported by the *Agence Nationale de la Recherche* (ANR-21-CE38-0019),
- The DLS-HN project, supported by the *Agence Nationale de la Recherche* (ANR-23-CE38-0004).

Last but not least, our deepest gratitude goes to those who lend their voice to phonetics : many thanks to the participants, for volunteering their time and effort in our experiments.

9. Bibliographical References

Alexei Baevski, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2021. Unsupervised speech recognition. *Advances in Neural Information Processing Systems*, 34:27826–27839.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477.

Martijn Bartelds, Wietse de Vries, Faraz Sanal, Caitlin Richter, Mark Liberman, and Martijn Wieling. 2022. Neural representations for modeling variation in speech. *Journal of Phonetics*, 92:101137.

Heather Bliss, Jennifer Abel, and Bryan Gick. 2018. Computer-assisted visual articulation feedback in l2 pronunciation instruction: A review. *Journal of Second Language Pronunciation*, 4(1):129–153.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 2426–2430. ISCA.

Tracey M Derwing. 2008. 13. curriculum issues in teaching pronunciation to second language learners. In *Phonology and second language acquisition*, pages 347–369. John Benjamins Publishing Company.

Fanny Ducel, Aurélie Névéol, and Karën Fort. 2025. "you'll be a nurse, my son!" automatically assessing gender biases in autoregressive language models in french and italian. *Language Resources and Evaluation*, 59(2):1495–1523.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65.

Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. 2021. Quantifying bias in automatic speech recognition.

James E Flege. 1995. Second language speech learning: Theory, findings, and problems. *Speech perception and linguistic experience: Issues in cross-language research*, 92(1):233–277.

James Emil Flege. 1981. The phonological basis of foreign accent: A hypothesis. *TESOL quarterly*, 15(4):443–455.

Timnit Gebru. 2019. Oxford handbook on AI ethics book chapter on race and gender. *CoRR*, abs/1908.06165.

Toni Giorgino. 2009. Computing and visualizing dynamic time warping alignments in r: the dtw package. *Journal of statistical Software*, 31:1–24.

Séverine Guillaume, Guillaume Wisniewski, Cécile Macaire, Guillaume Jacques, Alexis Michaud, Benjamin Galliot, Maximin Coavoux, Solange Rossato, Minh-Châu Nguyên, and Maxime Fily. 2022. Fine-tuning pre-trained models for Automatic Speech Recognition: experiments on a fieldwork corpus of Japhug (Trans-Himalayan family). In *ComputEL-5*, Dublin, Ireland.

William N Havard, Renauld Govain, Benjamin Lecouteux, and Emmanuel Schang. 2025. Self-supervised models of speech processing for haitian creole. *BABEL*, 1091(547):544.

Timothy J Hazen, Wade Shen, and Christopher White. 2009. Query-by-example spoken term detection using phonetic posteriorgram templates. In *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 421–426. IEEE.

Sara Kennedy and Pavel Trofimovich. 2017. Pronunciation acquisition. In *The Routledge handbook of instructed second language acquisition*, pages 260–279. Routledge.

Kartik Khandelwal, Preethi Jyothi, Abhijeet Awasthi, and Sunita Sarawagi. 2020. Black-box adaptation of asr for accented speech. *arXiv preprint arXiv:2006.13519*.

Eric Le Ferrand, Steven Bird, and Laurent Besacier. 2020. Enabling interactive transcription in an indigenous community. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3422–3428, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ekaterina Biteeva Lecocq. 2021. *Complexité et contrôle du geste linguo-palatal sous l'éclairage de sa variabilité. Le cas de la palatalisation en russe. Aspects phonétiques et phonologiques.* Ph.D. thesis, Université Grenoble Alpes.

Boris M Lobanov. 1971. Classification of russian vowels spoken by different speakers. *The Journal of the Acoustical Society of America*, 49(2B):606–608.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502.

David R Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. Panphon: A resource for mapping ipa segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484.

Alene Moyer. 2013. *Foreign accent: The phenomenon of non-native speech*. Cambridge University Press.

Murray J Munro. 2008. Foreign accent and speech intelligibility. In *Phonology and second language acquisition*, pages 193–218. John Benjamins Publishing Company.

Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. 2022. French CrowS-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8521–8531, Dublin, Ireland. Association for Computational Linguistics.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921. IEEE.

Jean-Marie Pierrel, Christophe Parisse, Jérôme Blanchard, Etienne Petitjean, and Frédéric Pierre. 2017. Ortolang: a french infrastructure for open resources and tools for language. In *Linköping Electronic Conference Proceedings*, volume 136, pages 102–112.

Dhananjay Ram, Lesly Miculicich, and Hervé Bourlard. 2020. Neural network based end-to-end query by example spoken term detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1416–1427.

Kazuya Saito, Stuart Webb, Pavel Trofimovich, and Talia Isaacs. 2016. Lexical correlates of comprehensibility versus accentedness in second language speech. *Bilingualism: Language and Cognition*, 19(3):597–609.

Nay San, Martijn Bartelds, Mitchell Browne, Lily Clifford, Fiona Gibson, John Mansfield, David Nash, Jane Simpson, Myfany Turpin, Maria Vollmer, et al. 2021. Leveraging pre-trained representations to improve access to untranscribed speech from endangered languages. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1094–1101. IEEE.

Florian Schiel and Andrea Baumann. 2006. Phondat 1, corpus version 3.

Yui Suzukida. 2021. The contribution of individual differences to l2 pronunciation learning: Insights from research and pedagogical implications. *RELC Journal*, 52(1):48–61.

Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

William Timkey and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4527–4546.

Steven H Weinberger and Stephen A Kunath. 2011. The speech accent archive: towards a typology of english accents. In *Corpus-based studies in language use, language learning, and language documentation*, pages 265–281. Brill.

Xin Xie and T Florian Jaeger. 2020. Comparing non-native and native speech: Are l2 productions more variable? *The Journal of the Acoustical Society of America*, 147(5):3322–3347.

Anna Zee, Marc Zee, and Anders Søgaard. 2024. Group fairness in multilingual speech recognition models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2213–2226.

Yuanyuan Zhang, Yixuan Zhang, Bence Mark Halpern, Tanvina Patel, and Odette Scharenborg. 2022. Mitigating bias against non-native accents. In *Interspeech*, pages 3168–3172.

10. Language Resource References

Dashkevich, Daria and Biteeva, Ekaterina and Fily, Maxime. 2026. *ru-fr_interference*. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

A. Optional Supplementary Materials: Appendices, Software, and Data

A.1. Software and programms

The whole pipeline for (i) automatic alignment, (ii) feature extraction, normalisation, DTW cost calculation and (iii) statistical post-processings is provided on our github: https://github.com/maxime-fily/RuFr-interference

A.2. Extra space for ethical considerations and limitations

The models used in this study are diverse: one multinlingual pretrained model and one monolingual fine-tuned model. It would have been better to have a third model (e.g., a monolingual pretrained model).

The limited size of our corpus restricted our ability to explore all aspects of speech variability. Limited size does not invalidate our results, but still, applying our method to larger corpora (Schiel and Baumann, 2006) would certainly help validate and extend our findings. In particular, the gender imbalance between the group of natives vs. non-natives, which we tried to address when we did the per-level sub-grouping, would certainly be less important on a larger corpus.

A.3. Morpheme length effect

We verify in figure 9 that the morpheme length in not a confounding factor for WFER. the distribution of length accros all WFER values confirms that length is not correlated with WFER.

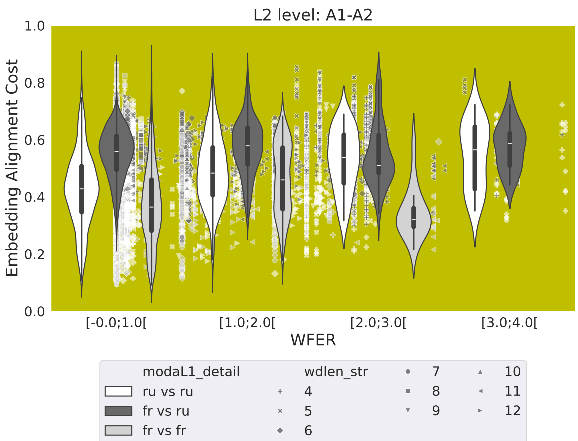


Figure 9: shape-coded length of morphemes for the Foreign accentedness effect: showing that WFER values are not correlated with length.

What LID Systems Say About Dialectal Variation. The Case Of Yiddish, Quechua and Mande

Johanna Cordova, Eric Jordan, Valentina Fedchenko
ERTIM - Inalco
2 rue de Lille, 75007 Paris, France
surname.name@inalco.fr

Abstract

This study investigates the ability of speech-based language identification (LID) systems to handle dialectal variation in low-resource settings and explores whether classification outcomes correspond to phonetic proximity, as well as an exploratory tool for dataset quality. We collected corpora for three macrolanguages Mande, Quechuan, and Yiddish, each presenting distinct internal variation, and evaluated three types of models: GMM, Whisper, and wav2vec2-based architectures. Models were tested both within language families and across the entire multilingual dataset to assess generalization. Layer-wise classifiers built on wav2vec2-XLSR embeddings were used to identify the layers most sensitive to phonetic or phonological features. Results show that simple GMM models can generalize well in small, highly similar datasets, while Whisper-based classifiers tend to overfit, particularly on closely related dialects. Wav2vec2-XLSR (layer 12 + MLP) captures better fine phonetic and prosodic distinctions, suggesting that embeddings encode nuanced pronunciation cues. For datasets with more diverse sources like Quechua, Whisper demonstrates better generalization. Overall, LID classifiers can both reveal linguistic patterns and highlight dataset quality issues, with model architecture and layer-specific representations shaping performance.

Keywords: Spoken Language Identification (SLID), dialectal variation, language classification

1. Introduction

While automatic speech processing achieves impressive results, including for an increasing number of low-resource languages, the more complex issue of handling dialectal variation remains largely underexplored and typically yields unsatisfactory results in large multilingual models (Joshi et al., 2024). The work presented in this article seeks to address two questions: is it possible to quantify the level of granularity in variation that speech processing systems based on embeddings are able to handle in a low-resource context? If so, can the classifications produced by these models be used as a new metric for measuring linguistic proximity between variants? To test these hypotheses, we begin by collecting corpora of variants for three highly distinct macrolanguages, each presenting different issues in terms of internal classification: the Mande languages, the Quechuan languages, and Yiddish dialects. Based on these corpora, we train or fine-tune three types of spoken language identification (SLID) models (GMM, wav2vec2, Whisper) in order to evaluate the classification and get an initial idea of the difficulty of the task with limited data. At the same time, we train a series of classifiers using embeddings extracted from each layer of an XLS-R model to determine whether the layers known to mainly encode phonological features (Pasad et al., 2021) yield higher classification accuracy. This will give us an idea of the extent to which the classification is based on phonetic/phonological criteria. The primary challenge of this task lies in the fact that numerous factors influence the results: the quality, type, quantity, and diversity of available data, as well as the impact of pretraining in the multilingual models used. Initially, our work will focus on identifying how these biases manifest in the languages being studied, thereby guiding more in-depth future research in the resulting directions. This work is therefore intended to be exploratory and preliminary, while providing the necessary data for further studies.

1.1. Quechuan languages

The Quechuan languages, spoken mainly in Peru, Bolivia, Ecuador and Colombia form a linguistic family with a complex internal structure. In terms of phylogenetic classification, the family is divided into two main branches: QI, grouping the central Peruvian languages, and QII (all other languages in the family) (Torero, 1970 [2002]; Cerrón-Palomino, 1987 [2003]; Blum et al., 2023). The number of distinct languages within the family remains unresolved. In Camacho Rios et al. (2024), it is proposed that between 12 and 17 languages can be distinguished, based on the criteria of mutual intelligibility, phonological and morphosyntactic distance, lexical distance, and sociolinguistic perception. The number of variants within these languages is especially difficult to estimate as the variations form a continuum (Mannheim, 2018). The work of SIL (Summer Institute of Linguistics) led to the creation of 44 ISO codes for modern Quechua languages, 43 of which are still spoken today. The criteria used to assign these ISO codes are unclear and many of them refer to linguistic variants

Proceedings of SPEAKABLE @ LREC 2026, pages 98–108
11 May 2026.

rather than distinct languages. Therefore, the current classification of Quechua languages leaves room for improvement. Any improvement on this classification should build on the existing work, ideally enhancing it through empirical foundations. We believe that the analysis of speech signal can be of particular interest to address this issue.

1.2. Yiddish dialects

Yiddish is a language of the Indo-European family, belonging to the Germanic branch, which arose through a historical language shift from Middle High German. Yiddish is considered by Ethnologue[1] as a "macrolanguage", divided into nearly extinct Western Yiddish (yih) and endangered Eastern Yiddish (ydd). Yiddish dialects are not assigned distinct ISO language codes. The present study focuses exclusively on Eastern Yiddish. Dialological research (Weinreich et al., 2008; Jacobs, 2005; Beider, 2015) has shown that the most salient isoglosses separating three major Eastern Yiddish dialects (Nothern-Eastern – NEY, Central-Eastern – CEY and Southern-Eastern – SEY) are primarily phonetic, particularly in the vowel system. Differences at the morphosyntactic and lexical levels are also attested, but they tend to be less systematic, less frequent, and less discriminative for dialect classification. Hasidic Yiddish is a somewhat special case, constituting a distinct sociolinguistic phenomenon: due to intense and ongoing contact with American English and Modern Hebrew, it exhibits a range of contact-induced features, including significant phonetic innovations (Nove, 2021).

1.3. Mande languages

The Mande languages constitute a genealogically coherent family traditionally classified within the Niger–Congo phylum, although their precise position within Niger–Congo has been subject to debate due to their typological distinctiveness. Mande languages are spoken across a wide area of West Africa, extending from Senegal and Mali to Côte d'Ivoire, Guinea, and Burkina Faso. The family is internally divided into several branches, usually grouped into Western, Eastern, Southern, and Southwestern Mande (Kastenholz, 1996; Vydrine, 2004). Bambara (bam), Wan (wan), and Ngen (gnj), the languages considered in the present study, belong to different subgroups within Western and Southern Mande and exhibit substantial phonological and morphosyntactic diversity (Vydrine, 2004; Nikitina and Treis, 2020; Korol, 2022). Unlike Quechua and Yiddish, each Mande language is assigned a distinct ISO 639-3 code, reflecting substantial mutual unintelligibility and perception of clear linguistic boundaries between varieties. Nevertheless, dialect continua and contact-induced variation are also attested, particularly in regions of intense multilingualism.

Typologically, Mande languages are predominantly isolating to slightly agglutinative, display strict SOV word order, and are characterized by tonal systems that play a central role in lexical and grammatical distinctions. Tone inventories and tonal processes vary considerably across the family, contributing to significant phonetic and phonological distance between languages (Vydrine, 2004).

1.4. Linguistic diversity and ISO code attribution in our dataset

The presented dataset of chosen language groups is treated as a structured ensemble of linguistic varieties whose distances are approximated through their ISO language codes. ISO codes represent a formalized attempt to capture perceived distances between linguistic variants, yet the criteria underlying these assignments vary considerably. Linguistic distance is understood here as a complex qualitative notion, grounded in accumulated linguistic evidence (phonological, morphosyntactic, and lexical), mutual interpretability, and the social judgment of speaker communities and linguistic institutions.

The dataset includes three distinct configurations of ISO coding practices. First, in the case of the Mande languages, ISO codes are predominantly assigned to what are widely regarded as separate languages, corresponding to relatively large phonological and grammatical distances. Second, Quechuan languages presents a scenario, where ISO codes have been assigned inconsistently and often correspond to dialectal or sub-language-level variation within a dense dialect continuum. Third, Eastern Yiddish dialects constitute a case in which no separate ISO codes are assigned, despite well-established dialectological, mostly phonetic, distinctions.

From a typological perspective, the dataset brings together highly diverse phonetic and morphosyntactic systems. It includes tonal Mande languages, which rely heavily on pitch contrasts; agglutinative Quechua languages, characterized by complex suffixal morphology; and inflectional Yiddish dialects, whose phonetic profiles reflect Middle and Eastern European areal features. These typological differences have direct consequences for how the languages are represented in acoustic language models.

[1] https://www.ethnologue.com/language/yid/

2. Datasets

To ensure comparability, all datasets created contain 2.5 hours of audio data per ISO code or variant. The files have all been converted to 16kHz mono WAV format and are between 5 and 30 seconds long. All datasets have a train, validation and test split, each containing utterances produced by different speakers. We have taken particular care in preparing the data in order to reduce speaker- and microphone-recognition biases. The overall quality of the corpora was measured using the Signal-To-Noise Ratio; a visualisation of the results is provided in Appendix 7, Figure 9. The overall synthesis of the corpus, including the train-dev-test split distribution and the number of speakers per split, is presented in Figure 11.

2.1. Quechuan languages

Two main sources with ISO code identification are available for Quechuan languages:

- Mozilla Data Collective[2] (MDC). This successor to Common Voice brings together speech datasets developed by speaker communities. For Quechua languages, 17 datasets are available, covering 16 ISO codes. All but one of these datasets consist of scripted speech: speakers read very short sentences (3.5 words in average) with simple syntax.
- Recordings of the Bible and New Testament. Many variants have recordings available online, some of which are prepared for NLP purposes[3]. This is the main data source for Quechuan languages in the large multilingual models.

Outside of these corpora, the volume of available audio content in Quechua varies significantly depending on the language. Southern Quechua (QIIC linguistic group) stands out as the majority language in terms of both speakers number and media visibility (social networks, radio, television), which facilitates data collection. For the LID experiments, we will focus on this language, comparing 4 closely related variants, for which we have collected between 15 and 30 minutes of additional data from the media or from recordings in our possession. These variants are divided into two sub-groups: Chanka (`quy`) and Collao (`quz`, `qxp`, `quh`). At the phonological level, they differ in that the Collao variants feature ejective and aspirated consonants. This distinction allows for rapid identification, even with short utterances. Within the Collao group, the differences are more subtle, as the variants share the same phonological system; our experiments will seek to determine whether the models are able to achieve this level of granularity in distinguishing between the variants.

[2] https://datacollective.mozillafoundation.org/datasets?q=quechua

[3] https://huggingface.co/datasets/Flux9665/BibleMMS

2.2. Yiddish dialects

Two Yiddish datasets were constructed for the experiments: one comprising the full training pipeline (train, development, and test splits), and a separate dataset for out-of-domain evaluation. Both datasets consist of three Eastern Yiddish varieties: Northern Eastern Yiddish (yid_ney), Central Eastern Yiddish (yid_cey), and Southern Eastern Yiddish (yid_sey). Data were drawn from the Corpus of Spoken Yiddish in Europe (CSYE, Bleaman and Nove (2025)). As the corpus consists of post-Holocaust interviews with native Yiddish speakers living in the United States, a preprocessing pipeline was applied to filter out interviewer speech produced by non-native speakers, as well as segments containing Yiddish–English code-switching.

At the initial stage, our experiments included Hasidic Yiddish data extracted from the Mozilla Common Voice corpus (Ardila et al., 2020). However, the preliminary observations suggested a strong bias toward the Hasidic variety, which systematically attracted the majority of the model's predictions. We hypothesized that the model was relying on non-linguistic cues. Therefore, to reduce source-related effects, we restricted the dataset to European Yiddish dialects, which are more homogeneous both linguistically and with respect to the types of recordings available.

The second out-of-domain test dataset was drawn from a different source corpus—Reading Electronic Yiddish Documents (REYD, Webber et al. (2022))—for the `yid_ney` and `yid_cey` varieties. This dataset represents a distinct type of data, consisting of read speech from books, and involves different speakers and recording conditions. For `yid_sey`, we used a YouTube interview with a native speaker of the dialect, which was diarized in order to bring it into a format comparable to the other datasets. The composition of this out-of-domain evaluation dataset is presented in Table 1.

Dialect	No. of speakers	No. of segments
`yid_ney`	2	461
`yid_cey`	2	387
`yid_sey`	1	397

Table 1: Out-of-domain test corpus for Yiddish dialects

2.3. Mande languages

The dataset comprising the three Mande languages, Bambara, Wan, and Ngen, represents a subset of a continuous fieldwork effort spanning more than twenty years within Mande-speaking communities. The data were collected in naturalistic settings and cover a range of speech recordings, including spontaneous narratives, elicited vocabularies, and translations of isolated phrases. This diversity of data provides a broad sample of speech styles and communicative contexts, while also introducing variability in prosodic and segmental realization.

The recordings for Wan and Ngen originate from unpublished personal fieldwork corpora[4]. Prior to inclusion in the present study, these data underwent preliminary processing, including semi-automatic segmentation of the audio into phrase-level units. In addition, metadata were compiled on the basis of field notes provided by the original data collectors, including information on discourse type, topical domain, and the number of speakers involved in each recording. This metadata was used both to ensure internal consistency of the dataset and to support controlled experimental splits where possible.

The Bambara data are substantially more voluminous and heterogeneous. A portion of this dataset is publicly available through the CORPORAN platform (Corpus oraux annotés), hosted by TGIR Huma-Num[5] (Vydrin, 2013, 2014). These recordings include annotated oral corpora collected across multiple regions and communicative settings. For the purposes of the present experiments, a curated subset of the Bambara recordings was selected to ensure comparability with the Wan and Ngen datasets in terms of segment duration, recording conditions, and distribution of recording types.

ISO Code	Split	No. of speakers	No. of segments
quy	train	> 7	654
	dev	> 2	167
	test	> 11	100
quz	train	> 12	797
	dev	9	118
	test	12	100
qxp	train	9	840
	dev	1	202
	test	9	100
quh	train	> 8	711
	dev	> 2	128
	test	> 8	100
bam	train	9	594
	dev	3	148
	test	2	148
wan	train	2	509
	dev	1	113
	test	1	112
gnj	train	2	544
	dev	2	146
	test	1	136
yid_ney	train	25	977
	dev	9	209
	test	5	203
yid_cey	train	14	903
	dev	5	220
	test	5	216
yid_sey	train	26	1018
	dev	5	212
	test	5	229

Table 2: Numbers of speakers and segments across all datasets and splits

3. State of the art in SLID

The different languages included in this study do not receive the same amount of coverage in the training data of the state of the art acoustic large language models. For the Mande languages, the Whisper model (Radford et al., 2022) includes training data for Bambara but lacks coverage of Wan and Ngen. Omni-ASR (Omnilingual et al., 2025) does not include Wan or Ngen, although it does incorporate other Mande languages that are phonetically similar to Wan, such as Mwan (moa). Eastern Yiddish, treated as a single entity in most models, receives no dialect-specific acoustic modelling, despite the phonetic salience of dialectal distinctions.

Quechuan languages receive comparatively better coverage: for the LID task, the MMS model reports support for 34 ISO codes. To assess the actual performance of this model, we evaluate it using the test split of our dataset, supplemented with a few utterances from the train split to ensure 100 files per language. The overall accuracy is relatively low, with a micro-F1 score of 0.46. The confusion matrix in Figure 1 reveals that the language most distinct from the other three (quy) is the best predicted. The overprediction of the quh class, which absorbs similar variants, may be due to an imbalance in the training data. This potential imbalance must be accounted for when analysing the results produced by pretrained models.

State-of-the-art dialect classification systems are

[4] A short fragment of Ngen corpus was published in Pangloss Collection of oral data: https://pangloss.cnrs.fr/corpus/Ngen. In written form, part of the Wan corpus has been incorporated into the Speech Reporting Corpus (Nikitina, 2023), a curated collection of narrative texts focusing on discourse reporting strategies in storytelling

[5] https://corporan.huma-num.fr/

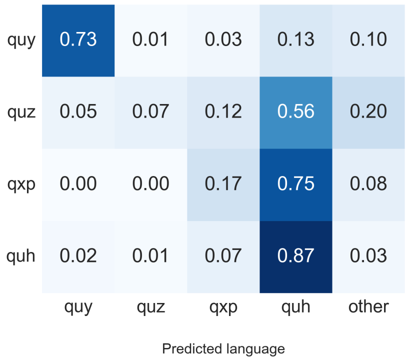


Figure 1: MMS LID classification result

often built on unsupervised acoustic representations such as traditional spectral features (Mel-Frequency Cepstral Coefficients), temporal derivatives (shifted-delta cepstra), or high-resolution signal processing outputs such as Single Frequency Filtering (SFF) and Zero-Time Windowing (ZTW), to capture phonetic and prosodic variability. When paired with robust classifiers, either classical statistical models (GMM, SVM) or neural architectures (Temporal Convolutional Networks, Time Delay Neural Networks, and variants like ECAPA-TDNN), the resulting systems achieve satisfactory performance on fine-grained dialect tasks: such combinations typically produce accuracies in the 80 % range according to multiple benchmarks in the literature (Agrawal et al., 2016; Chittaragi et al., 2018; Tibi and Messaoud, 2025; Kethireddy et al., 2022). Some research has already been conducted to understand which features these systems primarily rely on for classification decisions. For instance, (Bafna and Wiesner, 2025)'s work indicates that neural systems such as ECAPA-TDNN particularly encode accent-related information.

Conversely, handling dialectal variation is poorly suited to the architectural design of large language models: firstly, the lack of available corpora for non-standard varieties deprives tools of training data. Furthermore, the architectures of widely used models are designed to smooth out variations to ensure robustness against noise, often leading to outputs that are homogenised toward the closest majority variant. Consequently, attempts to identify dialects using models that encode the signal through embeddings yield mixed results, with a rapid decline in accuracy as dialectal similarity and the number of classes to discriminate increase (Joshi et al., 2024; Lonergan et al., 2023). The work of Van Dinh et al. (2024) on Vietnamese dialects serves as a particularly compelling example: the authors compiled a corpus of 102 hours of audio curated from publicly available sources for 63 regional variants. From this, they fine-tuned several flavours of pre-trained models (wav2vec2 and Whisper) for dialect identification, producing classification models with varying granularity. Each series of models is trained to predict 1) the region (3-class); 2) the dialect within each region (19 to 25-class); 3) the dialect at the global level (63-class). The results show high accuracy for the first task, but performance drops dramatically as the classification grows more complex. Significant F1-score variations in the performance of the 19-class models across regions suggest that there is a threshold of linguistic similarity beyond which dialectal variations are no longer recognised.

Understanding how embedding systems encode acoustic signals is a particularly important challenge, and several studies have addressed this question (Pasad et al., 2021; English et al., 2024), showing that:

- Early layers focus more on acoustic features (low-level signal properties),
- Middle layers begin to encode phonetic structure as well as tone and stress (de la Fuente and Jurafsky, 2024),
- Higher layers reflect more abstract linguistic units.

4. Experiment and results

Several speech-based language identification models were evaluated in this study in order to compare architectures with different inductive biases and to reveal possible correlations with linguistic distances and the system of ISO codes. The tested models include: (i) a Gaussian Mixture Model (GMM) baseline; (ii) *Whisper*, a transformer-based encoder–decoder model originally designed for automatic speech recognition (ASR) (Radford et al., 2022); and (iii) self-supervised speech representations based on `wav2vec2-base` (Baevski et al., 2020).

The models were evaluated under two experimental configurations: (a) separately within each language or dialect group, and (b) jointly on the entire multilingual dataset. This design allows us to assess model generalization under varying degrees of linguistic similarity and data scarcity. Specifically, we examine performance on: (i) highly similar and largely mutually intelligible dialects (Eastern Yiddish and Quechua) (ii) non-interintelligible languages within a genetically homogeneous family (Mande languages); and (iii) macro-groups of genetically and typologically diverse languages (Mande vs. Quechua vs. Yiddish).

4.1. GMM-based language identification

To establish a baseline free from biases induced by model pretraining, we first train a Gaussian

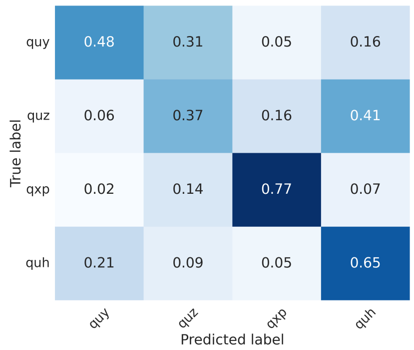
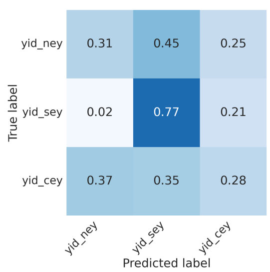


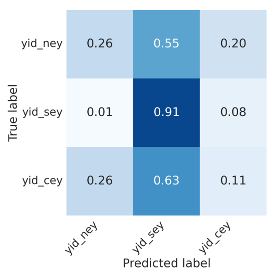
Figure 2: GMM predictions for Quechua variants

Mixture Model (GMM)-based classification system (Torres-Carrasquillo et al., 2004), using MFCC feature vectors extracted at the frame level from each speech utterance. Features are standardised using a scaler. The normalized MFCC vectors are then used to train class-specific GMMs via maximum likelihood estimation, whilst model hyperparameters (number of mixture components, covariance structure and covariance regularisation) are optimized through grid search.

The results of this classification are consistent across the tested languages: for languages with marked differences, such as Mande languages, the classification has a 100% accuracy. For Quechua variants, the accuracy drops to 0.57, with disparities among classes (see Figure 2): the `quz` variant is frequently confused with `quh`, with which it shares the same phonological system. The strong performance for `qxp` is likely due to the higher quality of its data and its lower internal variability. Accuracy is even lower for Yiddish dialects (see Figure 3) : 0.46 with in-domain test set and 0.43 when adding out-of-domain samples.



(a) In-domain test



(b) Out-of-domain test

Figure 3: GMM predictions for Yiddish dialects

4.2. Whisper-based language identification

Whisper is primarily designed as an ASR model and not as a dedicated LID system. In its original architecture, language identification is performed implicitly via the generation of a single language token at the beginning of decoding. As noted in prior work, this design introduces structural limitations for standalone LID tasks: because the language label corresponds to only the first generated token, the model relies predominantly on acoustic cues and makes only limited use of deeper linguistic context. As a consequence, Whisper's LID accuracy has been reported to reach approximately 80.3% in some evaluations, despite its strong ASR performance (Radford et al., 2022; Shen et al., 2023).

We tested Whisper in two configurations: (1) native language identification using the built-in LID classifier, and (2) LID based on acoustic embeddings extracted from intermediate layers of Whisper-small (layer 8) and Whisper-large (layer 20) (Yue et al., 2026), followed by an MLP classifier.

For the first configuration (1), we started by investigating Whisper-based LID on the Yiddish dataset, which represents a particularly challenging scenario due to the high phonetic similarity between dialects and the limited number of classes (three). An initial training run yielded an extremely high F1-score after the very first epoch (0.83), clearly indicating severe overfitting. To encourage the model to generalize over linguistic properties, we tested several increasingly constrained fine-tuning configurations on Whisper-small, always starting from pretrained weights. Our strategies varied in the degree of parameter freezing, from partial encoder fine-tuning to adaptation of the language identification (LID) head only. These progressively restrictive configurations effectively reduced the F1-score observed after the first training epoch to 0.45, indicating a decrease in overfitting and a reduced reliance on speaker- or corpus-specific cues.

Model training was conducted using a learning rate of 1×10^{-3}, with early stopping applied based on development set performance and a patience value of 3 epochs. To reduce overfitting and improve optimization stability, weight decay was set to 0.01 in the optimizer. Given the low-resource nature of the datasets and the resulting imbalance between language and dialect classes, a weighted cross-entropy loss function was employed.

The second configuration, particularly the MLP classifier built on representations from layer 20 of Whisper-large, shows reduced overfitting, it effectively separates more distantly related Mande languages, but still struggles to distinguish linguistically very similar varieties, such as Yiddish dialects. It nevertheless yields insightful results on the Quechua dataset, as illustrated in Figure 4.

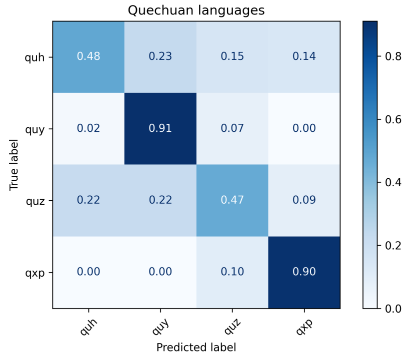


Figure 4: Confusion Matrix for Whisper-based LID (layer 20 + MLP) for Quechuan variants

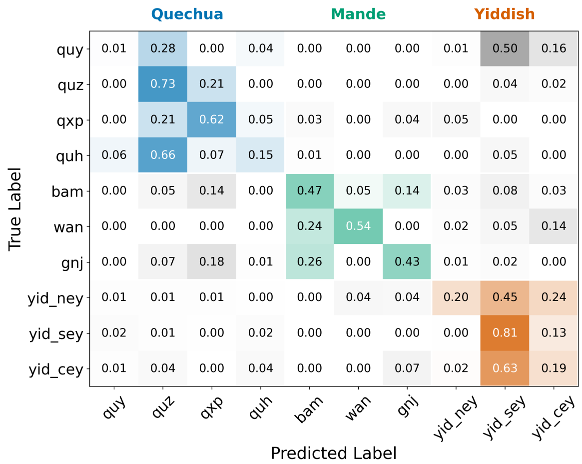


Figure 5: Confusion Matrix for Classifications on Macro test

4.3. Wav2Vec-based language identification

We used the widely adopted wav2vec2 architecture in 2 configurations. Firstly, the `wav2vec2-base` adapted for our task by fine-tuning the model alongside a classification head on our data (using the HuggingFace Wav2VecForSequenceClassification [6] config) and secondly, using the xlsr model extracting different layers of the model and training a classification head on the extracted embeddings.

Several configurations were tested for `wav2vec2-base`, however given the limited amount of data available from our datasets configuring the training process to allow the models to learn without overfitting the training data was of the utmost importance. With this in mind several configurations of hyperparameters and freezing of layers were tested. The most effective training configuration was obtained by freezing the CNN feature extractor and all but the final (12th) layer of the model.

The best balance between learning from the data while avoiding overfitting was found using a learning rate of 3×10^{-6} while performing a warmup on the first 6% of steps and using a weight decay of 0.1. The training was run for a total of 20 epochs with an early stopping patience of 5 epochs.

The most interesting result obtained for this model was on the macro test with all classes, where the model achieved an overall accuracy of 0.42 across the ten classes. However, perhaps more interesting than the pure performance of the model is how errors in classification tended to stay within language families as shown in Figure 5. This suggests that, in most cases, some common features within each of the language families is being learnt by the model. While the intra-family classification described above and this macro-performance indicate some linguistic criteria are playing a role in the classification performance, other factors of convergence cannot be entirely ruled out. Despite efforts to homogenize the data across language families, it must be noted that greater similarity in recording types (i.e. fieldwork data for Mande languages or Bible readings for Quechua) are also potential explanations for this intra-family classification performance.

[6] `https://huggingface.co/docs/transformers/en/model_doc/wav2vec2`

4.4. Classifying from layer-extracted embeddings

We extracted hidden representations from each of the 24 transformer layers of `wav2vec2-xls-r-300m` and performed mean pooling over the temporal dimension to obtain a one-dimension output. These layer-wise embeddings were then used as input to a simple multilayer perceptron classifier (PyTorch SimpleClassifier) with a dropout rate of 0.3. We trained a separate classifier for each layer's embeddings and evaluated their performance on the test split, in order to investigate whether the layers that encode the richest phonological information are also those that yield the highest dialect classification accuracy.

The results for Yiddish, shown in Figure 6 support this theory: a peak in performance is indeed observed for all variants at the twelfth layer of the model. This trend, however, is not observed for Quechua (Figure 7), where accuracy remains stable (and high from the earliest layers), possibly due to the greater presence of this language in the model's pretraining data.

5. Concluding discussion

These results suggest that, in a low-resource setting involving a small number of highly similar linguistic varieties, such as the three Eastern Yiddish

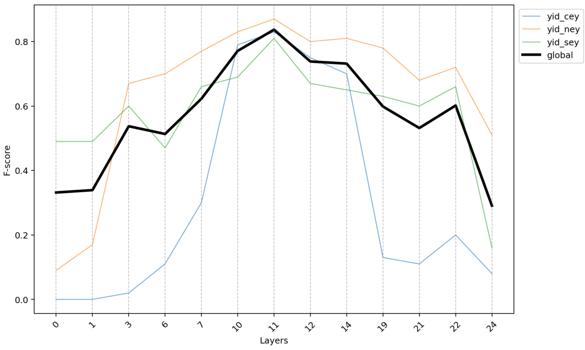


Figure 6: Fscore evolution through layers for Yiddish

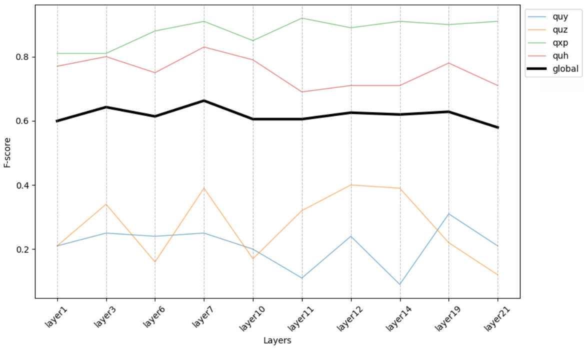
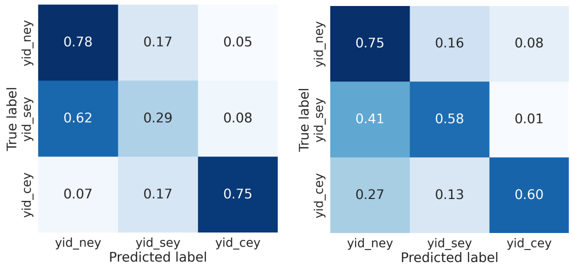


Figure 7: Fscore evolution through layers for Quechua



(a) In-domain test (b) Out-of-domain test

Figure 8: MLP predictions for Yiddish by classifier trained with embeddings from the 12th layer

dialects, the relatively simple GMM architecture exhibits a strong capacity for generalization. In contrast, Whisper-based classifiers appear to rely on a wider range of cues that may include extra-linguistic factors, limiting their ability to capture fine-grained phonetic distinctions between closely related varieties. The most granular analysis is provided by the wav2vec2-xlsr model (layer 12 + MLP). The model appears to rely more on finer phonetic and prosodic features than on coarse phonological cues: for instance, `yid_sey` speakers in our test dataset often exhibit what could be described as a "Russian accent" and are frequently bilingual in Russian, whereas `yid_cey` speakers are typically bilingual in Polish, a pattern not shared with `yid_ney` speakers. The wav2vec2-based classifier may capture these subtleties in pronunciation.

This interpretation is further supported by the layer-wise visualization of F1 scores (6), which shows that the F1 trajectory for `yid_cey` diverges from that of `yid_ney` and `yid_sey`, which are closer to each other (6). There is no reason to believe that the model was exposed disproportionately to `yid_ney` or `yid_sey` during training, since `yid_cey` is actually better represented in publicly available online Yiddish data than `yid_sey`. The layer-wise analysis of wav2vec2 F1 scores 6 indicates that performance is highest at layer 12, a layer often associated with phonetic sensitivity in probing studies (Pasad et al., 2021; San et al., 2024). While not conclusive, this pattern suggests that the wav2vec2-based classifier may rely primarily on phonetic representations when performing language identification generalization.

Conversely, for datasets with greater internal diversity of domain, such as the Quechuan languages, the Whisper-based model demonstrates superior generalization capabilities, likely benefiting from its broader representational capacity.

Finally, despite efforts to reduce bias introduced by the audio data available it cannot be excluded that the results from the macro-language classification may in fact be influenced by similarities between the files and their origins (fieldwork data, bible recordings etc.). Further work will aim to investigate the influence of these categories of data on classification performance. Nonetheless, this result indicates the promise of applying these models when processing even quite diverse corpora to determine possible similarities between audio files.

Future work will extend the present study to a broader range of varieties within the considered language groups, with particular emphasis on Quechuan and Mande languages. We plan to deepen the layerwise analysis of model predictions in order to better understand which linguistic properties are captured at different representational levels. In addition, we will systematically investigate the relationship between data quality and diversity and their impact on model performance. Finally, we aim to explore methods for extracting more fine-grained linguistic knowledge from speech models, with the goal of improving their usefulness for dialectological research.

6. Acknowledgments

We sincerely thank the linguists who generously shared their fieldwork datasets with our team for the purposes of this experiment: Tatiana Korol, Tatiana Nikitina and Valentin Vydrine. The work is supported by the French National Research Agency and Ministry of Higher Education, Research and Innovation (MESR).

7. Bibliographical References

S. Agrawal, Aruna Jain, and S. Sinha. 2016. Analysis and modeling of acoustic information for automatic dialect classification. *International Journal of Speech Technology*, 19:593–609.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Alexei Baevski, Henry Zhou, Abdel rahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *ArXiv*, abs/2006.11477.

Niyati Bafna and Matthew Wiesner. 2025. Lid models are actually accent classifiers: Implications and solutions for lid on accented speech. *arXiv preprint arXiv:2506.00628*.

A. Beider. 2015. *Origins of Yiddish Dialects*. Oxford Linguistics. Oxford University Press.

Isaac L Bleaman and Chaya R Nove. 2025. The corpus of spoken yiddish in europe: Goals, methods, and applications. *Language Documentation & Conservation*, 19.

Frederic Blum, Carlos Barrientos, Adriano Ingunza, and Zoe Poirier. 2023. A phylolinguistic classification of the quechua language family. *Indiana*, 40(1).

Gladys Camacho Rios, Simeon Floyd, and Félix Julca Guerrero. 2024. ¿cuántas lenguas quechuas hay? una estimación del número de lenguas quechuas. *Lexis*, 48(1):34–77.

Rodolfo Cerrón-Palomino. 1987 [2003]. *Lingüística quechua*. Centro de estudios rurales andinos" Bartolomé de Las Casas".

Nagaratna B Chittaragi, Ambareesh Prakash, and Shashidhar G Koolagudi. 2018. Dialect identification using spectral and prosodic features on single and ensemble classifiers. *Arabian Journal for Science and Engineering*, 43(8):4289–4302.

Antón de la Fuente and Dan Jurafsky. 2024. A layer-wise analysis of mandarin and english suprasegmentals in ssl speech models. In *Interspeech 2024*.

Patrick Cormac English, Erfan A. Shams, John D. Kelleher, and Julie Carson-Berndsen. 2024. Following the embedding: Identifying transition phenomena in wav2vec 2.0 representations of speech audio. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6685–6689.

Neil G. Jacobs. 2005. *Yiddish: A linguistic introduction*. Cambridge University Press.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2024. Natural language processing for dialects of a language: A survey. *ACM Computing Surveys*, 57:1 – 37.

Raimund Kastenholz. 1996. Sprachgeschichte im west-mande. köln: Rüdiger köppe verlag.

Rashmi Kethireddy, Sudarsana Reddy Kadiri, and S. Gangashetty. 2022. Deep neural architectures for dialect classification with single frequency filtering and zero-time windowing feature representations. *The Journal of the Acoustical Society of America*, 151 2:1077.

Tatiana Korol. 2022. Preliminary description of ngen pronominal elements. *Mandenkan*, 68:43–58.

Liam Lonergan, Mengjie Qian, Neasa Ní Chiaráin, Christer Gobl, and Ailbhe Ní Chasaide. 2023. Towards spoken dialect identification of irish. In *2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages*.

Bruce Mannheim. 2018. Three axes of variability in quechua. *The Andean World*, pages 507–523.

Tatiana Nikitina. 2023. *A narrative corpus of Wan*. CNRS-LLACAN & LACITO.

Tatiana Nikitina and Yvonne Treis. 2020. The use of manner demonstratives in discourse: A contrastive study of Wan (Mande) and Kambaata (Cushitic). In Åshild Næss, Anna Margetts, and Yvonne Treis, editors, *Demonstratives in discourse*. Language Science Press.

Chaya R. Nove. 2021. *Outcomes of language contact in New York Hasidic Yiddish*, pages 43–71. Berlin: Language Science Press.

Team Omnilingual, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe

Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, Kehan Lyu, Sagar Miglani, Vineel Pratap, Kaushik Ram Sadagopan, Safiyyah Saleem, Arina Turkatenko, Albert Ventayol-Boada, Zheng-Xin Yong, Yu-An Chung, Jean Maillard, Rashel Moritz, Alexandre Mourachko, Mary Williamson, and Shireen Yates. 2025. Omnilingual asr: Open-source multilingual speech recognition for 1600+ languages.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*.

Nay San, Georgios Paraskevopoulos, Aryaman Arora, Xiluo He, Prabhjot Kaur, Oliver Adams, and Dan Jurafsky. 2024. Predicting positive transfer for improved low-resource speech recognition using acoustic pseudo-tokens. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 100–112, St. Julian's, Malta. Association for Computational Linguistics.

Peng Shen, Xuguang Lu, and Hisashi Kawai. 2023. Generative linguistic representation for spoken language identification.

Nejib Tibi and Mohamed Anouar Ben Messaoud. 2025. Arabic dialect classification using an adaptive deep learning model. *Bulletin of Electrical Engineering and Informatics*.

Alfredo Torero. 1970 [2002]. *Idiomas de los Andes. Lingüística e historia*. Editorial horizonte.

Pedro A Torres-Carrasquillo, Terry P Gleason, and Douglas A Reynolds. 2004. Dialect identification using gaussian mixture models. In *Odyssey*, pages 297–300.

Nguyen Van Dinh, Thanh Chi Dang, Luan Thanh Nguyen, and Kiet Van Nguyen. 2024. Multi-dialect vietnamese: Task, dataset, baseline models and challenges. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7476–7498.

Valentin Vydrin. 2013. Bamana Reference Corpus (BRC). *Procedia - Social and Behavioral Sciences*, 95:75–80.

Valentin Vydrin. 2014. Bambara and maninka manding languages written corpora project ("projet des corpus écrits des langues manding : le bambara, le maninka") [in French]. In *TALN-RECITAL 2014 Workshop TALAf 2014 : Traitement Automatique des Langues Africaines (TALAf 2014: African Language Processing)*, pages 109–113, Marseille, France. Association pour le Traitement Automatique des Langues.

Valentin Vydrine. 2004. Areal and genetic features in west mande and south mande phonology: In what sense did mande languages evolve? *Journal of West African Languages*, XXX(2):113–125.

Jacob Webber, Samuel K. Lo, and Isaac L. Bleaman. 2022. Reyd – the first yiddish text-to-speech dataset and system. In *Interspeech 2022*, pages 2363–2367.

M. Weinreich, P. Glasser, P.E. Glasser, Y.I.J. Research, and S. Noble. 2008. *History of the Yiddish Language*. Number 1 in History of the Yiddish Language. Yale University Press.

Zhengjun Yue, Devendra Kayande, Zoran Cvetkovic, and Loweimi Erfan. 2026. Probing whisper for dysarthric speech in detection and assessment. In *2026 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2026*. IEEE.

Appendix

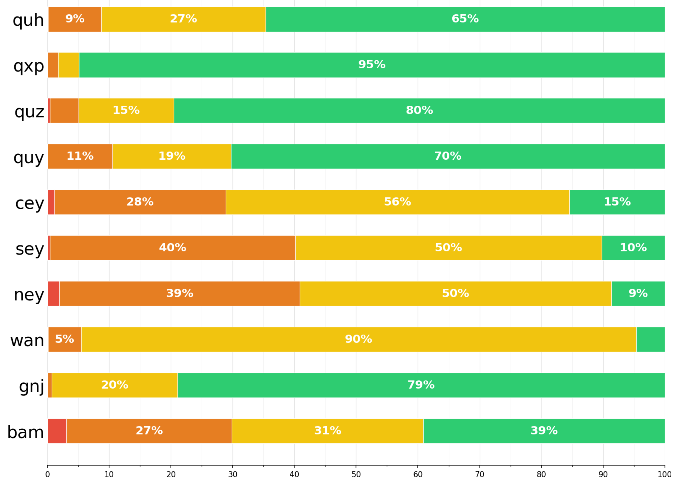


Figure 9: Estimation of corpus quality using the Signal-To-Noise Ratio (SNR).
● Poor ● Average ● Good ● Excellent

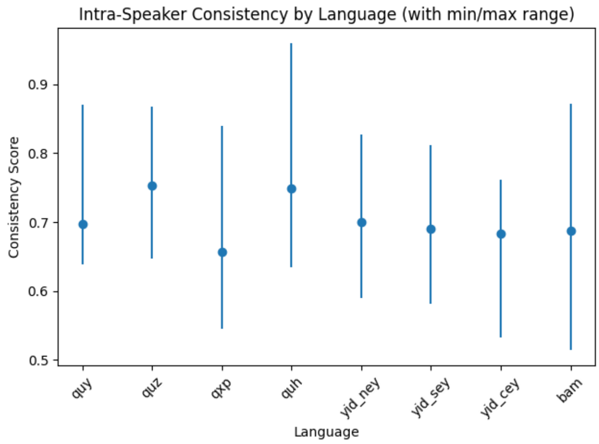


Figure 10: Intra-Speaker Temporal Consistency across the corpus. Data are not provided for Ngen and Wan because the speakers were not manually identified.

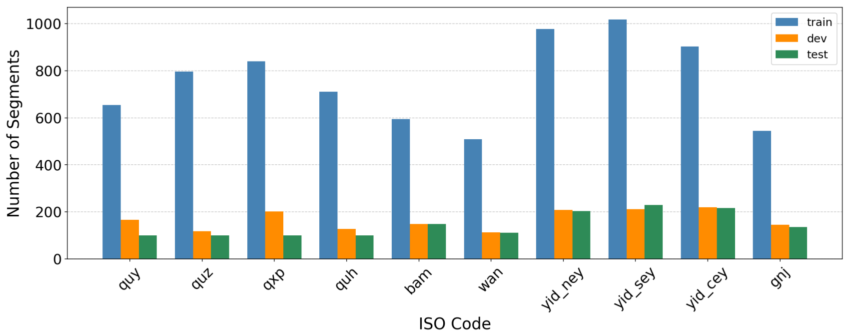


Figure 11: Segment distribution across all languages / dialects and splits

HARNESS: Lightweight Distilled Arabic Speech Foundation Models

Vrunda N. Sukhadia [1*], **Shammur Absar Chowdhury**[2]
[1]Amazon India, [2]Qatar Computing Research Institute, HBKU, Qatar
sukhadiavrunda@gmail.com, Shchowdhury@hbku.edu.qa

Abstract

Large self-supervised speech (SSL) models achieve strong downstream performance, but their size limits deployment in resource-constrained settings. We present HArnESS, an Arabic-centric self-supervised speech model family trained from scratch with iterative self-distillation, together with lightweight student variants that offer strong accuracy-efficiency trade-offs on Automatic Speech Recognition (ASR), Dialect Identification (DID), and Speech Emotion Recognition (SER). Our approach begins with a large bilingual Arabic-English teacher and progressively distills its knowledge into compressed student models while preserving Arabic-relevant acoustic and paralinguistic representations. We further study PCA-based compression of the teacher supervision signal to better match the capacity of shallow and thin students. Compared with HuBERT and XLS-R, HArnESS consistently improves performance on Arabic downstream tasks, while the compressed models remain competitive under substantial structural reduction. These results position HArnESS as a practical and accessible Arabic-centric SSL foundation for real-world speech applications.

Keywords: Self-supervised model, Distillation, Benchmark resources, Arabic downstream tasks

1. Introduction

Self-supervised learning (SSL) has transformed speech processing by learning transferable representations from large amounts of unlabeled audio. Large SSL models capture rich acoustic and linguistic structure and have shown strong performance across a wide range of speech tasks (Chen et al., 2022; Hsu et al., 2021; Baevski et al., 2022; Mohamed et al., 2022; Chung et al., 2021; wen Yang et al., 2021). These models can be used either as fixed feature extractors or fine-tuned with limited labeled data, making them especially attractive in low-resource settings.

The effectiveness of SSL models, however, depends heavily on the scale, diversity, and balance of the pretraining data. Multilingual SSL models such as XLS-R (Babu et al., 2021) have shown clear advantages for low-resource languages compared with monolingual models trained on high-resource languages such as English (Shi et al., 2023). At the same time, recent evidence suggests that multilingual models may disproportionately favor languages with greater pretraining coverage, which can limit gains for underrepresented languages (Storey et al., 2024). This motivates closer study of language-focused SSL models that better reflect the linguistic and acoustic properties of a target language.

Arabic is a particularly challenging case for speech modeling. It is spoken across 22 countries and exhibits substantial dialectal diversity, with many varieties differing in phonetics, morphology, and lexical usage. In addition, Arabic speech often includes influences from other languages, including English and French (Ali et al., 2021). This diversity makes Arabic speech processing difficult for generic multilingual models, which may not fully capture dialect-sensitive and culturally grounded speech patterns. These challenges motivate Arabic-centric SSL modeling that can better represent spoken Arabic while remaining robust to variation across regions and speaking styles.

At the same time, training and deploying language-focused SSL models remains expensive. Large-scale pretraining requires substantial compute, long training times, and broad unlabeled speech collections. These costs also make deployment difficult in practical and resource-constrained environments, where model size, memory use, and latency matter. Model compression is therefore essential for making such systems more accessible and usable.

Knowledge distillation has emerged as an effective approach for compressing large speech models while preserving much of their performance. In this setting, a smaller student model learns from a larger teacher model, leading to lower memory usage and faster inference with limited degradation in downstream quality. Prior work, including DistillHuBERT (Chang et al., 2022), FitHuBERT (Lee et al., 2022), DPHuBERT (Peng et al., 2023), SKILL (Zampierin et al., 2024), and related methods (Ashihara et al., 2022; Wang et al., 2022), has explored task-agnostic distillation for HuBERT-style models. However, such work has focused primarily on general-purpose compression, with limited attention to Arabic-centric SSL trained from scratch and systematically distilled into lightweight models.

While prior studies have applied self-supervised speech models such as HuBERT and wav2vec 2.0 to Arabic, Arabic-centric SSL remains underexplored, especially in the setting of large-scale

[*]This work was carried out at QCRI.

Proceedings of SPEAKABLE @ LREC 2026, pages 109–117
11 May 2026.

training from scratch and deployment-oriented compression. In this work, we focus on both aspects. We introduce **H**uBERT-based **Ar**abic a**n**d **E**nglish **S**elf-**S**upervised Speech (**HArnESS**), an Arabic-centric SSL model family trained from scratch on large-scale bilingual Arabic-English speech, and we study iterative self-distillation to build compact student models that retain strong performance on ASR, SER, and DID.

We adopt bilingual Arabic-English pretraining for two reasons. First, English corpora provide additional acoustic and phonetic diversity at scale, which can stabilize representation learning when Arabic resources are comparatively limited and heterogeneous. Second, Arabic speech in real-world settings often includes borrowed English words and code-switching, particularly in conversational and media domains. Our goal is therefore not to weaken Arabic-centric modeling, but to combine Arabic-focused coverage with the regularization benefits of broader bilingual pretraining.

Following the HuBERT training paradigm, we first train a large teacher model, HArnESS-L, with 24 encoder layers through iterative self-distillation. We then transfer its knowledge to smaller students, yielding HArnESS-S, a shallow variant, and HArnESS-ST, a shallow (S) and thin (T) variant. In addition, we investigate low-rank approximation of the teacher supervision signal to simplify the distillation target space and improve knowledge transfer to compact students.

We evaluate HArnESS-L, HArnESS-S, and HArnESS-ST on three downstream tasks spanning content, dialectal, and paralinguistic information, namely ASR, DID, and SER. We compare them against HuBERT-Large, trained primarily on English, and XLS-R, a multilingual SSL model (Babu et al., 2021). Our results show that Arabic-centric pretraining combined with iterative distillation provides an effective balance between task performance and model efficiency.

Our main contributions are as follows:

1. We introduce HArnESS, an Arabic-centric SSL model family trained from scratch, consisting of HArnESS-L (large), HArnESS-S (shallow), and HArnESS-ST (shallow and thin).

2. We study iterative self-distillation as a strategy for compressing Arabic-centric SSL models into lightweight deployment-oriented students.

3. We investigate compact supervision through low-rank approximation of the teacher signal and analyze its effect on student performance.

4. We benchmark the HArnESS family on ASR, DID, and SER, covering content, speaker-related, and paralinguistic speech tasks.

5. We publicly release the distilled models and benchmark resources to support future research.[1]

2. HArnESS Models

2.1. HArnESS Model

Figure 1 illustrates the HArnESS training pipeline, which follows a HuBERT-style iterative self-distillation procedure. At iteration i, we train a model $\mathbf{M}_i$ using discrete pseudo-labels produced from the previous iteration model $\mathbf{M}_{i-1}$. The core idea is masked-prediction pretraining: a subset of time frames is masked, and the model is optimized to predict the corresponding pseudo-labels.

Iterative self-distillation pipeline. Given an input utterance x, we first obtain frame-level embeddings from $\mathbf{M}_{i-1}$ and convert them into discrete targets via clustering (described below), yielding a pseudo-label sequence $z^{(i-1)} = \{z_t^{(i-1)}\}_{t=1}^{T}$ with $z_t^{(i-1)} \in \{1, \ldots, K\}$. We then train $\mathbf{M}_i$ to predict these targets from contextualized frame representations, where some frames are replaced by a mask token (or masked spans), forcing the model to use broader context.

Training regimen and compression schedule. We perform multiple iterations of refinement. In the first two iterations, we keep the model architecture unchanged to encourage progressively stronger acoustic abstractions. Starting from the third iteration, we distill and compress the model to obtain efficient variants. Specifically, we explore three compression axes: *(a)* reducing Transformer depth (d) to obtain a shallower model; *(b)* reducing model width (encoder dimension, enc_d) to obtain a thinner model; and *(c)* reducing attention capacity by decreasing the number of attention heads ($attn$). This schedule yields a teacher-like encoder in early iterations and a family of compact students in later iterations.

Model architecture. HArnESS consists of a convolutional (CNN) feature extractor followed by a stack of Transformer encoder layers. Similar to HuBERT encoders, the CNN front-end comprises 7 temporal convolution layers that transform raw audio into latent frame features. The Transformer encoder contains d layers with hidden dimension, emb$_d$. Each layer includes multi-head self-attention (MHA) with $attn$ heads and a position-wise feed-forward network (FFN). A linear prediction head

[1] https://huggingface.co/QCRI/distillHarness

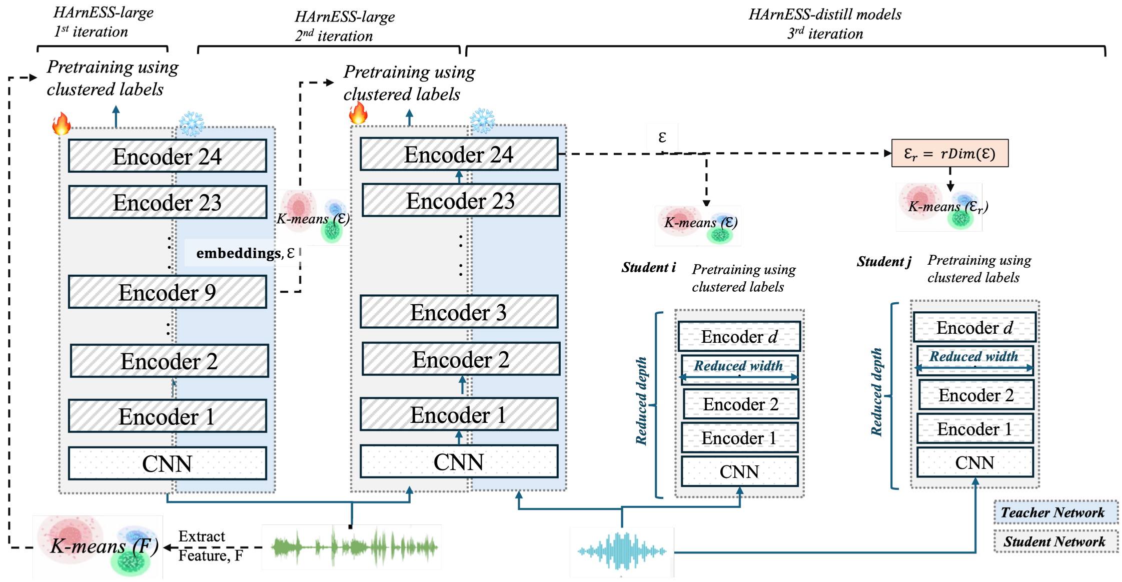


Figure 1: Overview of the iterative self-distillation framework used to build the HArnESS model family.

maps contextualized frame representations to a categorical distribution over K cluster IDs.

Training objective. HArnESS is trained with a HuBERT-style masked prediction objective. For each utterance, we mask a subset of time frames (or spans) in the input representation and train $\mathbf{M}_i$ to predict the corresponding discrete pseudo-labels generated from $\mathbf{M}_{i-1}$. We use standard cross-entropy classification over the cluster IDs. To improve stability, we compute the loss on both masked and unmasked frames and combine them with a fixed weighting: the masked-frame loss encourages contextual reasoning over surrounding speech, while the unmasked-frame loss provides an additional learning signal that helps prevent training collapse and improves convergence.

Pseudo-label generation. To obtain discrete targets for iteration i, we extract frame-level embeddings from the previous model $\mathbf{M}_{i-1}$ and cluster them using K-means, producing a pseudo-label sequence $z^{(i-1)} = \{z_t^{(i-1)}\}_{t=1}^{T}$ with $z_t^{(i-1)} \in \{1, \ldots, K\}$. Unless otherwise stated, we use last-layer embeddings, which provide the most abstract and stable representations.[2]

PCA for supervision signal compression. We also investigate PCA as a way to compress the teacher supervision signal before clustering. Instead of clustering the full teacher embedding, we optionally project it to a lower-dimensional space and then derive pseudo-labels from the projected representation. Concretely, given an embedding vector $h_t \in R^D$ from $\mathbf{M}_{i-1}$, we apply PCA to obtain a compressed representation $\tilde{h}_t \in R^{D'}$ $(D' \ll D)$ and then cluster $\tilde{h}_t$.

This serves two purposes. First, dimensionality reduction can remove noisy or redundant directions, which may improve clustering robustness. Second, it produces a simpler target space that better matches the capacity of compressed students, especially when model width (enc_d) is reduced. In this sense, PCA does not compress the student input directly; rather, it simplifies the discrete targets used during distillation.

Initialization and iteration-specific supervision. For the initial iteration $i = 1$, we bootstrap pseudo-labels by extracting MFCC features from raw speech x and clustering them to obtain $z^{(0)}$. For the next iteration ($i = 2$), we generate pseudo-labels from intermediate representations of $\mathbf{M}_0$, using the 9th Transformer layer embeddings for clustering. For all subsequent iterations ($i \geq 3$), we generate pseudo-labels using last-layer embeddings from $\mathbf{M}_{i-1}$, which provide the most abstract and stable representations.

For training $\mathbf{M}_i$ $(i \geq 1)$, we explore two weight initialization strategies: *(a)* random initialization (uniformly sampled weights), and *(b)* blocked-averaging initialization, where groups of student layers are initialized by averaging corresponding blocks of layers from $\mathbf{M}_{i-1}$. Blocked averaging provides a smoother transition across iterations and often improves stability, particularly when compressing depth/width in later iterations.

[2]We also explored averaged embeddings from selected layers and observed no consistent gains.

3. Experimental Setups

3.1. Pre-training Data

Iterations 1–2 (bilingual pretraining). We pretrain HArnESS on a mixture of publicly available Arabic and English speech corpora, including QASR (Mubarak et al., 2021), MGB3 (Ali et al., 2017), LibriSpeech (Panayotov et al., 2015), Common Voice (Arabic/English) (Ardila et al., 2020), and GigaSpeech (Chen et al., 2021), among others. The base pretraining pool consists of approximately **4K hours of Arabic** and **3.56K hours of English** speech. We further expand the training data through augmentation to reach approximately **23K hours** in total. Of this augmented portion, around **300 hours** come from additive background-noise augmentation, while the remainder is primarily produced through SpecAugment-based transformations.

To improve dialectal and cultural coverage, we also incorporate spoken content from 15 Arabic-speaking countries crawled from YouTube, covering diverse Arabic dialects. We provide a coarse dialect breakdown of the Arabic data by major region in Table 1, grouping samples into MSA, Gulf, Levantine, Egyptian, Maghrebi, mixed, and unlabeled categories where exact dialect labels are unavailable. All official development and test partitions are excluded from pretraining to avoid data leakage.

Category	Sub-category / Dialect	Duration (Hrs)
Original Clean Data		**7,566.00**
	English Subset	3,565.00
	Arabic Subset	4,001.00
	MSA / General Arabic	3,603.28
	Levantine	107.69
	Egyptian	109.20
	Gulf	77.13
	Maghrebi	69.11
	Other	34.59
Augmented Data		**15,434.00**
	Speed Perturbation ($0.9\times, 1.1\times$)	15,134.00
	Noise Augmentation (Arabic)	300.00
Total Training Volume		**23,000.00**

Table 1: Comprehensive breakdown of the 23,000-hour training corpus, including language distribution, dialectal variety, and augmentation strategies.

Iteration 3 (Arabic-only distillation). Our primary goal is to obtain lightweight Arabic-centric models. Accordingly, for the distillation/compression iteration, we use approximately **1,100 hours** of Arabic speech drawn from the QASR training data. For K-means training in this phase, we randomly sample **30%** of the iteration-3 data (approximately **300 hours**) to reduce clustering cost while maintaining linguistic diversity.

3.2. Downstream Tasks and Data

Benchmarking SSL speech encoders for English is supported by standardized suites such as SUPERB (wen Yang et al., 2021). In contrast, Arabic speech lacks an analogous standardized benchmark. To address this gap, we evaluate HArnESS across three representative Arabic tasks: **ASR** (content recognition), **dialect identification (DID)** (speaker information), and **speaker emotion recognition (SER)** (paralinguistic analysis).
ASR. We fine-tune on a **300-hour** subset of QASR and evaluate on the MGB2 (Ali et al., 2019) test set. To assess out-of-domain generalization, we additionally report performance on the MGB3 test set.
SER. We use KSUEmotion (Meftah et al., 2021), collected from 23 speakers with six emotion classes. The dataset is split into train (3.30 h), dev (0.83 h), and test (1.0 h).[3]
DID. We use the ADI5 dataset with five region-based dialect classes (MSA, Egyptian, Levantine, North African, and Gulf) and the official train/dev/test splits.
Metrics. We report word error rate (WER) for ASR and classification accuracy (Acc) for DID and SER.

3.3. Pre-training Hyperparameters

We train HArnESS using the `fairseq` codebase (Ott et al., 2019). Table 2 summarizes the key hyperparameters. Unless stated otherwise, we only change the supervision source and model capacity (Table 3).

3.4. Pre-training Procedure

Model configurations for the upstream encoders are summarized in Table 3.

Iterations 1–2 (HArnESS-L). For the first two iterations, we train the large model (**HArnESS-L**; 24 Transformer layers) on the 23k-hour bilingual mixture. Iteration 1 is trained for 500k steps and iteration 2 for 700k steps. For iteration 1 pseudo-labels, we cluster 39-dimensional MFCC features with $K{=}1000$ clusters. For iteration 2 pseudo-labels, we extract latent representations from the **9th Transformer layer** of the iteration-1 model and cluster them with $K{=}1000$ clusters to obtain refined targets.

Iteration 3 (compressed students). For iteration $i{=}3$, we train compressed models using **HArnESS-S** and **HArnESS-ST**, both with a 4-layer Transformer encoder and reduced capacity (Table 3).

[3]We will release our split for reproducibility.

Total pre-training audio (Iter 1–2)	23k hours (Arabic/English ≈ balanced)
K-means training subset (Iter 1–2)	300 hours
Distillation audio (Iter 3)	1,100 hours (Arabic-only)
K-means training subset (Iter 3)	30% ≈ 300 hours
# clusters (K)	1000
Feature type for i=0 targets	MFCC (39-dim)
Embedding layer for i=1 targets	Layer 9 embeddings of $\mathbf{M}_0$
Embedding layer for $i \geq 2$ targets	Last-layer embeddings of $\mathbf{M}_{i-1}$
Iteration 1 steps / GPUs / batch	500k / 24×H100 / 62.5s audio per GPU
Iteration 2 steps / GPUs / batch	700k / 24×H100 / 62.5s audio per GPU
Iteration 3 steps / GPUs / batch	300k / 8×H100 / 75s audio per GPU
Mask probability (p_{mask})	0.80
Mask span length (frames)	10
PCA dimension (D'; when enabled)	512

Table 2: Key pre-training hyperparameters.

Models	XR	HuL	H-L	H-S	H-ST	H-ST (PCA)
Supervision	–	–	$L9_{\text{emb}}$ ($i = 1$)	$L23_{\text{emb}}$ ($i = 2$)		PCA($L23_{\text{emb}}$) ($i = 2$)
CNN Encoder						
Strides	5, 2, 2, 2, 2, 2, 2					
Kernel Width	10, 3, 3, 3, 3, 2, 2					
Channels	512					
Transformer						
Depth (l)	24	24	24	4	4	4
Emb. Dim (emb$_d$)	1024	1024	1024	1024	512	512
FFN Dim (d_{ffn})	4096	4096	4096	2048	2048	2048
Attn. Heads (h_{attn})	16	16	16	16	16	16
Projection						
Dim. (d_p)	768	768	768	768	768	768
Params						
in M	300	316	316	65	28	28

Table 3: SSL Model Architecture Comparison. XR: XLS-R, HuL: HuBERT-Large, H-L: HArnESS-Large, H-S: HArnESS-Shallow, H-ST: HArnESS-Shallow and Thin. Dim. dimension, Emb.: Embedding. $L{*}_{emb}$: Embedding from layer * (e.g. 23) of model from iteration i.

Pseudo-labels are generated by clustering **last-layer** embeddings from the iteration-2 HArnESS-L teacher with K=1000 clusters. Thus, HArnESS-S and HArnESS-ST correspond to the third-iteration distilled models trained on 1,100 hours of Arabic speech.

3.5. Downstream Training

For downstream evaluation, we keep the SSL encoder frozen and use it strictly as a feature extractor. We obtain frame-level representations from all Transformer layers, average them to form an utterance-level representation, and train task-specific downstream models on top of these fixed features only. No gradients are propagated into the SSL encoder during downstream training.

3.5.1. DID and SER Architecture

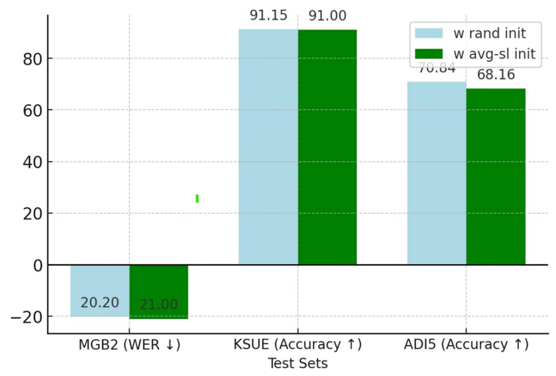
For DID and SER, we train a lightweight classifier on top of the frozen SSL features with a batch size of 4 for 10k steps. The classifier consists of three temporal convolution layers with kernel size 5, ReLU activations, and dropout of 0.4, followed by self-attention pooling, a feed-forward layer, and a final softmax layer. All hidden dimensions are set to 80. This setup isolates the quality of the learned SSL representations by keeping the encoder fixed and limiting the trainable parameters to the downstream classifier.

3.5.2. ASR Architecture

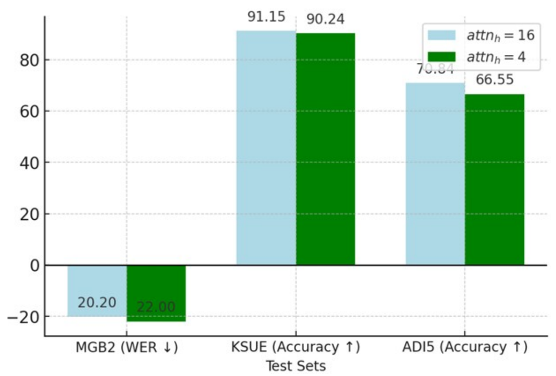
For ASR, we train an encoder–decoder model with a joint CTC/attention objective using the ESPnet toolkit.[4] The encoder consists of two Conformer layers and the decoder consists of two Transformer layers, each with 8 attention heads and 2048 linear units. We train for 70 epochs.

[4]Using ESPnet toolkit.

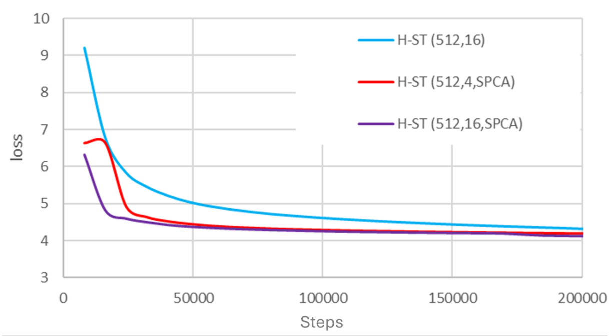
4. Results



(a) Weight Initialization



(b) Number of attention heads



(c) Effect of dimension reduction

Figure 2: Ablation results for the compressed student models. a) Effect of student initialization strategy. (b) Effect of reducing the number of attention heads. (c) Effect of applying PCA to teacher embeddings before clustering when generating pseudo-labels for student training. H-ST(emb_d, $attn_h$), SPA means PCA applied for supervision.

Comparison with Upper Bound: SOTA Model As contextual reference, Table 4 also reports representative published results from strong task-specific systems for Arabic ASR, DID, and SER. These results come from separate studies and are not directly comparable to our models because they differ in architecture, supervision, training data, and experimental protocol. We therefore use them only to contextualize performance, not to claim direct state-of-the-art results. For ASR, we report results alongside Fanar ASR (Fanar et al., 2025), a specialized system trained on more than 10K hours of MSA and dialectal Arabic speech. Under our much more constrained setup, where ASR fine-tuning uses only 300 hours of MSA data, HArnESS-L remains within about 5 WER points on MGB2 and MGB3, and HArnESS-S within about 10 points. For DID and SER, HArnESS-L also shows strong results relative to the published reference numbers, achieving 84.98% on ADI5 compared with 82.5% (Kulkarni and Aldarmaki, 2023), and 94.66% on KSUEmotion compared with 85.53% (on ResNet-based architecture). Although these comparisons are only approximate, they indicate that HArnESS-L is competitive with strong specialized systems, while the distilled students preserve much of the teacher's performance with substantially smaller capacity.

HArnESS-L *vs.* Existing SSLs for Arabic Compared with HuBERT-L and XLS-R, HArnESS-L performs better across the evaluated Arabic tasks, suggesting that Arabic-centric pretraining is beneficial for downstream Arabic speech processing. The compressed HArnESS variants also outperform the multilingual XLS-R baseline on several tasks, indicating that iterative distillation preserves useful task-relevant structure even under heavy compression.

Effects of structural compression and design choices. Figure 2 presents three ablation studies on student design, covering initialization, structural compression, and supervision compression. For iteration $i = 3$, we first examined the effect of student weight initialization and observed only minor differences in downstream performance (Figure 2). This indicates that initialization plays a limited role at this stage, and that performance depends more strongly on the distilled supervision signal.

We then evaluated the effect of reducing model depth. HArnESS-S achieves 79.4% structural compression relative to HArnESS-L while maintaining strong performance across tasks. Despite this compression, it still outperforms the multilingual and English SSL baselines, highlighting the effectiveness of Arabic-centric distillation. Compared with HArnESS-L, however, HArnESS-S shows a 4.7 absolute increase in WER, a 3.51-point drop in SER accuracy, and a 14.4-point drop in DID accuracy. The larger degradation on DID suggests that dialect-related cues are harder to preserve in shallower architectures.

Next, we reduced the number of attention heads from HArnESS-S (*attn* $= 16$) to HArnESS-S*

Models	ASR (WER ↓)		SER (Acc ↑)	DID (Acc ↑)
	MGB2	MGB3	KSUEmotion	ADI5
Published task-specific reference results (context only)				
Reference result	10.24 (Fanar et al., 2025)	21.31 (Fanar et al., 2025)	85.53% (Abouzeid et al., 2025)	82.5% (Kulkarni and Aldarmaki, 2023)
Our downstream evaluation with frozen SSL encoders				
HuBERT-L (*English*)	22.6*	51.2*	91.92%	64.14%
XLS-R (*Multilingual*)	22.60*	51.80*	73.32%	42.35%
HArnESS-L (*Bilingual: Arabic-English*)	**15.50***	**41.60***	**94.66%**	**84.98%**
Compressed HArnESS students distilled with ≈1000h Arabic-only data				
HArnESS-S ($\Delta S = 79.4\%$)	*20.20**	*52.80**	*91.15%*	*70.84%*
HArnESS-ST ($\Delta S = 93.7\%$)	23.20*	58.20*	89.02%	69.77%
HArnESS-ST$^{\equiv}$ ($\Delta S = 93.7\%$)	22.50*	55.60*	87.34%	61.64%

Table 4: Performance comparison on ASR, SER, and DID. ASR downstream results in our setup are obtained by training on a **300h QASR** subset, results denoted by a *. The top block lists previously published task-specific reference results from separate studies and is included only for context. These numbers are not from a unified baseline and are not directly comparable to the models evaluated in our setup. L denotes the large teacher model, S the shallow student, and ST the shallow-thin student. ΔS denotes overall structural compression relative to HArnESS-L.

Test Sets	emb_d=1024	emb_d=512	emb_d=256
MGB2 (WER ↓)	20.2	23.20	22.3
KSUEmotion (Acc ↑)	91.15%	89.02%	79.42%
ADI5 (Acc ↑)	70.84%	69.77%	53.41%
ΔS	70.43%	91.14%	96.52%

Table 5: Performance Comparison for different embedding dimensions. ΔS: Overall structural compression.

(*attn* $= 4$), which yields an additional 26.15% structural compression, reducing the model from 65M to 48M parameters. This change has only a limited effect on ASR and SER, but causes a larger drop on DID (Figure 2), again indicating that dialect-sensitive information is more susceptible to architectural compression.

Finally, we examined embedding-dimension reduction (Table 5). At extreme compression ($\Delta S = 96.52\%$ relative to HArnESS-L), performance drops sharply across tasks. This result suggests that overly aggressive dimensionality reduction weakens the representational capacity of the student and substantially limits downstream performance.

Effect of compressing the supervision signal. We also study whether simplifying the teacher supervision signal improves student training. Specifically, in iteration $i = 3$, we compare knowledge distillation with and without applying PCA to the teacher embeddings before clustering. As shown in Figure 2c, supervision derived from PCA-reduced embeddings converges faster than supervision from the original embeddings. This suggests that reducing redundancy in the teacher feature space produces a cleaner supervision signal, leading to more stable and efficient optimization while preserving effective knowledge transfer.

5. Conclusion

In this work, we introduced HArnESS, an Arabic-centric self-supervised speech model family designed to better capture the diversity of Arabic dialectal speech. Using an iterative self-distillation framework, we transferred knowledge from a large bilingual teacher model to compact shallow and shallow-thin student models while preserving Arabic-relevant speech representations. Experiments on Arabic ASR, SER, and DID show that HArnESS is competitive with, and in some cases stronger than, multilingual baselines such as HuBERT and XLS-R. The compressed HArnESS variants further offer an attractive efficiency-performance trade-off, making them suitable for more resource-constrained settings. Our downstream evaluation relies on frozen encoders, providing a controlled assessment of representation quality, but it does not fully reflect the gains that may emerge under end-to-end fine-tuning. Future work will extend the comparison to fine-tuned settings and broader baselines. We will publicly release the lightweight models and benchmarking resources to facilitate future research.

6. Bibliographical References

Ali Abouzeid, Bilal Elbouardi, Mohamed Maged, and Shady Shehata. 2025. Arabemonet: A lightweight hybrid 2d cnn-bilstm model with attention for robust arabic speech emotion recognition.

Ahmed Ali, Peter Bell, James Glass, Yacine Messaoui, Hamdy Mubarak, Steve Renals, and Yifan Zhang. 2019. The mgb-2 challenge: Arabic multi-dialect broadcast media recognition.

Ahmed Ali, Shammur Chowdhury, Mohamed Afify, Wassim El-Hajj, Hazem Hajj, Mourad Abbas, Amir Hussein, Nada Ghneim, Mohammad Abushariah, and Assal Alqudah. 2021. Connecting Arabs: bridging the gap in dialectal speech recognition. *Communications of the ACM*, pages 124–129.

Takanori Ashihara, Takafumi Moriya, Kohei Matsuura, and Tomohiro Tanaka. 2022. Deep versus wide: An analysis of student architectures for task-agnostic knowledge distillation of self-supervised speech models. In *23rd Annual Conference of the International Speech Communication Association, INTERSPEECH 2022*.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. In *Proceedings of Interspeech*.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR.

Heng-Jui Chang, Shu-wen Yang, and Hung-yi Lee. 2022. Distilhubert: Speech representation learning by layer-wise distillation of hidden-unit bert. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7087–7091. IEEE.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, pages 1505–1518.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250. IEEE.

Fanar, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, Majd Hawasly, Mus'ab Husaini, Soon-Gyo Jung, Ji Kim Lucas, Walid Magdy, Safa Messaoud, Abubakr Mohamed, Tasnim Mohiuddin, Basel Mousi, Hamdy Mubarak, Ahmad Musleh, Zan Naeem, Mourad Ouzzani, Dorde Popovic, Amin Sadeghi, Husrev Taha Sencar, Mohammed Shinoy, Omar Sinan, Yifan Zhang, Ahmed Ali, Yassine El Kheir, Xiaosong Ma, and Chaoyi Ruan. 2025. Fanar: An Arabic-Centric Multimodal Generative AI Platform. *arxXiv:2501.13944*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, pages 3451–3460.

Ajinkya Kulkarni and Hanan Aldarmaki. 2023. Yet another model for Arabic dialect identification. In *Proceedings of ArabicNLP 2023*, pages 435–440, Singapore (Hybrid). Association for Computational Linguistics.

Yeonghyeon Lee, KANGWOOK JANG, Jahyun Goo, Youngmoon Jung, and Hoi-Rin Kim. 2022. Fithubert: Going thinner and deeper for knowledge distillation of speech self-supervised learning. In *23rd Annual Conference of the International Speech Communication Association, INTERSPEECH 2022*, pages 3588–3592. ISCA.

Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, et al. 2022. Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1179–1210.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Yifan Peng, Yui Sudo, Shakeel Muhammad, and Shinji Watanabe. 2023. Dphubert: Joint distillation and pruning of self-supervised speech models. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2023, pages 62–66.

Jiatong Shi, Dan Berrebbi, William Chen, Ho-Lam Chung, En-Pei Hu, Wei Ping Huang, Xuankai Chang, Shang-Wen Li, Abdelrahman Mohamed, Hung yi Lee, and Shinji Watanabe. 2023. Ml-superb: Multilingual speech universal performance benchmark.

Edward Storey, Naomi Harte, and Peter Bell. 2024. Language bias in self-supervised learning for automatic speech recognition. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 37–42. IEEE.

Rui Wang, Qibing Bai, Junyi Ao, Long Zhou, Zhixiang Xiong, Zhihua Wei, Yu Zhang, Tom Ko, and Haizhou Li. 2022. Lighthubert: Lightweight and configurable speech representation learning with once-for-all hidden-unit bert. *Proc. Interspeech 2022*, pages 1686–1690.

Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. 2021. Superb: Speech processing universal performance benchmark.

Luca Zampierin, Ghouthi Boukli Hacene, Bac Nguyen, and Mirco Ravanelli. 2024. Skill: Similarity-aware knowledge distillation for speech self-supervised learning. *arXiv preprint arXiv:2402.16830*.

7. Language Resource References

Ahmed Ali, Stephan Vogel, and Steve Renals. 2017. Speech recognition challenge in the wild: Arabic MGB-3. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 316–322. IEEE.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Yujun Wang, Zhao You, and Zhiyong Yan. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio.

Ali Hamid Meftah, Mustafa A. Qamhan, Yasser Seddiq, Yousef A. Alotaibi, and Sid Ahmed Selouani. 2021. King saud university emotions corpus: Construction, analysis, evaluation, and comparison. *IEEE Access*, 9:54201–54219.

Hamdy Mubarak, Amir Hussein, Shammur Absar Chowdhury, and Ahmed Ali. 2021. QASR: QCRI Aljazeera Speech Resource. A Large Scale Annotated Arabic Speech Corpus. In *Proc. of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2274–2285, Online. Association for Computational Linguistics.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210.

When Does OmniASR Fail? A Fine-Grained Human Evaluation on Saudi Arabic Dialects

Hend Al-Khalifa

College of Computer and Information Sciences, King Saud University
Riyadh, Saudi Arabia
hendk@ksu.edu.sa

Abstract

Automatic Speech Recognition (ASR) evaluation has traditionally relied on Word Error Rate (WER), a metric that treats all errors equally and obscures critical failure modes. In this paper, we present a fine-grained human evaluation of Meta's recently released OmniASR system on Saudi Arabic dialects using the SADA dataset. Three trained annotators evaluated 103 audio samples, producing 264 annotations across two dimensions (comprehensibility and naturalness) while categorizing errors using a novel 10-category Arabic-specific error taxonomy. OmniASR achieved a mean WER of 42.2% and mean comprehensibility of 3.62/5, but exhibited a polarized performance pattern: 32.6% of transcriptions achieved perfect scores while 21.2% were essentially unusable. Error analysis reveals that hallucinations and deletions have the greatest negative impact on comprehensibility (−1.64 and −1.57 points respectively), roughly 6× more damaging than named entity errors. WER correlates with human comprehensibility ratings at a level comparable to inter-annotator agreement (r = −0.679 vs. pairwise annotator r = 0.61–0.67), but lacks the diagnostic granularity to reveal which error types drive quality degradation. These findings motivate the use of error-type-aware evaluation frameworks that complement WER with fine-grained analysis for Arabic ASR systems.

Keywords: Arabic ASR, OmniASR, Human Evaluation, Error Taxonomy, Saudi Dialects, SADA Dataset, Diglossia, WER

1. Introduction

Word Error Rate (WER) has been the dominant metric for evaluating Automatic Speech Recognition systems for decades. However, WER has significant limitations: it treats all errors equally, fails to capture semantic impact, and can obscure systematic failure patterns. These limitations are particularly pronounced for morphologically rich, dialectally diverse languages like Arabic, where a single morphological error may completely change the meaning of an utterance, while other errors may be cosmetic.

In November 2025, Meta released OmniASR (Keren et al., 2025), a multilingual ASR system supporting over 1,600 languages through a 7B-parameter Wav2vec 2.0 encoder combined with an LLM-based decoder. The system was designed for extensibility, enabling communities to add new languages with minimal data. While this represents a significant advancement in language coverage, the quality of transcription for individual languages, particularly Arabic dialects, remains understudied. The OmniASR paper reports aggregate Character Error Rates (CER) but provides limited analysis of error types or their perceptual impact on users.

Arabic presents unique challenges for ASR that go beyond those faced by well-resourced languages like English. The language exhibits diglossia (the coexistence of Modern Standard Arabic and regional spoken dialects), rich morphology with complex clitic attachment, and culturally sensitive content (e.g., religious phrases) that requires accurate transcription. These challenges motivate the need for evaluation frameworks that go beyond aggregate metrics.

This paper presents the first systematic human evaluation of OmniASR on Saudi Arabic dialects. We make three contributions: (1) a 10-category Arabic-specific error taxonomy that captures linguistically and culturally significant error types beyond what WER measures; (2) a human evaluation study with 264 annotations from 3 trained annotators on 103 audio samples from the SADA dataset (Alharbi et al., 2024); and (3) empirical findings demonstrating that error types have dramatically different impacts on perceived transcription quality, and that while WER correlates with human judgments at a level comparable to inter-annotator agreement, it cannot reveal which error types drive quality degradation.

The rest of the paper is organized as follows. Section 2 provides background on Arabic language characteristics relevant to ASR, including diglossia, morphological complexity, and code-switching. Section 3 reviews related work on Arabic ASR datasets, multilingual models, and evaluation methodologies. Section 4 describes our methodology, including the data sample, error taxonomy, annotation protocol, and ethical considerations. Section 5 presents our results on overall performance, error distribution, error impact analysis, inter-annotator agreement, and WER/CER analysis. Section 6 discusses the implications of our findings, and Section 7 concludes the paper.

Proceedings of SPEAKABLE @ LREC 2026, pages 118–124

2. Background

2.1 Arabic Diglossia and Dialectal Variation

Arabic is spoken by over 400 million people across more than 20 countries, making it the fifth most spoken language globally (Belinkov et al., 2019). However, 'Arabic' is not a single homogeneous language but rather a collection of varieties that exist in a state of diglossia (Ferguson, 1959). Native speakers acquire a regional spoken dialect as their first language, while Modern Standard Arabic (MSA), the standardized variety used in formal writing, news broadcasts, and official communication, is learned later through education (Badawi, 1973; Holes, 2004).

The differences between MSA and spoken dialects extend across all linguistic levels: phonology, morphology, syntax, and lexicon (Saiegh-Haddad, 2018). For example, the MSA word 'أريد' (ʔurīd, 'I want') becomes 'عايز' (ʕāyiz) in Egyptian, 'بدي' (biddi) in Levantine, and 'أبغى' (abġa) in Gulf Arabic. These are not merely pronunciation differences but entirely distinct lexical items. Such variation poses fundamental challenges for ASR systems, which must decide whether to transcribe what was actually said (dialect) or normalize to MSA, and the 'correct' choice depends on the use case.

Saudi Arabia alone encompasses multiple dialect groups: Najdi (central), Hijazi (western), and Gulf/Khaleeji (eastern), each with distinct phonological and lexical features. The SADA dataset we use in this study captures this diversity through content from Saudi television programs.

2.2 Arabic Morphological Complexity

Arabic morphology is based on a root-and-pattern system where most words derive from a triconsonantal root (e.g., k-t-b for writing-related words) combined with vowel patterns and affixes to create meaning (Boudelaa & Marslen-Wilson, 2013). This creates a high degree of morphological productivity but also ambiguity. Additionally, Arabic exhibits extensive use of clitics (attached pronouns, conjunctions, and prepositions that combine with base words). For example, 'وكتابي' (wa-kitāb-ī) encodes 'and my book' in a single orthographic unit.

For ASR, this morphological richness creates challenges at multiple levels. Segmentation errors (incorrect word boundaries) can create or destroy meaning. Clitic attachment errors change grammatical relationships. Dialectal morphological variants (e.g., different verb conjugation patterns) may be transcribed as MSA equivalents, altering the register and authenticity of the output. These considerations motivated several categories in our error taxonomy.

2.3 Arabic Code-Switching and Religious Language

Modern Arabic speech frequently involves code-switching with English, particularly in technical, business, and youth-oriented contexts. Terms like 'update,' 'email,' and 'meeting' are commonly inserted into Arabic discourse with varying degrees of phonological adaptation. ASR systems must decide whether to preserve these as English or attempt Arabic transliteration; both choices can create errors.

Additionally, religious phrases permeate everyday Arabic speech: greetings ('السلام عليكم', as-salāmu ʕalaykum), expressions of intent (' إن شاء الله', in shāʔ Allāh, 'God willing'), and responses to news ('الحمد لله', al-ḥamdu li-llāh, 'praise be to God'). These phrases carry cultural and religious significance, and errors in their transcription, particularly truncation or misspelling, can be perceived as disrespectful. This motivated our inclusion of 'Religious Phrases' as a distinct error category.

3. Related Work

Arabic ASR has benefited from several large-scale datasets in recent years. The MGB Challenge series introduced broadcast speech corpora covering multiple dialects (Ali et al., 2016), while the SADA dataset provides 668 hours of transcribed Saudi television content across Najdi, Hijazi, and Gulf dialects, with rich metadata on speaker characteristics (Alharbi et al., 2024). Benchmarking efforts such as the Open Universal Arabic ASR Leaderboard (Wang et al., 2024) have evaluated models including Whisper, MMS, and Wav2vec 2.0 variants, consistently finding that performance varies significantly across dialects, with MSA typically achieving lower error rates than dialectal varieties.

The most recent advance in multilingual ASR is OmniASR (Keren et al., 2025), which extends coverage to over 1,600 languages through scaled self-supervised pretraining with a 7B-parameter Wav2vec 2.0 encoder combined with an LLM-inspired decoder. The system achieves character error rates below 10% for 78% of supported languages and introduces zero-shot capabilities for adding new languages with minimal paired examples. However, its evaluation focuses primarily on aggregate CER metrics across languages rather than detailed error analysis for specific language families like Arabic.

This reliance on aggregate metrics reflects a broader limitation in ASR evaluation. Researchers have long recognized that WER treats all errors equally regardless of their semantic impact. Alternative approaches include Semantic Error Rate, which attempts to weight errors by importance (McCowan et al., 2004), and human evaluation frameworks from machine translation that separate adequacy from fluency (Koehn & Monz, 2006), an approach we adapt here as comprehensibility and naturalness. More recently, calls for 'slice-aware' metrics have emphasized the need to reveal performance variation across demographic groups and linguistic conditions (Tatman, 2017). Despite these advances, fine-grained error taxonomies for Arabic ASR remain underexplored, with most evaluations continuing to rely solely on WER or CER. Our work addresses this gap by introducing an Arabic-specific error taxonomy and correlating error types with human quality judgments.

4. Methodology

4.1 Data

We used the SADA dataset (Alharbi et al., 2024), which contains 668 hours of transcribed Arabic audio from Saudi television programming, covering Najdi, Hijazi, and Gulf dialects with metadata on speaker characteristics. From this dataset, we sampled 103 audio files for human evaluation, stratified by dialect to ensure representation across Arabic varieties. Our annotated sample included Najdi (57.3%), Gulf/Khaliji (11.7%), Hijazi (10.7%), Egyptian (6.8%), Levantine (2.9%), and other varieties including MSA, Sudanese, Tunisian, and Saidi (10.6%). Audio files ranged from short utterances (4 words) to longer segments (up to 19 words), with an average length of approximately 10 words per utterance.

4.2 ASR System

We used Meta's OmniASR system via the HuggingFace Spaces interface (facebook/omniasr-transcriptions). OmniASR employs a 7B-parameter Wav2vec 2.0 encoder pretrained on massive multilingual speech data, combined with an LLM-based decoder that leverages language model priors for improved transcription (Keren et al., 2025). We used the default inference settings without language conditioning, simulating a realistic zero-shot usage scenario where users may not specify the dialect.

4.3 Error Taxonomy

We developed a 10-category error taxonomy specifically designed for Arabic ASR evaluation, drawing on linguistic literature on Arabic diglossia, morphology, and sociolinguistic variation. The taxonomy captures both general ASR errors (deletions, insertions, substitutions) and Arabic-specific phenomena (dialect-MSA substitution, religious phrases, and morphological errors). Table 1 presents the complete taxonomy with examples.

Error Type	Description	Example	Transliteration
Hallucination	Words inserted not spoken	العرض متاح → العرض متاح مجانا	*'available' → 'available for free'*
Deletion	Spoken words omitted	لا أوافق → أوافق	*'I do NOT agree' → 'I agree'*
Substitution	Semantic word replacement	التذكرة صالحة → التذكرة صعبة	*'ticket is valid' → 'ticket is difficult'*
Segmentation	Incorrect word boundaries	بالرغم → بال رغم	*'despite' incorrectly split*
Morphology	Clitic/conjugation errors	كتابي → كتاب	*'my book' → 'book'*
Named Entities	Names, places, orgs	الرياض → الرواد	*'Riyadh' → 'al-Rawad'*
Religious Phrases	Islamic expressions	إن شاء الله → إنشاء الله	*'God willing' → misspelled*
Dialect↔MSA	Register substitution	عايز → أريد	*Egyptian dialect → MSA*
Code-switching	Foreign word errors	GitHub → جتهاب	*English → Transliterated Arabic*
Numbers	Digits, dates, prices	١٢ مارس → ٢١ مارس	*'March 12' → 'March 21'*

Table 1: Arabic ASR Error Taxonomy. Arrow (→) indicates ground truth → erroneous output.

4.4 Annotation Protocol

Three senior undergraduate IT students who had completed an NLP course and are native Arabic speakers were trained as annotators. Training included detailed Arabic-language guidelines with the error taxonomy, examples, rating scale anchors, and practice items with feedback. For each audio-transcription pair, annotators: (1) listened to the audio without viewing the transcription; (2) reviewed the OmniASR transcription; (3) rated Comprehensibility (1-5): How easily can the transcription be understood? (4) rated Naturalness (1-5): How natural does the Arabic text sound? (5) selected all applicable error types from the taxonomy. Annotation was conducted using Google Forms, with 62 audio files evaluated by all three annotators to enable inter-annotator agreement analysis.

5. Results

5.1 Overall Performance

OmniASR achieved moderate overall performance on our sample: mean comprehensibility of 3.62/5 (σ=1.35) and mean naturalness of 4.05/5 (σ=1.26). However, the distribution exhibits a polarized pattern with elevated frequency at both extremes (Figure 1). Perfect scores (5/5 on both dimensions) were achieved on 32.6% of transcriptions, while 21.2% scored ≤2 on comprehensibility, essentially unusable for any downstream application. Although the middle range (scores 2–4) accounts for 54% of ratings, the proportion of extreme scores is notably higher than expected under a normal distribution. This polarized pattern suggests OmniASR either succeeds well or fails catastrophically, rather than degrading gradually.

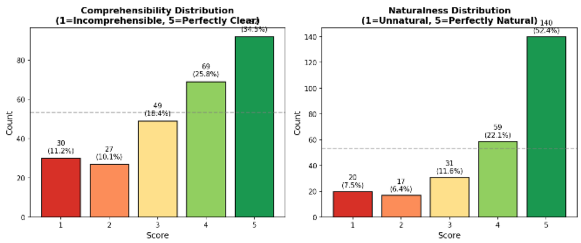


Figure 1: Distribution of comprehensibility and naturalness scores across 264 evaluations.

5.2 Error Type Distribution

Figure 2 shows the distribution of error types across 442 total error instances. The most common errors were semantic substitutions (22.9%), deletions (22.4%), hallucinations (18.3%), and segmentation errors (16.5%). Morphological errors accounted for 11.1%. Arabic-specific error types were relatively rare: named entities (5.4%), religious phrases (1.4%), dialect↔MSA substitution (0.7%), numbers (0.7%), and code-switching (0.5%).

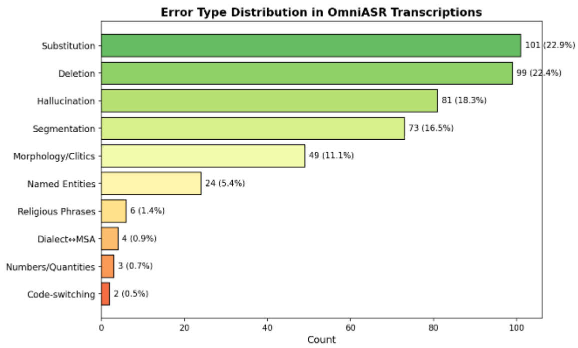


Figure 2: Distribution of error types in OmniASR transcriptions (n=442 error instances).

5.3 Error Impact Analysis

Critically, not all errors affect comprehensibility equally (Figure 3). To quantify impact, we computed the difference in mean comprehensibility between transcriptions where each error type was present versus absent. Because our data include repeated judgments by individual raters on multiple audio files, these mean-difference estimates should be interpreted as descriptive rather than inferential; a mixed-effects regression with random intercepts for annotator and audio file would be needed for formal hypothesis testing. With this caveat, hallucinations show the greatest negative impact (−1.64 points when present vs. absent), followed by deletions (−1.57), segmentation errors (−1.33), and substitutions (−1.18). Notably, named entity errors have minimal impact on human comprehensibility ratings (−0.27), suggesting readers can often infer correct names from context. However, this low impact on comprehensibility should not be taken to mean that named entity errors are inconsequential: for downstream NLP tasks such as named entity recognition, information extraction, or question answering, entity errors may be among the most damaging, effectively inverting the impact ranking observed here.

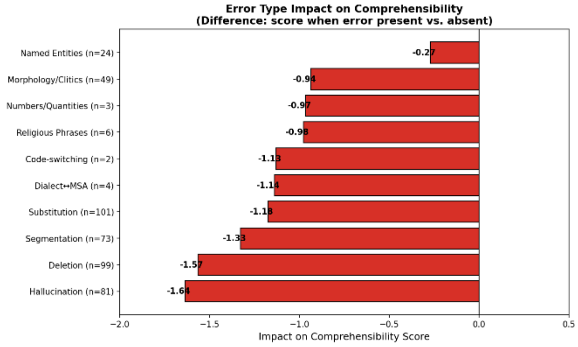


Figure 3: Impact of each error type on comprehensibility.

5.4 Error Count and Co-occurrence

We found a strong negative correlation between error count and comprehensibility (Pearson r = −0.776, $p < 0.001$; Figure 4). Transcriptions with zero errors averaged 4.86/5, declining to 4.20/5 for one error, 3.26/5 for two errors, 2.53/5 for three errors, and 1.84/5 for four errors. The most common error pairs were deletion+hallucination (58 co-occurrences), deletion+substitution (46), and hallucination+segmentation (45).

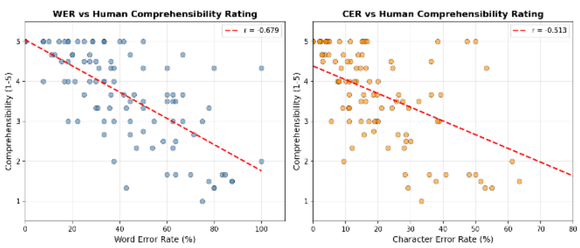


Figure 4: Mean comprehensibility by number of errors per transcription.

5.5 Dialect Effects

Performance varied by dialect. Based on human ratings, Najdi dialect achieved the highest comprehensibility (3.83), followed by Gulf Arabic (3.44) and Egyptian (3.40). MSA samples scored perfectly (5.00) but the sample size was too small (n=4) for reliable conclusions.

5.6 Inter-Annotator Agreement

For the 62 files evaluated by all three annotators, we computed pairwise Pearson correlations. For comprehensibility, correlations ranged from 0.61 to 0.67 (moderate-to-good agreement). Within ±1 point agreement was achieved on 79.6% of pairwise comparisons, with exact agreement on 40.9%. Naturalness showed more variation (r = 0.26 to 0.59), consistent with its more subjective nature. These agreement levels support the reliability of the annotation task while acknowledging inherent subjectivity in quality judgments.

5.7 Automatic Metric Analysis (WER/CER)

To complement our human evaluation, we computed Word Error Rate (WER) and Character Error Rate (CER) by aligning OmniASR transcriptions with SADA ground truth for 100 matched audio files. OmniASR achieved a mean WER of 42.2% (σ=24.0%, median=37.5%) and mean CER of 21.1% (σ=17.3%, median=16.1%). The WER distribution reveals substantial variation: 19% of files achieved excellent performance (WER <20%), while 37% showed poor performance (WER >50%).

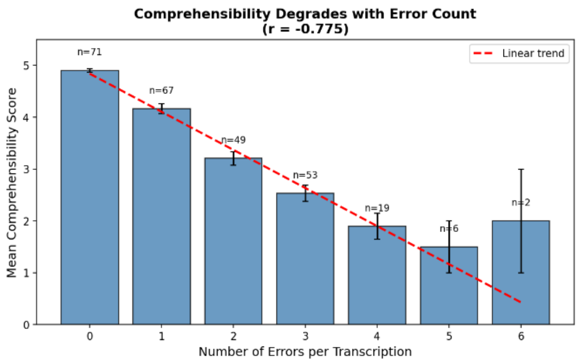


Figure 5: Correlation between automatic metrics (WER/CER) and human comprehensibility ratings.

We examined the correlation between automatic metrics and human judgments (Figure 5). WER showed a strong negative correlation with comprehensibility (r = −0.679), while CER showed a moderate correlation (r = −0.513). Notably, the WER–comprehensibility correlation (|r| = 0.679) is comparable to inter-annotator agreement for comprehensibility (pairwise r = 0.61–0.67), indicating that WER predicts human judgments approximately as well as individual annotators predict each other. This suggests that WER is a reasonable single-number proxy for overall perceived quality. However, WER by design treats all word-level errors as equivalent and cannot reveal which error types are most damaging. The value of our fine-grained taxonomy lies precisely in this diagnostic capability: identifying that hallucinations and deletions are 6× more impactful than named entity errors, information that no aggregate metric can provide regardless of its correlation with human judgments.

6. Discussion

6.1 The Polarized Performance Problem

Perhaps our most noticeable finding is OmniASR's polarized performance distribution: approximately one-third of transcriptions are essentially perfect, while one-fifth are essentially unusable. While the majority of scores (54%) fall in the middle range (2–4), the elevated frequency at both extremes is notable. This pattern would be obscured by aggregate metrics like mean WER (42.2%) or CER (21.1%). For practical applications, users need to know not just average quality but the probability of catastrophic failure.

This polarization likely reflects how neural ASR models process speech. When the acoustic signal clearly matches patterns in training data, the model produces confident, accurate transcriptions. When there is a mismatch due to unfamiliar dialect features, acoustic conditions, or speaker characteristics, the model may 'fall off a cliff' rather than degrading gracefully. The strong correlation between error count and comprehensibility (r = −0.776) supports this interpretation: errors cluster together in problematic utterances rather than being distributed uniformly.

6.2 Why Hallucinations and Deletions Are Most Damaging

Our analysis revealed that hallucinations (−1.64 impact) and deletions (−1.57) are by far the most damaging error types for comprehensibility, roughly 6× more impactful than named entity errors (−0.27). Hallucinations are particularly problematic because they introduce false information that readers have no way to detect or correct. In high-stakes applications (medical, legal, journalistic), hallucinated content could have serious consequences. The prevalence of hallucinations (18.3% of errors) suggests that OmniASR's LLM-based decoder may be over-generating content to create fluent output.

Deletions are damaging because they remove information entirely, potentially changing meaning. The Arabic sentence 'لا أوافق' (I do not agree) becomes 'أوافق' (I agree) if the negation particle is deleted. This type of semantically critical deletion, where function words such as negation markers are omitted, was observed in our data and illustrates how deletions can invert meaning entirely. The high co-occurrence of deletions with hallucinations (58 pairs) suggests a common failure mode: the model skips content it cannot confidently transcribe and compensates by generating plausible filler.

6.3 The Fluency Trap

OmniASR achieved notably higher naturalness scores (mean 4.05) than comprehensibility scores (mean 3.62). This gap reveals that the LLM-based decoder produces fluent, natural-sounding Arabic text that may nonetheless be semantically incorrect. This 'fluency trap' has been documented in neural machine translation (Martindale & Carpuat, 2018). For ASR evaluation, this finding argues strongly for separating fluency/naturalness from accuracy/comprehensibility in evaluation frameworks. An important question for future work is whether this fluency trap is specific to LLM-based decoding. CTC-based models, which decode frame-by-frame without autoregressive language model priors, would likely produce less fluent but potentially more faithful transcriptions. Comparing LLM-based and CTC-based architectures using our evaluation framework could reveal whether there is a systematic trade-off between naturalness and comprehensibility across decoding strategies.

6.4 WER and the Role of Fine-Grained Analysis

Our WER/CER analysis reveals a nuanced picture of WER's role in ASR evaluation. The WER–comprehensibility correlation ($r = -0.679$) is comparable to inter-annotator agreement on comprehensibility (pairwise $r = 0.61$–0.67), indicating that WER predicts human judgments at a level on par with the agreement among individual annotators. WER therefore serves as a reliable single-number summary of overall quality. However, the value of our error taxonomy is not in replacing WER but in complementing it with diagnostic information that WER by design cannot provide. WER treats all word-level errors as equivalent, yet our analysis shows that hallucinations and deletions are roughly 6× more damaging to comprehensibility than entity errors. A weighted WER scheme informed by these impact differences could provide a more informative automatic metric, and our taxonomy offers the empirical basis for developing such weightings.

6.5 Implications for Arabic ASR Evaluation

Our findings have several implications for Arabic ASR evaluation: (1) Weighted Error Metrics: evaluation metrics should weight errors by their perceptual impact; (2) Distributional Reporting: beyond mean WER, report the proportion of catastrophic failures and successes; (3) Dialect-Specific Evaluation: performance should be reported separately for MSA and each major dialect group; (4) Arabic-Specific Error Categories: track morphological errors, religious phrase accuracy, and dialect-MSA substitution; (5) Task-Dependent Impact Profiles: the error impact ranking we report reflects human comprehensibility, but downstream applications may exhibit entirely different sensitivity profiles. For instance, named entity errors had minimal impact on comprehensibility (−0.27) yet could be highly damaging for entity-centric tasks such as information extraction and question answering. Future work should evaluate the same transcriptions through downstream NLP pipelines to produce task-specific impact rankings that complement our human-centered analysis.

7. Conclusion

We presented the first fine-grained human evaluation of Meta's OmniASR system on Saudi Arabic dialects, introducing a 10-category error taxonomy designed for Arabic's linguistic and cultural characteristics. OmniASR achieved 42.2% WER and 3.62/5 mean comprehensibility, but exhibited a polarized performance pattern with 32.6% perfect transcriptions and 21.2% unusable ones. Error analysis reveals that hallucinations and deletions are approximately 6× more damaging than named entity errors for human comprehensibility, though the impact ranking may differ for downstream NLP tasks. While WER correlates with human judgments at a level comparable to inter-annotator agreement ($r = -0.679$ vs. pairwise annotator $r = 0.61$–0.67), it cannot identify which error types drive quality degradation. Our error taxonomy complements WER by providing this diagnostic granularity, supporting the development of error-type-aware evaluation frameworks for Arabic ASR.

8. Limitations and Ethical Considerations

Our study has several limitations. First, our sample size (103 audio files, 264 evaluations) is modest; a larger-scale study would provide more robust estimates, particularly for rare error types and dialect-specific breakdowns. Second, our annotators were senior undergraduate students who had completed an NLP course and received task-specific training rather than professional linguists; inter-annotator agreement for comprehensibility ($r = 0.61$–0.67) falls within ranges reported in similar ASR evaluation studies, but expert annotators might identify errors more consistently. Third, our error impact estimates are based on mean differences between present/absent conditions; a mixed-effects regression with random intercepts for annotator and audio file would provide more rigorous statistical inference, and we note this as a direction for follow-up analysis. Fourth, we evaluated only OmniASR, an LLM-based system; comparative evaluation against CTC-based models (e.g., Wav2Vec 2.0 CTC) and attention-based models (e.g., Whisper) would reveal whether the fluency trap and error type distributions we observe are architecture-specific. Finally, our sample was primarily Saudi dialects with some Egyptian content; evaluation on Levantine and Maghrebi dialects would further test the generalizability of our error taxonomy.

The annotation task was conducted as part of a graded course assignment in an NLP course. Grades were based on task completion and annotation quality (consistency and thoroughness) rather than on achieving particular results. The SADA dataset used in this study is publicly available under CC BY-NC-SA 4.0 license.

9. References

Ali, A., Bell, P., Glass, J., Messaoui, Y., Mubarak, H., Renals, S., & Zhang, Y. (2016). The MGB-2 Challenge: Arabic Multi-Dialect Broadcast Media Recognition. In Proceedings of the 2016 IEEE Spoken Language Technology Workshop (SLT), pp. 279-284. IEEE. https://doi.org/10.1109/SLT.2016.7846277

Alharbi, S., Alowisheq, A., Tüske, Z., Darwish, K., Alrajeh, A., Alrowithi, A., Bin Tamran, A., Ibrahim, A., Aloraini, R., Alnajim, R., Alkahtani, R., Almuasaad, R., Alrasheed, S., Alsubaie, S., & Alonaizan, Y. (2024). SADA: Saudi Audio Dataset for Arabic. In Proceedings of ICASSP 2024 - IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 10286-10290. IEEE. https://doi.org/10.1109/ICASSP48485.2024.10446243

Badawi, E. S. (1973). Mustawayat al-'arabiyya al-mu'asira fi Misr [Levels of Contemporary Arabic in Egypt]. Cairo: Dar al-Ma'arif.

Belinkov, Y., Barrón-Cedeño, A., Magidow, A., Shmidman, A., & Romanov, M. (2019). Studying the history of the Arabic language: Language technology and a large-scale historical corpus. Language Resources and Evaluation, 53(4), 771-805. https://doi.org/10.1007/s10579-019-09460-w

Boudelaa, S., & Marslen-Wilson, W. D. (2013). Morphological structure in the Arabic mental lexicon: Parallels between standard and dialectal Arabic. Language and Cognitive Processes, 28(10), 1453-1473. https://doi.org/10.1080/01690965.2012.719629

Ferguson, C. A. (1959). Diglossia. Word, 15(2), 325-340. https://doi.org/10.1080/00437956.1959.11659702

Holes, C. (2004). Modern Arabic: Structures, Functions, and Varieties (Revised ed.). Georgetown University Press.

Keren, G., Kozhevnikov, A., Meng, Y., Ropers, C., Setzler, M., Wang, S., Adebara, I., Auli, M., Balioglu, C., Chan, K., Cheng, C., Chuang, J., Droof, C., Duppenthaler, M., Duquenne, P.-A., Erben, A., Gao, C., Gonzalez, G. M., Lyu, K., ... Yates, S. (2025). Omnilingual ASR: Open-Source Multilingual Speech Recognition for 1600+ Languages. arXiv preprint arXiv:2511.09690. https://arxiv.org/abs/2511.09690

Koehn, P., & Monz, C. (2006). Manual and Automatic Evaluation of Machine Translation between European Languages. In Proceedings of the Workshop on Statistical Machine Translation, pp. 102-121. Association for Computational Linguistics.

Martindale, M. J., & Carpuat, M. (2018). Fluency Over Adequacy: A Pilot Study in Measuring User Trust in Imperfect MT. In Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (AMTA), Vol. 1, pp. 13-25.

McCowan, I., Moore, D., Dines, J., Gatica-Perez, D., Flynn, M., Wellner, P., & Bourlard, H. (2004). On the Use of Information Retrieval Measures for Speech Recognition Evaluation. IDIAP Research Report 04-73. IDIAP.

Saiegh-Haddad, E. (2018). MAWRID: A Model of Arabic Word Reading in Development. Journal of Learning Disabilities, 51(5), 454-462. https://doi.org/10.1177/0022219417720460

Tatman, R. (2017). Gender and Dialect Bias in YouTube's Automatic Captions. In Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, pp. 53-59. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1606

Wang, Y., Alhmoud, A., & Alqurishi, M. (2024). Open Universal Arabic ASR Leaderboard. https://arxiv.org/abs/2412.13788

SpeechLM for Automatic Speech Recognition in Low-resource Languages

Md Abdur Razzaq Riyadh[1], Eneko Agirre[1], Eva Navas[1], Claudia Borg[2]
[1]HiTZ Center, University of the Basque Country, Spain
[2]Dept. of Artificial Intelligence, University of Malta, Malta
{mdabdurrazzaq.riyadh,e.agirre,eva.navas}@ehu.eus, claudia.borg@um.edu.mt

Abstract

Multi-modal Speech Language Models (SpeechLMs) are a recent advancement in natural language processing. These SpeechLMs are instruction-tuned and optimized for general tasks. Their usefulness for Automatic Speech Recognition (ASR), particularly in relatively low-resource scenarios, remains largely understudied. This work developed SpeechLM for ASR in Basque and Maltese and studied the impact of language-adapted Large Language Model (LLM) and speech encoder within the SpeechLM for ASR. Using supervised learning, we fine-tuned LLaMA-Omni, a SpeechLM, for ASR. We have conducted comprehensive hyperparameter tuning and experimented with language-adapted SpeechLM components to improve performance and evaluated our best models on in-distribution datasets for both languages and an out-of-distribution dataset for Basque. LLaMA-Omni achieved 8.09% WER in Basque and 25.65% WER for Maltese on average across multiple test splits. The in-distribution results show that SpeechLM outperforms a fine-tuned ASR system under specific constraints, whereas it underperforms the baseline model on out-of-distribution Basque, indicating weaker overall robustness. We also find that a language-adapted LLM within SpeechLM improves in out-of-distribution settings when compared to the off-the-shelf LLM within SpeechLM.

Keywords: SpeechLM, ASR, LLM, Low-resource, Basque, Maltese

1. Introduction

ASR is the computational process of converting spoken language into written text, enabling machines to interpret and respond to human speech. It is an important task in natural language processing as speech is a natural and prevalent way of communication, while text is the more common modality of information processing. To bridge these modalities, SpeechLMs as a research field focus on seamless interaction between speech and language models, where the language model has an intrinsic capability to understand and generate speech. To directly understand speech, SpeechLMs encode raw audio signals or waveforms and convert them into discrete tokens or continuous representations. In this study, we adapt a SpeechLM to leverage its speech recognition capabilities, focusing primarily on the development and evaluation of ASR systems with a particular emphasis on the Basque and Maltese languages. The goal of this work is to analyze whether a general-purpose SpeechLM can be effectively adapted for supervised ASR in under-resourced languages, and to understand which components have a greater impact in different data regimes. Specifically, we have worked with LLaMA-Omni (Fang et al., 2025), a multi-modal speech language model that supports speech and text modalities as input and can produce output in both speech and text. We fine-tune LLaMA-Omni with supervised learning for ASR in different experimental setups to understand the impact of language-adapted components and evaluate them in terms of accuracy and robustness, where out-of-distribution data was available.

Our work demonstrates that LLaMA-Omni can be adapted for ASR, yielding better results than the fine-tuned speech encoder in some cases. On Basque in-distribution test sets, LLaMA-Omni outperforms the fine-tuned baseline model, `whisper-large-v3` (the speech encoder used within LLaMA-Omni) by 22%, demonstrating the effectiveness of SpeechLM architecture. However, the fine-tuned baseline outperforms LLaMA-Omni significantly on out-of-distribution evaluation. In Maltese, the performance decreases by 50% on the in-distribution test set due to a significantly smaller training dataset, highlighting a limitation in the SpeechLM architecture when handling data scarcity.

2. Related Work

Recent progress in SpeechLM has enabled end-to-end speech understanding and generation without relying on traditional cascaded ASR and text-to-speech pipeline. Among them, LLaMA-Omni (Fang et al., 2025) proposed a novel architecture toward low-latency, high-quality speech understanding and generation by directly mapping speech inputs to both textual and spoken outputs. However, their evaluation focuses exclusively on speech-to-text instruction-following and

Proceedings of SPEAKABLE @ LREC 2026, pages 125–131

speech-to-speech instruction-following tasks in English. This trend is common across SpeechLM research: evaluating on instruction-following generative tasks and downstream tasks such as ASR, text-to-speech, speech translation, etc., in high-resource languages. Although Soundwave (Zhang et al., 2025b) focuses on efficient training with limited resources and employs the same whisper-LLaMA 3.1-8B-Instruct architecture as LLaMA-Omni, its evaluation remains restricted to English speech tasks. EchoX (Zhang et al., 2025a) builds on Soundwave to emphasize acoustic robustness for knowledge-based question answering, again focusing on English. Despite its multilingual capabilities, GLM-4 Voice (Zeng et al., 2024) is restricted to high-resource training and evaluation regimes.

Although interest in SpeechLMs as general-purpose generative models is increasing, their performance on low-resource downstream tasks such as ASR remains largely unexplored. In this paper, we investigate whether recent SpeechLM architectures can be efficiently adapted to ASR for two low-resource languages: Basque and Maltese. We also isolate the effect of language adaptation at different components of the system. For the LLM, we incorporate Latxa Instruct (Etxaniz et al.), an instruction-tuned Basque LLM derived from the LLaMA 3.1-8B-Instruct (Grattafiori et al., 2024). Similarly, we analyze the impact of adapting the speech encoder, whisper. Our results provide the first systematic evaluation of SpeechLM-based ASR in low-resource language settings.

3. Methodology

We aim to investigate the adaptability of SpeechLM for low-resource ASR and further examine the impact of component-specific language-adaptation on ASR performance. Specifically, we address the following research questions: i) Can a general-purpose SpeechLM be adapted for supervised ASR in Basque and Maltese? ii) Does language adaptation of the LLM improve ASR performance in Basque? iii) Does adapting the speech encoder provide additional performance gains? While the first and third research questions focus on both Basque and Maltese, the second one only focuses on Basque, as a language-adapted LLM was not available for Maltese.

In this empirical study, we designed three controlled experiments using LLaMA-Omni to answer each of our research questions. The first experiment aimed to adapt LLaMA-Omni for ASR in Basque and Maltese, separately. Toward that goal, we fine-tuned the already trained English LLaMA-Omni[1] with supervised learning on respective datasets for each language. The goal of this experiment was to examine the effectiveness of adapting a SpeechLM for speech recognition in low-resource languages. LLaMA-Omni has four primary components: speech encoder, speech adaptor, LLM, and speech decoder, as presented in fig. 1. The second and third experiments allowed us to understand the impact of language-adapted LLM and speech encoder within the SpeechLM architecture on its performance in ASR. In the second experiment, we kept the same setting as in the first experiment, but replaced the LLaMA 3.1-8B-Instruct with Latxa Instruct (Etxaniz et al.), a continually pre-trained Basque-adapted LLaMA 3.1-8B-Instruct. We refer to this newly constructed SpeechLM architecture Latxa-Omni. We hypothesize that because Latxa Instruct has a better understanding of Basque, it should perform better at handling Basque-specific transcriptions, thereby improving ASR performance. The final experiment replicated the setup of the first experiment, with the exception that we swap the speech encoder with a version fine-tuned on the same training dataset for Basque and Maltese while keeping the remaining components fixed.

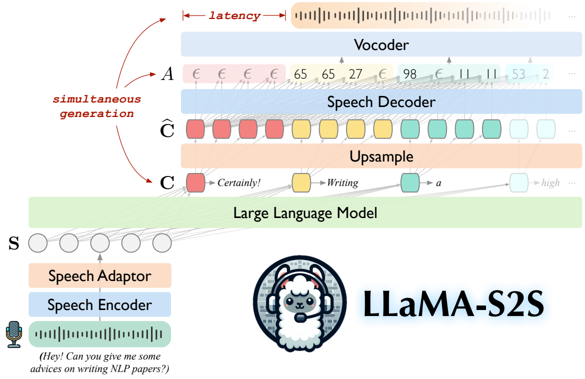


Figure 1: LLaMA-Omni's modular architecture, from Fang et al. (2025)

To study how training data size affects SpeechLM-based ASR, we conducted a control experiment by fine-tuning LLaMA-Omni on a 35-hour subset of Basque speech, matching the amount of training data available for Maltese. By keeping all other aspects identical to the first experiment, we could isolate the impact of extreme data constraints while controlling for language-specific factors.

The SpeechLM LLaMA-Omni combines speech and text into a unified input for the LLM. While the input audio for ASR is provided as speech to the LLM, the text prompt must be predetermined.

[1] https://huggingface.co/ICTNLP/Llama-3.1-8B-Omni/

Our preliminary experiments showed that the specific prompt used was not very important, as long as it was used consistently. For all our experiments, we fixed the prompt to be `<speech> Please directly repeat the sentence in <language>`, where `<speech>` gets replaced by the sequence of speech representation vectors from the speech adaptor and `<language>` is either ”Basque” or ”Maltese”.

Following Fang et al. (2025), only the speech adaptor and the LLM are fine-tuned in all experiments, whereas the speech encoder is kept frozen, including in the third experiment in which it was replaced. The speech decoder was frozen throughout, as it is only relevant for speech generation and is not used in our text-only ASR setup. We conducted a hyperparameter search, focusing on learning rate and batch size. We explored learning rates of $(1 \times 10^{-4},\ 5 \times 10^{-5},\ 2 \times 10^{-5},\ 1 \times 10^{-5},\ 5 \times 10^{-6})$ and batch sizes of $(32, 64, 128, 256)$. All experiments on Basque run for 6 epochs and on Maltese for 4 epochs, a value determined through preliminary trials. Each training run took approximately 30 hours. We used the cross-entropy loss (Goodfellow et al., 2016) optimized with AdamW (Loshchilov and Hutter, 2019) optimizer in combination with a cosine learning rate scheduler and a 5% warm-up phase. During inference, nucleus sampling was used with the temperature set to 0.6 and a probability of p = 0.9.

We used `whisper-large-v3` (Radford et al., 2022) as the baseline for both languages. As `whisper-large-v3` is the underlying speech encoder in LLaMA-Omni, this setup isolates the impact of the SpeechLM architecture on ASR performance. For Basque, the baseline was fine-tuned by Aholab[2], and for Maltese, we fine-tune it ourselves.

4. Data

In this work, we have used multiple datasets, and their summaries are presented in tables 1 and 2 for Basque and Maltese, respectively. For the training, validation, and in-distribution evaluation of Basque speech recognition systems, an aggregated dataset titled Composite Corpus EU v2.1[3] was utilized. This publicly accessible resource compiles subsets from several other publicly available datasets, specifically Common Voice 18 (Ardila et al., 2020), Basque Parliament Speech Corpus 1.0 (Varona et al., 2024), and OpenSLR 76 (Kjartansson et al., 2020). To evaluate the robustness of our system in Basque, we utilized Faktoria dataset, from EJIE[4] for out-of-distribution evaluation. Faktoria is a radio talk show and it differs substantially from our training data both in domain and acoustic characteristics. The recordings include spontaneous conversational speech, background music, and noticeable dialectal variation, making the dataset particularly challenging. The dataset is divided into two subsets: ‘Faktoria (EUS)’, containing standard Basque, and ‘Faktoria (Other)’, consisting of samples from various Basque dialects. Collectively, the Basque corpus comprises a total of 676 hours of training data and 63 hours of testing data.

The primary training dataset for Maltese was a speech-text parallel corpus, titled ‘MASRI’ (Hernandez Mena et al., 2020). It is a publicly accessible resource[5] developed to advance research in Maltese ASR. Along with MASRI, we also used Maltese data from Common Voice 22. There was no out-of-distribution dataset readily available for Maltese. In total, the Maltese datasets provide

Dataset	Set	Hour	# Samples
CV 18	Train	300.0	198498
	Test	24.0	14312
	Validation	1.0	620
Basque Parl.	Train	370.0	185699
	Test	3.0	1521
	Validation	1.0	550
OpenSLR	Train	6.0	3229
	Test	1.0	526
	Validation	1.0	521
Faktoria	Test (EUS)	27.5	19142
	Test (other)	7.5	4863

Table 1: Summary of the Basque datasets used in this study, including Common Voice 18, Basque Parliament Speech Corpus 1.0, OpenSLR, and Faktoria test sets.

Dataset	Set	Hour	# Samples
Headset v2	Train	6.50	4979
Farfield	Train	9.50	4476
Booths	Train	2.50	1256
MEP	Train	1.25	656
TUBE	Train	13.25	8956
	Test	1.00	663
	Validation	1.00	638
CV 22	Train	2.37	1910
	Test	2.36	1661
	Validation	2.09	1625

Table 2: Maltese datasets used in this study. Datasets Headset v2, Farfield, Booths, MEP, and TUBE subsets are from the MASRI corpus.

[2]https://huggingface.co/HiTZ/whisper-large-v3-eu
[3]https://huggingface.co/datasets/asierhv/composite_corpus_eu_v2.1
[4]https://www.ejie.euskadi.eus
[5]https://www.um.edu.mt/projects/masri

35.37 hours of training material and 3.36 hours for testing.

5. Results

The goal of this work is to investigate the adaptability of SpeechLM for ASR in Basque and Maltese. To that end, we use Word Error Rate (WER) for evaluation, where lower values indicate better ASR performance. The baselines for both languages were fine-tuned on the same training set for respective languages, referred in section 4. We also evaluate them on the same test sets from tables 1 and 2, reported in tables 3 and 4, and their performance fluctuates by language, yielding an WER of 10.43% for Basque and 16.74% for Maltese. For all model comparisons, we report the average over multiple splits.

5.1. Quantitative Evaluation

Experiment 1: Off-the-shelf Adaptation. The result from the first experiment is presented as 'LLaMA-Omni' in tables 3 and 4. We examined the efficacy of adapting a standard SpeechLM through supervised fine-tuning. In Basque, LLaMA-Omni outperformed the baseline by approximately 22%. Conversely, the Maltese model exhibited a performance degradation of roughly 53%, suggesting that the architectural adaptation is highly sensitive to the target language or available data volume.

Experiment 2: Language-adapted LLMs. The second question explores whether replacing the LLM in the SpeechLM with a Basque-specific, instruction-tuned LLM leads to improved ASR performance. As shown in table 3, Latxa-Omni achieved the best overall performance with a WER of 7.98%, maintaining a significant lead over the baseline, though showing only marginal gains over the standard LLaMA-Omni in in-distribution settings.

Experiment 3: Adapted Speech Encoders. The third experiment explores the impact of a language-adapted speech encoder within the SpeechLM architecture. This experiment's result is presented as 'LLaMA-Omni (+ adapted SE)' in tables 3 and 4 which replaces the speech encoder adapted for each language. Integrating a language-adapted speech encoder yielded mixed results. For Basque, the model performed 21% better than the baseline, mirroring previous experiments. And in Maltese, this variant remained approximately 50% worse than the baseline. This suggests that while speech encoder adaptation can be beneficial, it cannot fully compensate for other bottlenecks in low-resource settings.

Data Scarcity Analysis. To investigate the Maltese performance gap, we conducted a controlled experiment in the Basque using a reduced corpus (LLaMA-Omni (35H) in table 3) which was sampled randomly. This model exhibited a marked performance degradation, with WER increasing by 76% relative to the baseline and 127% compared to LLaMA-Omni, which was fine-tuned on the entire corpus. These results support the hypothesis that data scarcity is the primary factor limiting the adaptation of SpeechLM architectures for Maltese.

5.2. Out-of-distribution Evaluation

To evaluate the robustness of our systems, we evaluate our models on an out-of-distribution dataset. We could only do this for Basque, utilizing the Faktoria corpus described in section 4. No out-of-distribution datasets were available for Maltese. Initial observations on the results presented in table 5 reveal a notable degradation in all models' performance on Faktoria. Interestingly, the baseline outperforms all the LLaMA-Omni variants from experiments 1-3 on both subsets, which contrasts our results in the in-distribution setting, reported in table 3. Moreover, the substantially higher WER observed on the Faktoria (other) subset by all models, including the baseline, compared to Faktoria (EUS only), clearly demonstrates the models' lack of robustness to dialectal variation. However, models with language-adapted components outperform LLaMA-Omni, while Latxa-Omni has the highest margin of 13% compared to the LLaMA-Omni from the first experiment. These results suggest that LLaMA-Omni negatively impacts generalizability ASR when compared to `whisper-large-v3`. Nevertheless, using language-adapted components generalize better than LLaMA-Omni, demonstrating better robustness. This pattern did not emerge in the in-distribution evaluation for Basque and Maltese.

5.3. Qualitative Analysis

Basque. To better understand the quality of the transcriptions produced across experiments in Basque, we analyzed both the best-performing and worst-performing transcription cases from the test sets, along with the most commonly mistaken words. In cases where a model achieved perfect transcriptions (WER = 0%), we observed that the corresponding audio featured clear pronunciation and minimal background noise. These recordings were typically well-articulated, contributing to the model's ability to recognize and transcribe the speech accurately. On the other hand, in poorly transcribed samples, particularly those with high WER, many samples contained speech that was phonetically ambiguous or acoustically similar to

Model/Corpus	CV 18	Basque Parl.	OpenSLR	Average
whisper-large-v3	5.08%	13.72%	**12.48%**	10.43%
LLaMA-Omni (35H)	16.62%	11.04%	27.45%	18.37%
LLaMA-Omni	4.80%	5.00%	14.47%	8.09%
Latxa-Omni	4.79%	**4.91%**	14.23%	**7.98%**
LLaMA-Omni (+ adapted SE)	4.86%	4.91%	14.78%	8.18%

Table 3: Basque ASR evaluation results for the baseline and LLaMA-Omni models. The first row shows the baseline system; the second shows LLaMA-Omni trained on only 35 hours of Basque data. The remaining rows correspond to Experiments 1-3: LLaMA-Omni fine-tuned on the full corpus, LLaMA-Omni with the LLM substituted, and LLaMA-Omni with an adapted speech encoder. Columns report WER on CV18, Basque Parliament Speech Corpus 1.0, and OpenSLR, along with their average.

Model/Corpus	CV 22	MASRI	Average
whisper-large-v3	**11.44%**	**22.04%**	**16.74%**
LLaMA-Omni	21.08%	30.22%	25.65%
LLaMA-Omni (+ adapted SE)	20.93%	29.33%	25.13%

Table 4: Maltese ASR results for the baseline and LLaMA-Omni experiments. The first row shows the baseline system. The next two rows correspond to Experiment 1 (LLaMA-Omni fine-tuning) and Experiment 3 (LLaMA-Omni with an adapted speech encoder). Columns report WER on CV22, MASRI, and their average.

other words, making it difficult for the model to differentiate. This suggests that the language model has difficulty handling words that sound alike, especially when there is little context to help disambiguate them. The transcription quality of Latxa-Omni and LLaMA-Omni (+ adapted SE) model is highly comparable to our fine-tuned 'LLaMA-Omni (2e-5, 256)'. Both models tend to struggle with a similar set of samples, primarily those containing phonetically close or noisy audio. In several samples, Latxa-Omni even performed worse than LLaMA-Omni. Interestingly, many of the word substitutions made by Latxa-Omni overlap with those observed in LLaMA-Omni's outputs, particularly for phonetically similar words.

Maltese. To understand the transcription quality in Maltese, we compare LLaMA-Omni and LLaMA-Omni (+ adapted SE) and find that the transcription quality is quite similar. Both model struggle with code-switching and transcribe phonetically instead of using the original spelling. The original transcriptions in the MASRI dataset contain filler words, but our fine-tuned models generally ignore them. The most common errors involve confusion between phonetically similar words. Another source of errors involves phonetic mis-segmentation, producing unintelligible tokens. These errors often take the shape of Italian-like morphological forms, reflecting the influence of contact language on the model's output. For example, "Ħalli niddistingwu" is transcribed as "Hallini d-distingu" by LLaMA-Omni. This example illustrates that while the speech encoder is able to phonetically capture Maltese sounds, the limited size of the dataset constrains LLM's ability to acquire and generalize Maltese grammatical rules.

6. Conclusions

This work investigated the adaptability of SpeechLM for ASR in the low-resource languages of Basque and Maltese, specifically examining how architectural components and data scale influence model performance. Our findings reveal that while LLaMA-Omni is highly effective and achieves comparable performance to traditional baselines when sufficient data is available, it is significantly more sensitive to data scarcity. The performance degradation observed in Maltese and the reduced-corpus Basque experiments suggests that the large-scale architecture of LLaMA-Omni requires a higher data size threshold for effective supervised fine-tuning than specialized ASR models.

The study further demonstrated a marked improvement in generalization to out-of-distribution samples and dialectal variations for Basque over off-the-shelf fine-tuning. This indicates that, in low-resource and linguistically diverse scenarios, the LLM plays a more critical role in handling distribution shifts than acoustic features alone. Despite these advancements, the reliance on substantial labeled datasets remains a limitation for truly low-resource scenarios. Future work should focus on optimizing adapter designs and exploring cross-lingual transfer mechanisms to reduce the data burden for SpeechLM adaptation. Finally, this research underscores that advancing ASR for under-resourced languages necessitates a shift toward

Model/Corpus	Faktoria (EUS only)	Faktoria (others)
whisper-large-v3	**18.56%**	**29.42%**
LLaMA-Omni	25.23%	54.81%
Latxa-Omni	21.87%	31.54%
LLaMA-Omni (+ adapted SE)	22.88%	45.14%

Table 5: Out-of-distribution (OOD) evaluation result for Basque. The first row is the baseline model's OOD evaluation. The remaining are models from the first, second, and third experiment respectively.

a more holistic integration of language-specific linguistic knowledge within the SpeechLM framework.

7. Acknowledgments

This work was partially supported by the Erasmus Mundus Master's Programme in Language and Communication Technologies (LCT), the Basque Government (IKER-GAITU project), the Spanish Government (Project AIA2025-163322-C61 funded by MICIU/AEI/10.13039/501100011033).

8. Bibliographical References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Julen Etxaniz, Oscar Sainz, Naiara Miguel, Itziar Aldabe, German Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, and Aitor Soroa. Latxa: An open language model and evaluation suite for basque. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14952–14972. Association for Computational Linguistics.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2025. Llama-omni: Seamless speech interaction with large language models. In *International Conference on Representation Learning*, volume 2025, pages 57607–57624.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. 2024. The Llama 3 Herd of Models.

Carlos Daniel Hernandez Mena, Albert Gatt, Andrea DeMarco, Claudia Borg, Lonneke van der Plas, Amanda Muscat, and Ian Padovani. 2020. MASRI-HEADSET: A Maltese corpus for speech recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6381–6388, Marseille, France. European Language Resources Association.

Oddur Kjartansson, Alexander Gutkin, Alena Butryna, Isin Demirsahin, and Clara Rivera. 2020. Open-Source High Quality Speech Datasets for Basque, Catalan and Galician. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 21–27, Marseille, France. European Language Resources association.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision.

Amparo Varona, Mikel Penagarikano, Germán Bordel, and Luis Javier Rodriguez-Fuentes. 2024. A Bilingual Basque–Spanish Dataset of Parliamentary Sessions for the Development and Evaluation of Speech Technology. *Applied Sciences*, 14(5):1951.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. 2024. GLM-4-Voice: Towards Intelligent and Human-Like End-to-End Spoken Chatbot.

Yuhao Zhang, Yuhao Du, Zhanchen Dai, Xiangnan Ma, Kaiqi Kou, Benyou Wang, and Haizhou

Li. 2025a. Echox: Towards mitigating acoustic-semantic gap via echo training for speech-to-speech llms.

Yuhao Zhang, Zhiheng Liu, Fan Bu, Ruiyu Zhang, Benyou Wang, and Haizhou Li. 2025b. Soundwave: Less is More for Speech-Text Alignment in LLMs.

Improving low-resource ASR using bilingual fine-tuning with language identification: a cross-linguistic evaluation

Reihaneh Amooie[1], **Yun Hao**[1], **Wietse de Vries**[1], **Jelske Dijkstra**[2], **Matt Coler**[1]
Martijn Wieling[1,3]
[1]University of Groningen, [2]Fryske Akademy, [3]Vrije Universiteit Brussel
{r.amooie, yun.hao, wietse.de.vries, m.coler, m.b.wieling}@rug.nl, jdijkstra@fryske-akademy.nl

Abstract

This study explores how bilingual fine-tuning affects automatic speech recognition (ASR) in low-resource languages. We evaluate this method across nine linguistically and geographically diverse language pairs, covering a range of language families and writing systems. To distinguish the two languages, during training, we pre-pend each input text with a language identification token. At inference, the model jointly predicts both the language and transcription from the speech input alone. As texts for which the language is incorrectly determined show low ASR performance, we also conduct a follow-up experiment in which the language identification token is provided both during training and inference. Our results show that bilingual fine-tuning can be beneficial when language identification accuracy is high, and that in cases where language identification performance is low, including the language identification token at inference helps to improve ASR performance.

Keywords: low-resource ASR, cross-lingual transfer, language identification, self-supervised learning

1. Introduction

Automatic Speech Recognition (ASR) has progressed rapidly in recent years. However, much of that progress remains limited to high-resource languages, such as English. Low-resource languages still face challenges, including data scarcity and dialectal variation. A promising approach to address this challenge is cross-lingual transfer learning, where knowledge from high-resource languages supports low-resource ones.

Both language-aware and language-agnostic multilingual fine-tuning strategies have been explored as promising approaches to improve ASR in low-resource settings by enabling cross-lingual transfer. Yang et al. (2023), for instance, proposed a sparse multilingual model with language-specific sub-networks and overlapping shared paths, allowing high-resource data to support lower-resource languages. Similarly, San et al. (2024) demonstrated that supplementing a low-resource language with donor data can be highly effective: continued pre-training with only 10 hours of Punjabi plus 60 hours of Hindi nearly matched the performance of using 70 hours of Punjabi alone.

Another line of work explicitly incorporates language information. External language identification (LID) can aid disambiguation but increases latency (Waters et al., 2019), leading to approaches that integrate implicit or explicit LID into ASR systems. Examples include meta-learning frameworks (Hsu et al., 2020; Xiao et al., 2021), adapter-based methods (Hou et al., 2020; Winata et al., 2020), and multi-task learning setups (Hou et al., 2020; Chen and Mak, 2015). In multi-task learning, Chen et al. (2023) used an auxiliary CTC objective so that earlier encoder layers focused on language identification while later layers generated transcriptions conditioned on language identity. Relatedly, Liu et al. (2023) employed hierarchical Softmax with a Huffman Tree structure, exploiting similarities in linguistic unit frequency distributions across related languages to boost low-resource performance. A recent study by Amooie et al. (2025) shows that Frisian ASR performance seems to improve when multilingual fine-tuning corpora are augmented with pre-pended language identification (LID) tokens, enabling the model to condition on language identity during fine-tuning and evaluation.

However, it remains unclear whether bilingual fine-tuning with explicit language identification consistently improves ASR across diverse linguistic contexts. Prior work has not systematically examined this approach across language families, scripts, or varying degrees of linguistic similarity. Factors such as linguistic proximity, writing script compatibility, and the model's ability to disentangle languages have not been systematically examined. In this work, we therefore investigate when and under what conditions bilingual fine-tuning using two related languages with explicit language identification (LID) tokens (prepended to each training sample to indicate its language) benefits low-resource languages. We evaluate the approach across nine typologically diverse pairs of related languages spanning five language families and multiple writing systems. For five target–donor pairs (for which the available data allows for sub-sampling), we train multiple models using random subsets of data to assess the robustness of any observed gains.

Proceedings of SPEAKABLE @ LREC 2026, pages 132–138
11 May 2026.

2. Data

In addition to Frisian (with the aim of reproducing the results of Amooie et al., 2025), we included eight additional target languages from five language families. We deliberately selected eight target languages representing diverse language families, typological features, and writing systems to evaluate the method across various low-resource scenarios.

For each target language, we first selected the most similar donor language available in Common Voice 17.0 Ardila, Rosana, and others (2019) based on the lexical-phonetic distances (i.e. the LDND distance) from the ASJP database (Wichmann et al., 2010, see also De Vries et al., 2021 for a similar approach). We adopt a bilingual fine-tuning setup rather than a multilingual one, as prior work Amooie et al., 2025 has shown that multilingual (involving more than two languages) fine-tuning does not clearly improve over bilingual fine-tuning.

All audio was extracted from Common Voice 17.0 and sampled at 16 kHz. To avoid bias toward higher-resource languages, we down-sampled both languages in each pair to 3,000 utterances. We control for the number of utterances (3,000 or fewer, depending on available resources) to ensure a comparable number of training instances across languages. While this does result in some variation of the total duration due to differences in utterance length distributions, we opted for consistency in the number of samples seen during training. When fewer utterances were available (i.e. thereby representing a very-low-resource scenario), we matched both languages to the lower count. For five language pairs with over 3,000 utterances available (FY-NL, DA-SV, GL-IT, UK-BE, SK-CS), we repeated training 10 times using different random seeds to assess the statistical significance and robustness of the improvements. Table 1 shows all relevant information for each of these datasets.

Each model was evaluated on the test split of Common Voice 17.0 (Ardila, Rosana, and others, 2019) for the target low-resource language in each pair. We did not down-sample the test sets. However, the development splits, which were used to monitor training progress and tune hyperparameters, were down-sampled to match the number of training samples.

2.1. Language pairs

In this section, we briefly introduce the selected language pairs with linguistic information from Glottolog (Hammarström et al., 2024) and WALS (Dryer and Haspelmath, 2024).

Table 1: Fine-tuning datasets (subsets from Common Voice 17.0; Ardila, Rosana, and others (2019)). Rows with a single language correspond to the target language, while rows with language pairs combine the target language (first) with a source language (second). Sim. script: both languages use similar scripts; Dur.: duration; # Utt.: number of utterances; # Spk.: number of speakers.

Languages	ISO	Sim. script	Dur.	# Utt.	# Spk.
Frisian	FY	-	4.25h	3000	192
Frisian, Dutch	FY-NL	yes	8.27h	6000	205
Danish	DA	-	3.44h	3000	4
Danish, Swedish	DA-SV	yes	6.58h	6000	20
Galician	GL	-	4.12h	3000	26
Galician, Italian	GL-IT	yes	8.21h	6000	30
Ukrainian	UK	-	4.04h	3000	29
Ukrainian, Belarusian	UK-BE	yes	8.15h	6000	38
Slovak	SK	-	3.21h	3000	9
Slovak, Czech	SK-CS	yes	7.19h	6000	20
Serbian	SR	-	1.47h	1880	6
Serbian, Bulgarian	SR-BG	yes	3.98h	3760	8
Slovenian	SL	-	1.34h	1390	9
Slovenian, Polish	SL-PL	yes	3.19h	2780	34
Malayalam	ML	-	1.42h	1260	2
Malayalam, Tamil	ML-TA	no	3.28h	2520	9
Finnish	FI	-	2.60h	2080	6
Finnish, Estonian	FI-ET	yes	6.56h	4160	240

- **Germanic**

  Frisian (target) and *Dutch* (donor) were included as West Germanic languages to reproduce the results of Amooie et al. (2025). Both use the Latin alphabet (24 and 26 characters, respectively, depending on how they are counted), with Frisian featuring unique diacritics such as â, ê, and ô. Dutch was chosen as donor due to its high similarity to Frisian.

  Danish (target) and *Swedish* (donor) represent North Germanic languages, both using the Latin alphabet (29 letters) but differing in their extra characters, e.g., ä and ö in Swedish versus æ and ø in Danish.

- **Romance**

  Galician (target) and *Italian* (donor) are Romance languages that both use the Latin script but follow different orthographic conventions. For instance, Galician includes the letter ñ and uses acute accents to mark non-default stress, whereas Italian lacks ñ but employs grave/acute accents (e.g., è, é, ò, ó), chiefly to indicate vowel quality and final stress.

- **Slavic**

  We included four pairs of Slavic languages that differ in script and spelling traditions. *Ukrainian–Belarusian* and *Serbian–Bulgarian* are written in Cyrillic, while *Slovak–Czech* and *Slovenian–Polish* use the Latin alphabet. Within each pair, the languages share the same script, but differ in orthographic conventions. For example, Ukrainian and Belarusian vary in certain characters and soft sign usage, while Serbian and Bulgarian differ in some Cyrillic letters and vowel notation.

- **Dravidian**

  Malayalam (target) and *Tamil* (donor) are Dravidian languages with writing systems that, while historically connected, developed distinct characteristics. Both scripts trace back to Brahmi, but Malayalam evolved through Grantha and retains more of its features, whereas Tamil underwent greater simplification and standardization.

- **Uralic**

  Finnish (target) and *Estonian* (donor) are Uralic languages that use Latin alphabets with distinct diacritics. For example, Estonian employs š, ž, and o, which Finnish lacks, while Finnish uses å (mainly in Swedish loanwords). Estonian also omits c, q, w, x, and y.

3. Method

3.1. Finetuning procedure

We fine-tuned the pre-trained XLS-R 1B model (Babu et al., 2021), which is based on the Wav2Vec 2.0 architecture (Baevski et al., 2020) and contains a convolutional feature encoder followed by a transformer-based context network. During fine-tuning, we froze the convolutional feature encoder to retain the pre-trained acoustic representations and updated only the transformer layers. This approach aligns with standard practices in fine-tuning Wav2Vec 2.0 models for ASR tasks. To train the models, we used a learning rate of 0.00008 and a batch size of 8 with 16 gradient accumulation steps. All experiments were conducted using 16-bit floating point precision on a single NVIDIA A100 GPU with 40 GB of RAM.

We compared each model trained using the bilingual data to a baseline model trained only using the target language data. When the size of the dataset doubles, the training time necessarily increases. To ensure that any performance gain is caused by adding donor language data rather than longer training, we held the number of training epochs constant (50) in all experiments. This was to ensure that every training example is seen 50 times in both monolingual and bilingual runs, so the bilingual model performs two times more updates only because it contains two times more training examples. This also prevents potentially overfitting the monolingual model.

3.2. Language identification procedure

To provide explicit language context during training and inference, we pre-pended a language identification (LID) token to each utterance (sentence) during training. This LID token serves as a ground-truth label indicating its language (e.g., [FY-NL] for Frisian, and [NL] for Dutch).

In this approach, at inference, we do not provide the language identification token. The decoder first predicts the LID token, and then the transcription. By doing so, language identifications and transcriptions are learned and inferred jointly.

3.2.1. Follow-up experiment: Providing the correct LID during inference

Proceeding from the assumption that samples for which the language was correctly predicted have better ASR performance, we also add a small experiment for a subset of language pairs in which we condition the model directly on the target language identity. To achieve this, we extend `Wav2Vec2ForCTC` with a simple language-specific bias embedding (one vector per language, sized to the vocabulary) and add this bias to the CTC logits at every time step. Each training and test utterance is assigned a numeric language ID (e.g., `0` for Danish, `1` for Swedish), which the model receives as an additional argument (`langid`) during fine-tuning and inference. During both phases, the model looks up the bias vector corresponding to the given language. This bias vector functions as a language-specific prior that shifts the model's output distribution toward that language's characteristic phonemes and orthographic patterns. As a result, the decoder is more likely to predict characters and sequences typical of the target language. The encoder remains completely shared across languages, so this mechanism introduces only a minimal number of additional parameters while explicitly informing the decoder about the correct language. This conditioning helps the decoder stay within the correct language space and prevents cross-language confusion during decoding.

3.3. Evaluation metrics

ASR performance is evaluated using the word error rate (WER). The benefit of bilingual fine-tuning is quantified as ΔWER, defined as monolingual WER minus bilingual WER. Consequently, more positive values indicate a greater improvement (i.e. reduced WER). LID performance was evaluated using accuracy, defined as the proportion of test utterances whose language was correctly identified as the target language.

4. Results and Discussion

Table 2 and Figure 1 present the overall comparison of bilingual fine-tuning with language identification (LID) against monolingual baselines across all language pairs. For a subset of five language pairs (FY-NL, DA-SV, GL-IT, UK-BE, and SK-CS), we conducted multi-run experiments with 10 randomly selected subsets of 3000 training samples per language pair. As the remaining four language pairs (SR-BG, SL-PL, ML-TA, and FI-ET) had fewer than 3000 training samples per language pair available (i.e. representing very-low-resource languages; see Table 1) and subsetting was thus not possible, we only report the results of the single run including all data for these pairs.

In Figure 1, the blue bars denote the (average) improvement in word error rate (ΔWER), while the orange markers indicate the corresponding LID accuracy. For the multi-run pairs, error bars show the variance across shuffles. For the remaining very-low-resource language pairs, results are based on a single run. The bars are sorted from left to right by decreasing ΔWER. Overall, we observe a diverse pattern: a small majority of language pairs, such as Frisian–Dutch, Galician–Italian, and Danish–Swedish, show positive ΔWER values, indicating consistent gains from bilingual fine-tuning, whereas a smaller number of pairs including Finnish–Estonian, Slovenian-Polish and Ukranian-Belarusian exhibit negative ΔWER, suggesting that bilingual fine-tuning may also harm performance. However, it seems that bilingual fine-tuning only appears to harm performance when LID accuracy is not very high (i.e. lower than 95%). Furthermore, script differences do not appear to hinder transfer, as the ML-TA pair shows a positive ΔWER despite differing scripts. Likewise, the number of training speakers also did not exhibit a strong effect, as Tamil has relatively few speakers in the training data, yet helped to improve ASR performance for Malayalam, whereas Estonian has many speakers in the training data, but did not improve ASR performance for Finnish.

For the five language pairs for which multiple runs were possible (due to having more data to randomly sample from), Table 2 also summarizes the average differences in WER between the bilingual and monolingual models across the 10 shuffles. Single-sample t-tests of the differences in WER (H_0: $\Delta\text{WER}=0$) show that ΔWER is significantly greater than zero for FY–NL, DA–SV, and GL–IT, but that there is no significant difference for the other two language pairs (UK-BE, and SK-CS). Consequently, for three out of five pairs, all

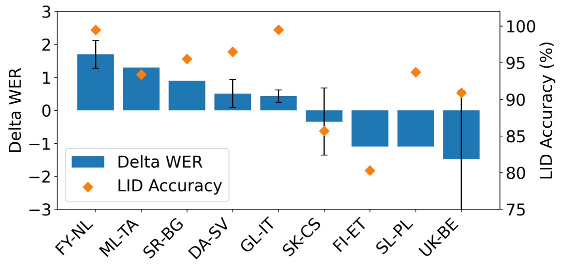


Figure 1: Average ΔWER with error bars and LID Accuracy by language pair

showing an average LID accuracy over 95%, the bilingual approach appears to be beneficial. This finding suggests that LID accuracy may be an important factor in determining ASR performance in our bilingual approach. In the following section, we will investigate this in more detail.

4.1. Relation between LID accuracy and ΔWER

As language pairs with higher LID accuracy often exhibit a larger ΔWER, we examine this relationship more systematically. Consequently, we determined the (Pearson) correlation between the LID accuracy and ΔWER across the 10 random shuffles for the five language pairs included in Table 2. Table 3 reports the numerical results of this analysis, whereas Figure 2 provides the visualization in a scatter plot. The (regression) lines visualize the strength and direction of the correlation, both overall and per language pair. All correlations are positive (also reflected by the angle of the lines in Figure 2). While the statistical significance varies by language pair, this is caused by the relatively low number of runs (i.e. 10) per language pair. The general pattern across all languages, however, is significant ($p < 0.001$).[1]

These findings suggest that poor LID performance is a bottleneck limiting the benefits of bilingual fine-tuning. To test whether providing correct LID at inference could overcome this bottleneck, we conducted a follow-up experiment on three language pairs with varying LID accuracy.

4.2. Providing LID at inference

As higher LID generally appeared to be associated with a greater improvement in WER for the bilingual model, we also investigated a potentially beneficial effect of providing LID at inference. For this, we conducted an additional experiment for three language

[1]Note that due to the dependencies in our data (i.e. multiple observations for five language pairs), the overall significance had to be determined via a linear mixed-effects regression analysis with language pair as a random-effect factor.

Table 2: WER comparison of the monolingual baseline vs. the bilingual model. A positive ΔWER indicates an improvement of the monolingual model compared to the bilingual model. For the five language pairs for which 10 random subsets of training data were used, the p-value reflects the significance of a (two-tailed) single-sample t-test of the difference in WER with nine degrees of freedom. Significant p-values (< 0.05) are marked in boldface. Dist. indicates the linguistic distance between the pairs, and LID acc. indicates the accuracy of the inferred language identification across 10 random shuffles.

Language pair	WER$_{\text{monolingual}}$ (SD)	WER$_{\text{bilingual}}$ (SD)	ΔWER (SD)	$p(t)$	Dist.	LID acc. (%)
FY–NL	16.1 ($\pm$0.4)	14.4 ($\pm$0.3)	+1.7 ($\pm$0.4)	**< 0.001**	52.0	99.5
DA–SV	21.3 ($\pm$0.3)	20.7 ($\pm$0.5)	+0.5 ($\pm$0.4)	**0.004**	52.4	96.5
GL–IT	10.8 ($\pm$0.1)	10.4 ($\pm$0.1)	+0.4 ($\pm$0.2)	**< 0.001**	49.9	99.5
UK–BE	27.9 ($\pm$0.9)	29.4 ($\pm$1.5)	-1.5 ($\pm$2.1)	0.051	48.1	90.9
SK–CS	25.2 ($\pm$0.4)	25.5 ($\pm$0.9)	-0.3 ($\pm$1.0)	0.319	32.8	85.7
SR–BG	15.3	14.4	+0.9	N/A	48.0	95.5
SL–PL	19.5	20.6	-1.1	N/A	46.4	93.7
FI–ET	25.4	26.5	-1.1	N/A	47.6	80.0
ML–TA	75.2	73.9	+1.3	N/A	34.8	92.4

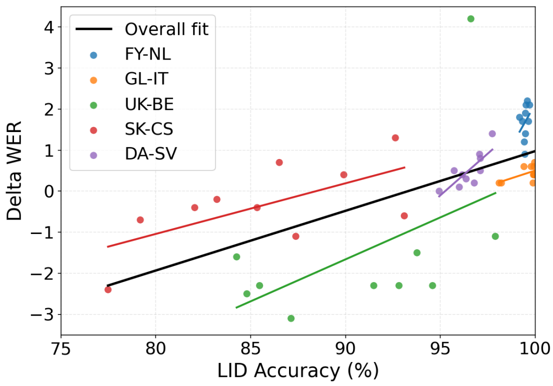


Figure 2: Relationship between LID accuracy (%) and the performance difference between bilingual finetuning compared to the monolingual baseline.

Table 3: Pearson correlation analysis between ΔWER and LID accuracy. Significant p-values (< 0.05) are marked in boldface.

Language Pair	Pearson r	$p(r)$	R^2
FY–NL	0.298	0.403	0.089
DA–SV	0.792	**0.006**	0.627
GL–IT	0.582	0.078	0.338
UK–BE	0.501	0.141	0.251
SK–CS	0.642	**0.045**	0.412
ALL	0.621	**< 0.001**[1]	0.386

pairs (FY-NL, DA-SV, and SK-CS) for which 3000 samples per language were available. For each language pair, we conducted a single experiment to limit the required computational time.

The potential benefit of providing LID during inference will both be dependent on the LID performance for a language pair, as well as the difference in WER for samples for which the LID was identified correctly versus those for which the LID was identified incorrectly. Specifically, for language pairs with a lower LID performance, and those with a greater difference in WER between correctly versus incorrectly identified samples, there will be a greater potential for improvement.

Table 4: WER for three language pairs under two proposed settings: Predicted LID at inference and Given LID at inference. For each setting we report the WER (%). For the predicted LID, we provide two additional values between parentheses: the WER for the samples where the LID was correctly predicted (corr.), and the WER for the samples where the LID was incorrectly predicted (incorr.).

Pair	WER$_{\text{Given LID}}$	WER$_{\text{Predicted LID}}$ (corr.; incorr.)
FY–NL	**14.3**	**14.3** (14.1; 53.3)
DA–SV	**20.4**	22.2 (20.7; 59.7)
SK–CS	**25.7**	26.3 (24.5; 35.4)

The three language pairs we selected varied along these dimensions. Specifically, LID performance for FY-NL was very high (over 99%), whereas it was lower for DA-SV (about 95%), and lowest for SK-CS (about 89%). The WER of samples for which the LID was incorrectly identified was always higher than for those whose LID was correctly identified, but this performance gap also differed per language pair. For FY-NL and DA-SV, the difference was large (about 40%; 14.1% for FY-NL samples with a correct LID vs. 53.3% for FY-NL samples with an incorrect LID, and 20.7% vs. 59.7%, respectively, for DA-SV), whereas it was much smaller for SK-CS (about 11%; 24.5% vs. 35.4%, respectively), likely due to the increased similarity between the two languages (see Table 2).

Taking these two aspects together, we expect a greater potential for improvement for both DA-SV and SK-CS compared to FY-NL. Indeed, our results shown in Table 4 show this assumption to be correct. Due to the very high LID, explicitly including LID at inference for FY-NL resulted in equal performance (14.3%) compared to not including it. By contrast, for both DA-SV (WER reduction of 1.8%; from 22.2% to 20.4%), as well as SK-CS (WER reduction of 0.6%; from 26.3% to 25.7 %), the results improved by explicitly including LID at inference.

4.3. Limitations

While we identified clear correlation between ASR and LID performance, we did not investigate why LID performance was lower for some language pairs than for others. While we expect this pattern to be similar across model architectures, we did not investigate the effect of model architecture, the fine-tuning approach we used, or the stopping criterion. Likewise, we did not investigate whether the bilingual approach, in cases where the LID was correctly identified, may have resulted in new mistakes compared to the monolingual model. We will address these questions in future work. It is clear, however, that a lower LID performance will result in performance degradation, as the WER will be much higher for samples where language was incorrectly identified (as is shown in Table 4). While adding a third or even a fourth language has shown to improve results slightly (Amooie et al., 2025), it also tends to result in lower LID performance. Manually providing LID for each sample solves this issue, but may not always be possible in automated pipelines.

5. Conclusion

In this study, we investigated bilingual fine-tuning with LID tokens for low-resource ASR across nine diverse language pairs. Our analysis shows that performance gains are not consistent across all settings but are strongly associated with LID accuracy. Reliable language disambiguation, particularly in related languages, leads to clear improvements, while lower LID accuracy limits or even counteracts the benefits, as the ASR performance is lower for samples with an incorrect LID. In that case, preliminary results showed that providing the correct language identification at inference helps to improve performance compared to the monolingual baseline. Consequently, in cases of little data for one language, ASR performance for that language may be increased by supplementing the monolingual data with data of a similar language, as long as LID tokens are provided, or LID identification performance is sufficiently high (i.e. over 95%). In sum, through this study, we have extended prior evidence from Frisian (Amooie et al., 2025) to a broader set of languages and show the benefit of bilingual training with language identification for low-resource ASR.

6. Bibliographical References

Reihaneh Amooie, Wietse de Vries, Yun Hao, Jelske Dijkstra, Matt Coler, and Martijn Wieling. 2025. Evaluating Standard and Dialectal Frisian ASR: Multilingual Fine-tuning and Language Identification for Improved Low-resource Performance. *arXiv preprint arXiv:2502.04883*.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick Von Platen, Yatharth Saraf, Juan Pino, et al. 2021. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. *arXiv preprint arXiv:2111.09296*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Dongpeng Chen and Brian Kan-Wing Mak. 2015. Multitask Learning of Deep Neural Networks for Low-resource Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(7):1172–1183.

William Chen, Brian Yan, Jiatong Shi, Yifan Peng, Soumi Maiti, and Shinji Watanabe. 2023. Improving Massively Multilingual ASR with Auxiliary CTC Objectives. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Wietse De Vries, Martijn Bartelds, Malvina Nissim, and Martijn Wieling. 2021. Adapting Monolingual Models: Data can be Scarce when Language Similarity is High. *arXiv preprint arXiv:2105.02855*.

Matthew Dryer and Martin Haspelmath. 2024. The world atlas of language structures online.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2024. glottolog/glottolog: Glottolog database 5.1.

Wenxin Hou, Yue Dong, Bairong Zhuang, Longfei Yang, Jiatong Shi, and Takahiro Shinozaki. 2020. Large-scale End-to-End Multilingual Speech Recognition and Language Identification with Multi-task Learning. *Babel*, 37(4k):10k.

Jui-Yang Hsu, Yuan-Jui Chen, and Hung-yi Lee. 2020. Meta Learning for End-to-End Low-resource Speech Recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7844–7848. IEEE.

Qianying Liu, Zhuo Gong, Zhengdong Yang, Yuhang Yang, Sheng Li, Chenchen Ding, Nobuaki Minematsu, Hao Huang, Fei Cheng, Chenhui Chu, et al. 2023. Hierarchical Softmax for End-to-End Low-resource Multilingual

Speech Recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Nay San, Georgios Paraskevopoulos, Aryaman Arora, Xiluo He, Prabhjot Kaur, Oliver Adams, and Dan Jurafsky. 2024. Predicting Positive Transfer for Improved Low-resource Speech Recognition using Acoustic Pseudo-tokens. *arXiv preprint arXiv:2402.02302*.

Austin Waters, Neeraj Gaur, Parisa Haghani, Pedro Moreno, and Zhongdi Qu. 2019. Leveraging Language ID in Multilingual End-to-End Speech Recognition. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 928–935. IEEE.

Søren Wichmann, Eric W Holman, Dik Bakker, and Cecil H Brown. 2010. Evaluating Linguistic Distance Measures. *Physica A: Statistical Mechanics and its Applications*, 389(17):3632–3639.

Genta Indra Winata, Guangsen Wang, Caiming Xiong, and Steven Hoi. 2020. Adapt-and-Adjust: Overcoming the Long-tail Problem of Multilingual Speech Recognition. *arXiv preprint arXiv:2012.01687*.

Yubei Xiao, Ke Gong, Pan Zhou, Guolin Zheng, Xiaodan Liang, and Liang Lin. 2021. Adversarial Meta Sampling for Multilingual Low-resource Speech Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14112–14120.

Mu Yang, Andros Tjandra, Chunxi Liu, David Zhang, Duc Le, and Ozlem Kalinli. 2023. Learning ASR Pathways: A Sparse Multilingual ASR Model. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

7. Language Resource References

Ardila, Rosana, and others. 2019. *Common voice: A massively-multilingual speech corpus.* Mozilla Foundation. Mozilla Foundation, Common Voice, 17.0.

Leveraging Speech Models for Audio-based Lexical Retrieval in Dictionaries: the Case of the Teochew Language

Siman Chen, Ilaine Wang, Maxime Fily, Pierre Magistry
ERTIM, Inalco
Paris, France
{siman.chen, ilaine.wang, maxime.fily, pierre.magistry}@inalco.fr

Abstract

This study presents our attempt to apply Query by Example - Spoken Term Detection methodologies to a real-world, low-resource scenario: building an audio-based query functionality for the Teochew dictionary WhatTCSay. This functionality enables users to retrieve dictionary entries without prior knowledge of the writing systems in Teochew, thereby enhancing the accessibility of the dictionary and facilitating language revitalization efforts within Teochew communities. To address the retrieval task, we investigate two approaches: (i) an Automatic Speech Recognition (ASR)-based approach using text-to-text matching, and (ii) a Dynamic Time Warping (DTW)-based acoustic framework for audio-to-audio retrieval. In the first approach, we compare an automatic romanization of the spoken query against the gold romanization from the dictionary; in the second, we directly match the user's spoken query against audio recordings from the dictionary pronounced by a native speaker. Retrieval performance is evaluated using recall at rank k. Results show that text-to-text matching achieves better performance than audio-to-audio matching; however, the two approaches were not optimized under fully comparable conditions, as the ASR-based approach benefited from additional optimization, which was not equally available for the DTW method.

Keywords: Speech Models, Low-Resource Languages, Automatic Speech Recognition, Dynamic Time Warping, Teochew Language, Query-by-Example, Spoken Term Detection

1. Introduction

Dictionaries represent an invaluable linguistic resource for any language. Yet, looking up a word in a dictionary is far from a trivial task and actually requires a certain level of prior knowledge: reading and writing skills, but also familiarity with orthographic conventions, the ability to identify lemmas, and, in the case of paper dictionaries, knowledge of language-specific sorting rules. Such prerequisites are often taken for granted for languages with a well-established written tradition and institutional support, where formal school plays a central role. This assumption, however, does not hold for minority languages, and, as a matter of fact, for the vast majority of the world's languages.

Heritage languages are particularly affected in this regard, as they are typically spoken at home, transmitted orally across generations, and associated with a limited vocabulary and lower prestige.

In order to make dictionaries accessible to speakers who lack the aforementioned skills, we propose a speech-based search functionality. Our approach builds on Query-by-Example Spoken Term Detection and is applied to Teochew, a Sinitic language for which the diaspora community has already developed a dedicated dictionary.

2. Background

2.1. Teochew as a Multivariational Heritage Language

Teochew is a Sinitic language belonging to the Southern Min branch of the Sino-Tibetan family, spoken primarily in eastern Guangdong, China.

From the 18th to the 20th centuries, successive waves of migration have spread the Teochew people over the world, resulting first in large diaspora communities across Southeast Asian countries such as Thailand, Cambodia, Singapore and Malaysia, and then in Western countries, including the United States, France, and Australia (Live, 1995; McFarland, 2021).

Based on several accounts, Tan (2020) estimates the number of speakers to be between 5 and 7 million in Thailand, the largest Teochew community in Southeast Asia, and between 80,000 and 150,000 speakers in France. These figures should be treated with caution as they stem from outdated sources[1] and may not reflect the current situation, especially since population censuses do not target spoken languages but rather ethnicity, and having Teochew origins does not necessarily imply Teochew language proficiency.

[1] Estimates for Thailand were taken from sources dating from 2001 and 2004, while the lowest estimates for France date back to 1989. See Tan (2020) for more details on these questions.

Proceedings of SPEAKABLE @ LREC 2026, pages 139–149
11 May 2026.

Teochew is indeed a heritage language (HL) in all of those countries, which means that one should not presume that the Teochew language is passed down to the younger generations. A HL is often defined as a home language (Valdés, 2000), as opposed to a majority language (Montrul, 2010), and “is no longer the present dominant language where [the speaker] lives” (ElHawari, 2020). While HLs are often studied in the context of migration, in our case, Teochew is also a HL in China, spoken at home, as opposed to Mandarin.

As a result of its worldwide spread and the lack of standardization, distinct varieties have emerged between speakers in the language’s homeland in southern China as well as in the global diaspora. Those varieties are shaped by long-term language contact and often exhibit phonetic interference from dominant languages as well as substantial lexical borrowing. While in China interference mostly occurs with Mandarin, in the diaspora we can observe traces of languages including Thai, Khmer, Malay, French and English.[2]

2.2. Teochew as a Language with Multiple Writing Systems

Despite being predominantly used and transmitted as an oral language, Teochew is not a non-written language. In fact, multiple writing systems are used by Teochew speakers: sinograms and the Latin-based transcriptions (or romanizations).

Sinograms. As a Sinitic language, Teochew can be written using sinograms. Yet, unlike Mandarin and Cantonese, Teochew has not benefited from a sustained effort to maintain its written form in sinograms, resulting in a heavily incomplete character set and the lack of consensus for which characters should be used for many words. Such cases include colloquial words that are not used in Mandarin, such as /ta˧pɔʊ˧/ ‘boy’[3], or loanwords such as /ma˧ta˧˥/ ‘the police’, which comes from Malay.

Romanization. Multiple Latin-based transcription systems co-exist for Teochew. Historically, the *pe̍h-ūe-jī system* (also called Swatow Church Romanization) created by missionaries in the late 19th century is the first complete romanization system developed for Teochew. It was later followed by the *Guangdong Peng’im*[4] (GD), created by the Provincial Education Department of Guangdong in 1960 and based on *hànyǔ pīnyīn* used for Mandarin. On their Discord server, the international Teochew diaspora community encourages the use of *Gaginang Peng’im* (GGN), a system based on the GD but with a few modifications to better fit the pronunciation of speakers of the Southeast Asian diaspora. This version notably abandons the diacritically marked vowel <ê>, making it more accessible for typing across different keyboard layouts (see Table 1 for a full comparison of GD and GGN). As an example, the word *jap8 ghueh4* in GGN (‘October’) would be *zab8 ghuêh4* in GD. In both systems, the eight tones are represented using numbers.

The transcription used in the Teochew dictionary WhatTCSay is GGN, while our training set is transcribed using GD (see 4.1). In our text-based pipeline, we include a converter from GD to GGN on the transcription output.

initials		nuclei		codas	
GD	GGN	GD	GGN	GD	GGN
b	b	a	a	**b**	**p**
p	p	ai	ai	**g**	**k**
bh	bh	ao	ao	m	m
g	g	**e**	**eu**	h	h
gh	gh	**ê**	**e**	ng	ng
d	d	i	i	n	n
t	t	ia	ia		
s	s	io	io		
z	**j**	u	u		
c	**ch**	ua	ua		
r	**y**	uai	uai		
l	l	**uê**	**ue**		
m	m	ui	ui		
n	n	o	o		
ng	ng	oi	oi		
h	h	ou	ou		

Table 1: Romanization mapping between Guangdong Peng’im (GD) and Gaginang Peng’im (GGN). Non-matching graphemes are marked in bold.

Several other romanization systems exist, though they remain less widely adopted. In her dissertation on Teochew, (Tan, 2020) prefers using *cháoyǔ pīnyīn*, a transcription system created in 2010 by Chen Enquan for their dictionary. Birnie-Smith (2016) reports that Teochew Indonesians are transcribing Teochew using a system based on their usage of the Latin alphabet for Malay. Such *ad hoc* informal adaptations are widely used in the diaspora, as observed in Teochew online

[2]See McFarland (2021, 2022) for an account of studies on Southeast Asia Teochew varieties, and a thorough study of Cambodian Teochew, and Tan (2020) on the possible influence of French.

[3]See the 9 alternative forms for this word in https://en.wiktionary.org/wiki/%E4%B8%88%E5%A4%AB#Pronunciation_3 which include Hokkien-based and Teochew-based propositions.

[4]The word *peng’im* is the Teochew counterpart of *pinyin* (拼音). In this article, we use *peng’im* for any romanization.

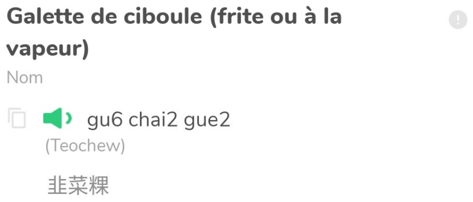


Figure 1: Screenshot of the entry for *gu6 chai2 gue2* ('chive cakes') in the French Android version of WhatTCSay3.

communities. This is one of the reasons certain communities are encouraging the use of either GD or GGN, to maximize mutual intelligibility among speakers from different regions.

2.3. WhatTCSay, a Teochew dictionary app

WhatTeochewSay (WhatTCSay) is a dictionary app for mobile phones originally crowdfunded in 2012[5] and is collaboratively developed by heritage speakers in North America and France. The current release of the app (WhatTCSay3) contains more than 6,700 entries. Each entry is composed of a word (or expression) transcribed in GGN, with a definition (either in English or French) and a part of speech. Optionally, some entries also have the corresponding sinograms, and an audio recorded by the founder. Figure 1 shows a complete entry as an example. The recording feature (indicated by the megaphone icon) is crucial for speakers who cannot read peng'im, but is only available for about 65% of the dictionary in the current version.

The next release will contain more than 10k entries and take into account accents from different regions of Guangdong as well as from different diaspora communities. To keep up with this extension, automatically generated audio will be provided for entries lacking a native speaker recording, based on the text-to-speech system developed by Magistry et al. (2024).

WTCS is a popular app among the diaspora, with almost 20k downloads in total reported by the developer in 2026. Despite encouragement to learn and use either peng'im system in the online communities, most Teochew speakers do not have proficiency in those systems, and have difficulties in reading the peng'im in the app for words that do not have an audio recording associated. It is also almost impossible for them to guess the correct spelling in the case they want to look up a word they heard. Building an audio-based lexical retrieval feature is the next step in making the app definitely accessible.

In this paper, we explore two different strategies to identify dictionary entries based on audio queries. The first approach adopts an Automatic Speech Recognition (ASR)-based pipeline in which audio queries are transcribed into peng'im and subsequently matched against candidate textual entries. The second approach relies on direct audio-to-audio matching using acoustic embeddings extracted from self-supervised speech models within a Dynamic Time Warping (DTW)-based framework.

Our main contributions are as follows: 1) a comparison of ASR-based and DTW-based methods for speech lexicon retrieval; 2) an evaluation on real-world Teochew speech data; 3) a demonstration of how self-supervised speech representations can support low-resource query-by-example retrieval.

3. Query-by-Example Spoken Term Detection

The speech and lexical retrieval task in this paper is redefined as a Query-by-Example Spoken Term Detection (QbE-STD) task. The only difference is that instead of matching all audio documents from a corpus which contain a spoken query provided by a user (Hazen et al., 2009), we match the spoken query directly with a corpus of isolated words. Existing approaches generally follow two paradigms: (1) acoustic matching, where speech features are extracted from audio signals and aligned with query representations (Naik et al., 2020; Le Ferrand et al., 2021; San et al., 2021), and (2) ASR-based pipelines, where audio is first transcribed and then matched using text-based retrieval methods (Parada et al., 2009; Lee et al., 2015; Macaire et al., 2022).

3.1. Dynamic Time Warping

The acoustic approach typically adopts a two-stage framework: extracting frame-level acoustic features from both queries and target audio, and computing a detection score for each query–target pair, typically using DTW-based template matching (San et al., 2021). DTW is an algorithm for measuring the similarity between two temporal sequences that may vary in speed or duration, which is widely adopted in speech recognition (Sakoe and Chiba, 1978).

Previous studies by Le Ferrand et al. (2020) on two endangered languages—Mboshi (Congo) and Kunwinjku (Australia) showed that classical acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs) and perceptual linear prediction

[5] https://www.theteochewstore.org/blogs/latest/123903619-whattcsay-teochew-language-learning-app-now-available-for-free-the-story-behind

(PLP) features, can outperform neural and self-supervised representations from *Wav2Vec* models trained on those languages. This is likely due to the extremely limited amount of available training data, which constrained the effectiveness of self-supervised representation learning.

San et al. (2021) investigated spoken term detection using data from seven Australian Aboriginal languages and a regional variety of Dutch. They systematically evaluated feature extraction approaches based on *Wav2Vec 2.0* representations from both the English monolingual and multilingual *XLSR-53* models. They subsequently implemented a DTW-based detection stage. Results showed that embeddings from the 11th Transformer layer of the English *Wav2Vec 2.0* model achieved the best retrieval accuracy, outperforming MFCC and bottleneck features by 56–86%, even under mismatched speaker and recording conditions.

Their findings suggest that representations extracted from a model trained on a single language or a set of phonologically similar languages may be more beneficial for QbE-STD than a large multilingual model such as *XLSR-53* trained on a diverse set of 53 languages. This is particularly when supervised fine-tuning in the target language is not readily feasible.

3.2. Automatic Speech Recognition

Another approach to the QbE-STD task is based on ASR technique. ASR refers to the use of machines to convert human speech into corresponding text (Benzeghiba et al., 2007). With the rapid advancements in deep learning and the availability of large-scale datasets (Karpagavalli and Chandra, 2016; Ardila et al., 2020), state-of-the-art ASR systems now deliver high-precision transcription for resource-rich languages, including English, Mandarin, and others (Radford et al., 2023; Seed-ASR, 2024).

However, comparable performance in low-resource settings has only recently become achievable through large-scale and multilingual models (Pratap et al., 2024). Recent advancements in self-supervised multilingual speech models have led to a popular paradigm for solving low-resource speech recognition problems. Self-supervised learning (SSL) enables the model to learn robust acoustic representations from unlabeled data during pretraining (Conneau et al., 2021). Several studies have demonstrated that multilingual ASR architectures, such as *Wav2Vec 2.0 XLS-R* (Babu et al., 2022) and *Whisper* (Radford et al., 2023), can be effectively fine-tuned on small language-specific datasets to achieve competitive recognition accuracy. In this paper, We experiment with the multilingual *Wav2Vec2 XLS-R-300M* model (Babu et al., 2022), considering both the pre-trained base model and its fine-tuned counterpart trained on our target language Teochew.

Macaire et al. (2022) propose an ASR-driven QbE framework in which speech segments are first transcribed, and subsequently matched with corpus transcriptions using Smith-Waterman algorithm (Lecouteux et al., 2012). Their findings indicate that for french-related creole languages such as Gwadloupéyen and Morisien, a French monolingual model fine-tuned with extremely limited annotated data (as little as 10 minutes) can achieve usable performance, highlighting the potential of self-supervised ASR for low-resource linguistic documentation.

Motivated by recent progress in ASR and speech representation learning, this work investigates two complementary strategies for identifying dictionary entries from audio queries. The first approach adopts an ASR-based pipeline in which spoken queries are transcribed into peng'im and matched against candidate textual entries. The second approach performs direct audio-to-audio matching using acoustic embeddings within a DTW-based framework.

4. Methodology

4.1. Datasets

In this study, we use three different datasets.

Train. For model training, we employ the Teochew Wild corpus[6] (Pan et al., 2025), the first publicly released Teochew speech dataset with sinograms and Guangdong Peng'im (GD), a romanization system presented in Section 2.2. The corpus consists of 18.9 hours of speech collected from online media, including recordings from 20 Teochew speakers from China (11 male and 9 female), with a total of approximately 12,500 sentence-level utterances.

Test. Our test data set for assessing the retrieval results is a set of recordings made by the founder of WhatTCSay (WTCS), a speaker from the diaspora community. The dictionary consists of 4,603 mp3 files, totaling 1.28 hours of speech and 9,423 syllables (Lim et al., 2024). As expected for recordings for a dictionary, most of this data consist of single words pronounced in isolation (see Section 2.3 for more information on the dictionary).

Evaluation. To evaluate the retrieval performance of our system, we constructed an evalua-

[6] https://huggingface.co/datasets/panlr/teochew_wild

tion set consisting of 172 audio stimuli corresponding to 31 different words present in the dictionary.

We originally presented a list of 40 words (definitions in French or in English, transcription in GGN) to Teochew heritage speakers. Words were selected to cover the Teochew's full phonemic inventory at least three times. Speakers were asked to pronounce as many words as possible, depending on their knowledge. We then filtered out noisy samples and samples for which the words were pronounced by less than four speakers, so that we can have comparable data, while ensuring that the resulting 31 words still covered the Teochew phonemic inventory.

These recordings were collected by six Teochew speakers using their own mobile phones, including five from the diaspora and one from China (for a total of three male and three female). The 172 resulting stimuli are distributed as follows: 35 monosyllabic stimuli, 101 disyllabic stimuli and 36 stimuli with three or more syllables. None of them were included in the training data. This reflects realistic usage conditions, and forms an out-of-domain evaluation set.

4.2. Experimental Pipeline

4.2.1. Fine-Tuning

Fine-tuning a pre-trained model is a common approach for low-resource settings. Rather than training a model from scratch, fine-tuning adapts the parameters of a large pre-trained model to the target task using a relatively small amount of labeled data. This approach is particularly effective when the source and target domains are related, allowing the model to transfer previously learned representations while specializing for the new linguistic and acoustic characteristics (Baevski and Mohamed, 2020).

For the current work, we choose to fine-tune *Wav2Vec2-XLS-R-300M* for both ASR transcription and feature extraction. As a multilingual extension of the original *Wav2Vec2* framework, *XLS-R* scales pre-training to 436K hours of speech across 128 languages, enabling the model to learn robust cross-lingual representations and improve generalization in multilingual and low-resource settings (Babu et al., 2022).

All Teochew recordings were loaded and uniformly resampled to 16 kHz to ensure consistent input features. The corresponding transcripts were pre-processed by removing punctuation marks that did not contribute to phonetic realization and could not be represented acoustically. We further normalized the text to retain only linguistic symbols relevant to the Teochew peng'im romanization scheme. The hyperparameters are given in Table 2 and the results are shown in Table 3.

Parameter	Value
pretrained_model	Wav2Vec2-XLS-R-300M
attention_dropout	0.1
hidden_dropout	0.1
feat_proj_dropout	0.1
mask_time_prob	0.05
layerdrop	0.1
ctc_loss_reduction	mean
train_batch_size	16
num_train_epochs	50
fp16	True
learning_rate	3e-4

Table 2: Values of the hyperparameters used to fine-tune the Wav2Vec2-XLS-R-300M model on Teochew.

ASR Backbone	Val WER	Test WER
Wav2Vec2-XLS-R-300M (fine-tuned)	13.51	12.52

Table 3: ASR performance of the fine-tuned Wav2Vec2-XLS-R-300M model on peng'im transcriptions (WER, %).

4.2.2. ASR-Based Method

We employed the fine-tuned *XLS-R* model trained on Guangdong Peng'im (GD) to transcribe our test set, which consists of 172 stimuli produced by six speakers. It has 35 monosyllabic words, 101 disyllabic words, 36 words of three or more syllables.

To enable comparison with the WhatTCSay dictionary, we used the Gaginang Peng'im (GGN) system for consistency. Model outputs in GD were further converted to GGN using a rule-based converter that supports conversion between different Teochew romanization systems[7]. The workflow of this method is illustrated in Figure 2.

We then used the Levenshtein distance (Levenshtein, 1966) to measure the difference between two peng'im sequences. Initial weights were set to their default values, with substitution, insertion, and deletion costs equal to 1.

To better reflect Teochew writing system, we incorporated digraphs adapted to GGN that correspond to single phonemes (e.g., the consonant clusters <bh>, <gh>, <ng>, <ch>, which respectively correspond to the phonemes /b/, /g/, /ŋ/ and /ʃ/, and the vowel unit <eu>, which is not a diphthong but corresponds to /ɯ/).

We further analyzed the ASR outputs generated for the 4,603 audio recordings in the dictionary by comparing them with their reference peng'im transcriptions following the syllabic structure (onset, nucleus, coda). The edit-distance weights were then adjusted to account for frequent phonetic confusions that share articulatory features. We placed

[7] https://github.com/learn-teochew/parsetc

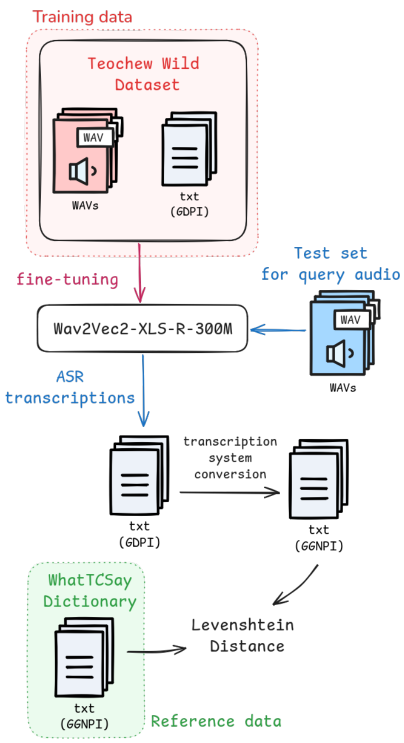


Figure 2: Workflow for the ASR-based methodology

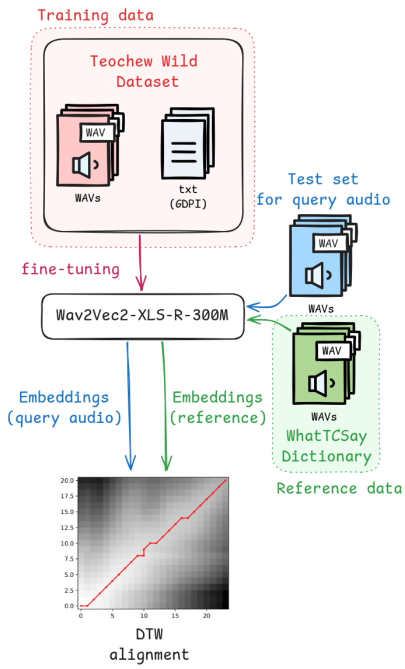


Figure 3: Workflow for the DTW-based methodology

a stronger emphasis on vowel variation, as vowels are known to be highly unstable across pronunciations (Dolgopolsky, 1964). To mitigate the risk of overfitting, six-fold cross-validation was conducted. Based on the resulting performance, the vowel substitution cost was reduced to 0.7. The four most frequent vowel confusion observed in the transcription output (i↔e, a↔o, eu↔o, eu↔a)[8] were assigned further lower weights (0.25). These adjustments also reflect speaker perception, as those contrasts are often considered equivalent by speakers in different dialectal contexts. For example, the word *new* is pronounced /siŋ/ by some speakers and /sɛŋ/ by others.

4.2.3. DTW-Based Method

This method (Figure 3) aims to retrieve dictionary audio entries by direct spoken audio matching without relying on textual representations as an intermediate step. We compare the original *Wav2Vec2-XLS-R* model and the fine-tuned version for feature extraction.

In our setup, no temporal pooling was applied (`pooling = None`) in order to preserve the full sequential structure of the speech representations. This allows DTW to perform frame-level alignment on fine-grained acoustic sequences. Dictionary candidates were subsequently ranked according to their DTW distance, where lower scores indicate higher similarity.

4.3. Evaluation Metrics

For evaluation, we compute Recall at rank k (R@1, R@5, R@10), which measures the proportion of queries for which the correct dictionary entry is present among the top-k retrieved results (Bruno et al., 2002). This metric closely reflects real-world dictionary search scenarios, where the correct entry is expected to appear among the top-ranked results.

5. Results

5.1. ASR Performance

We used the fine-tuned model to transcribe the 172 stimuli and performed top ten retrieval under different query settings. In 74% of cases, the target entry does appear among the top ten results (see Table 4). However, a preliminary analysis revealed that the eight-tone system seems to introduce notable ASR errors. Consequently, we performed a tone-level evaluation by comparing the ASR outputs with the reference peng'im transcriptions. Across 9,343 aligned syllables, the system exhibited a tone error rate of 51.42%. This can be partly attributed to tonal variation across accents, as well as inconsistencies between the training corpus and the dictionary. Indeed, tone sandhi occurs in Teochew and the training corpus is anno-

[8]Where <e> is /ɛ/, <o> is /ɔ/ and <eu> is /ɯ/.

tated with lexical tones (e.g., *jap8 ghueh4*), while the dictionary shows tones after sandhi rules were applied (e.g., *jap4 ghueh4*). Such tonal discrepancies introduce additional noise when computing Levenshtein distance. To mitigate this effect, we removed tone information from the queries and further evaluated a weighted Levenshtein condition. The results for three different query settings are presented in Table 4, where we can see that we do indeed manage to meaningfully improve recall, reaching 82% at recall@10.

ASR Query Setting	R@1	R@5	R@10
Query with tones	0.50	0.67	0.74
Query removing tones	0.55	0.67	0.76
Query removing tones + weighted Levenshtein	**0.57**	**0.77**	**0.82**

Table 4: ASR query retrieval performance under different query normalization settings.

5.2. DTW Performance

We leveraged speech representations extracted from both the pre-trained *XLS-R* model and its fine-tuned counterpart trained on Teochew data. And we applied a dynamic time warping (DTW)–based retrieval approach. Evaluation was conducted on an unseen, out-of-domain teochew test set comprising six speakers not included in the training data. Figure 4 reports the retrieval performance for both models using features extracted from all 24 transformer layers of each model. The fine-tuned model achieves nearly 50% relative improvement compared to the pre-trained baseline.

The base model achieves its best recall scores at intermediate layers, particularly around layer 13. In contrast, the fine-tuned model shows an improvement toward deeper layers, with the final layer yielding the highest recall scores, indicating that task-specific fine-tuning reshapes the representational hierarchy (see Table 5). Interestingly, we also observe that intermediate layers of the fine-tuned model (layers 12–17) perform comparably well in retrieving the correct entry within the top-10 results. Overall, model differences become most pronounced in the final layers where the effects of the learning objective are strong.

Method	R@1	R@5	R@10
XLS-R-300M (layer 13)	0.20	0.33	0.36
XLS-R-300M-ft (layer 24)	**0.42**	**0.59**	**0.63**

Table 5: Comparison between the two speech models from layer 1-24.

When comparing the two methods, the retrieval results show a consistent performance advantage

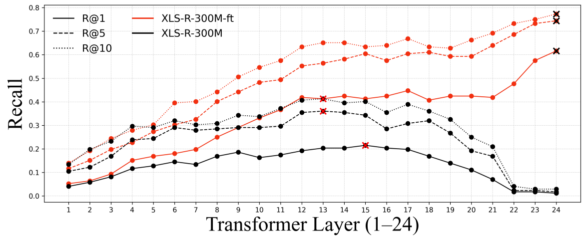


Figure 4: Retrieval results from both models' transformer layer 1-24 using DTW; the crosses indicate highest score achieved on dataset.

Method	R@1	R@5	R@10
ASR-based retrieval	**0.57**	**0.77**	**0.82**
DTW-based retrieval	0.42	0.59	0.63

Table 6: Comparison between ASR-based and DTW-based retrieval on the six-speaker evaluation set.

for the ASR-based method. It yields relative improvements of 35.7%, 30.5%, and 30.2% for R@1, R@5, and R@10 respectively.

Method	1-syll	2-syll	3+-syll
ASR-based retrieval	**0.50**	**0.75**	**0.86**
DTW-based retrieval	0.35	0.58	0.68

Table 7: Average recall (mean of R@1, R@5 and R@10) across syllable-length groups for both methods.

Table 7 illustrates how syllable length affects performance in the DTW-based method. The results show that across all syllable categories, the ASR-based method consistently outperforms the DTW-based method, and both show higher recall as word length increases. Overall, this advantage aligns with our expectations: even when ASR transcription introduces substitutions or deletions, Levenshtein distance can still recover the correct entry by matching partially shared segments. In contrast, DTW is more sensitive to local acoustic distortions and temporal misalignments (Permanasari et al., 2019).

6. Discussion

The ASR-based pipeline produces textual transcriptions that enable Levenshtein-based retrieval, even under partial overlap. This makes the approach more robust to accent variation, and provides more effective retrieval cues than purely acoustic similarity. This is particularly important for Teochew, a multivariational language characterized by pronunciation variation across speakers. However, the ASR-based pipeline requires converting transcription outputs from one peng'im system to another before applying Levenshtein distance, and removing tones, which introduces additional process. In contrast, the DTW-based approach operates directly on speech representations, which has the advantage of bypassing orthographic or writing conventions. However, this method can be highly sensitive to factors such as background noise, volume variation, and pronunciation differences (Permanasari et al., 2019). We further assess the impact of acoustic artifacts on retrieval performance, we selected a subset of stimuli that contained noticeable silence, background noise, or both. The results shows that 8 out of 10 queries showed rank improvement when silence or noise is removed, which indicates that preprocessing has a positive effect.

Real-Time Retrieval. When deploying the system in real-life applications, apart from the evaluation metrics on recall scores, other factors like real-time retrieval and user experience also play an important role. During a workshop organized with the local Teochew community in Paris to test our system, we observed that users were generally satisfied as long as the spoken word appeared in the results page.

In addition, mismatches sometimes arose when speakers pronounced multi-word expressions as a single unit. For instance, *jiahbeung* ('to eat (rice)'), does exists in this version of the dictionary, but as separate entries: *jiah* ('to eat') and *beung* ('rice'). In such cases, prompting users to pronounce words separately after an unsuccessful search could improve usability.

Another possible strategy to reduce frustration is to offer two options after each query: repeating the same word or pronouncing a different one. Repeated attempts could potentially be leveraged to refine the retrieval process by incorporating multiple trials of the same spoken query.

Layer-Wise Analysis of SSL Representations. Our findings further suggest that the DTW-based retrieval method relying on speech representations is not as fully unsupervised as initially assumed. Using representations extracted from the fine-tuned model trained on Teochew consistently outperform those from the base model. This suggests that the ASR fine-tuning objective may contribute positively to the audio lexicon retrieval task, particularly when the fine-tuning language aligns with the target language.

Our DTW-based method shows that layers contain useful abstractions and generalizations of acoustic information. Performance varies across layers, indicating that the choice of representation level has an impact on retrieval quality.

For both the pre-trained *XLS-R* model and its fine-tuned counterpart, intermediate layers perform well in retrieving the correct entry within the top-10 results (see Figure 4). This is consistent with previous layer-wise analyses of self-supervised speech models, which suggest that middle layers often encode strong phonetic information Pasad et al. (2021) and may serve as competitive representations for speech-related tasks, and that the last layer might not always be the optimal choice (Bartelds et al., 2022; Cho et al., 2023). For example, Hao et al. (2024) report that the optimal layer for acoustic-to-articulatory inversion (AAI) task is typically located around two-thirds of the model depth.

However, the optimal layer to use for downstream evaluation may vary depending on factors such as optimization, data, and downstream task (Bordes et al., 2023). Further investigation is required to better understand these layer effects, particularly in low-resource settings.

We also observed that retrieval performance differs dramatically between models in the final layers. This behavior may be linked to ASR fine-tuning, during which the final layers become tailored to the ASR learning objective that learn to differentiate between different phonemes. Such task-oriented representations may positively transfer to the query-by-example retrieval setting, where distinguishing fine-grained phonetic differences is essential.

7. Conclusion and Perspectives

This study addresses the Teochew language, an under-resourced Sinitic language within the NLP community. We explore different strategies to identify dictionary entries based on audio queries.

The first approach uses an ASR-based pipeline to transcribe audio queries into peng'im for lexical matching, while the second performs direct audio-to-audio matching using DTW over self-supervised speech embeddings. Overall, the ASR-based method achieves superior performance.

Despite requiring additional linguistic normalization, the ASR-based method offers higher interpretability and enables integration with downstream text-based applications. In future work, it would be interesting to incorporate a language model during ASR fine-tuning to further constrain and refine the transcription outputs.

Based on ASR error patterns, we further adapted the Levenshtein distance to better account for vowel variability, which also reflects phonetic variation observed across dialectal pronunciations. As a result, this pipeline becomes less out-of-domain and more guided by supervised linguistic knowledge compared to the DTW-based approach. These observations also raise questions regarding how to incorporate similar vowel-weighting strategies into DTW-based matching.

The DTW-based method, however, has the advantage of bypassing orthographic inconsistencies and spelling variation, which are common challenges for under-resourced languages lacking standardized writing systems. By operating directly on speech representations, our results show that embeddings extracted from the fine-tuned *XLS-R* model significantly outperform those from the pre-trained base multilingual model, which suggests that ASR fine-tuning may facilitate positive transfer to query-by-example retrieval tasks.

These findings also suggest that DTW-based retrieval is not as fully unsupervised as initially assumed, as performance remains strongly influenced by supervised ASR fine-tuning. A layer-wise analysis further reveals that the pre-trained model reaches peak retrieval performance around layer 13, whereas the fine-tuned model achieves peak performance at deeper layers (around layers 15–17 and the final layer). These observations motivate future work exploring detailed error analysis and visualization of phonetic features encoded in these layers, as well as strategies such as truncating or reinitializing higher transformer layers prior to fine-tuning.

8. Limitations

This work does not explicitly investigate how tonal information is represented or modeled. Future research should examine how ASR systems capture tonal features in Sinitic languages, such as contour tones and tone sandhi. This can provide insights into 'unretrievable' queries.

9. Acknowledgments

This work is funded by the DiLSi ANR project ANR-23-CE38-0004-01. It was granted access to the HPC/AI resources of [CINES/IDRIS/TGCC] under the allocation 2025-AD011014016R1 made by GENCI.

We are also grateful to the "Les Jeunes Teochew de France" association in Paris, for allowing us to recreate the "salle mauve" on their Discord for our recording needs. We would like to thank especially the members who generously contributed to the construction of the evaluation set used in the experiments.

10. Bibliographical References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proceedings of Interspeech 2022*, Incheon, South Korea.

Alexei Baevski and Abdelrahman Mohamed. 2020. Effectiveness of Self-supervised Pre-Training for ASR. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 7694–7698. IEEE.

Martijn Bartelds, Wietse de Vries, Faraz Sanal, Caitlin Richter, Mark Liberman, and Martijn Wieling. 2022. Neural Representations for Modeling Variation in Speech. *Journal of Phonetics*, 92:101137.

Mohamed Benzeghiba, Renato De Mori, Olivier Deroo, Stephane Dupont, Teodora Erbes, Denis Jouvet, Luciano Fissore, Pietro Laface, Alfred Mertins, Christophe Ris, et al. 2007. Automatic Speech Recognition and Speech Variability: A Review. *Speech communication*, 49(10-11):763–786.

Jessica Rae Birnie-Smith. 2016. Ethnic Identity and Language Choice across Online Forums. *International Journal of Multilingualism*, 13(2):165–183.

Florian Bordes, Randall Balestriero, Quentin Garrido, Adrien Bardes, and Pascal Vincent. 2023. Guillotine regularization: Why removing layers is needed to improve generalization in self-supervised learning. *Transactions on Machine Learning Research*.

Nicolas Bruno, Surajit Chaudhuri, and Luis Gravano. 2002. Top-k Selection Queries over Relational Databases: Mapping Strategies and Performance Evaluation. *ACM Transactions on Database Systems (TODS)*, 27(2):153–187.

Cheol Jun Cho, Peter Wu, Abdelrahman Mohamed, and Gopala K Anumanchipalli. 2023. Evidence of Vocal Tract Articulation in Self-Supervised Learning of Speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Proceedings of Interspeech 2021*, pages 2426–2430.

Aron B. Dolgopolsky. 1964. Gipoteza drevnejšego rodstva jazykovych semej severnoj evrazii s verojatnostej točky zrenija [a probabilistic hypothesis concerning the oldest relationships among the language families of northern eurasia]. *Voprosy jazykoznanija*, 2:53–63.

Rasha ElHawari. 2020. *Teaching Arabic as a heritage language*. Routledge.

Yun Hao, Reihaneh Amooie, Wietse de Vries, Thomas Tienkamp, Rik van Noord, and Martijn Wieling. 2024. Exploring Self-Supervised Speech Representations for Cross-lingual Acoustic-to-Articulatory Inversion. In *Interspeech 2024*, pages 4603–4607. ISCA.

Timothy J. Hazen, Wade Shen, and Christopher White. 2009. Query-by-Example Spoken Term Detection using Phonetic Posteriorgram Templates. In *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 421–426. IEEE.

Shunmugam Karpagavalli and Edy Chandra. 2016. A Review on Automatic Speech Recognition Architecture and Approaches. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 9(4):393–404.

Eric Le Ferrand, Steven Bird, and Laurent Besacier. 2020. Enabling interactive transcription in an indigenous community. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3422–3428.

Eric Le Ferrand, Steven Bird, and Laurent Besacier. 2021. Phone Based Keyword Spotting for Transcribing Very Low Resource Languages. In *Proceedings of the 19th Annual Workshop of the Australasian Language Technology Association*, pages 79–86, Online. Australasian Language Technology Association.

Benjamin Lecouteux, Georges Linares, and Stanislas Oger. 2012. Integrating imperfect transcripts into speech recognition systems for building high-quality corpora. *Computer Speech & Language*, 26(2):67–89.

Lin-Shan Lee, James Glass, Hung-Yi Lee, and Chun-An Chan. 2015. Spoken Content Retrieval-Beyond Cascading Speech Recognition with Text Retrieval. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(9):1389–1420.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Yu-Sion Live. 1995. Les Chinois de Paris: groupes, quartiers et réseaux. In Antoine Marès and Pierre Milza, editors, *Le Paris des étrangers depuis 1945*, pages 343–357. Éditions de la Sorbonne.

Cécile Macaire, Didier Schwab, Benjamin Lecouteux, and Emmanuel Schang. 2022. Automatic speech recognition and query by example for creole languages documentation. *Findings*.

Pierre Magistry, Ilaine Wang, and Ty Eng Lim. 2024. Experiments on Speech Synthesis for Teochew, Can Taiwanese Help? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6849–6854.

Joanna Rose McFarland. 2021. Language contact and lexical changes in Khmer and Teochew in Cambodia and Beyond. In Tom Hoogervorst and Caroline Chia, editors, *Sinophone Southeast Asia: Sinitic Voices across the Southern Seas*. Brill.

Joanna Rose McFarland. 2022. *Aspects of Cambodian Teochew Grammar: A Radical Construction Grammar Account*. Ph.D. thesis, Nanyang Technological University.

Silvina Montrul. 2010. Current Issues in Heritage Language Acquisition. *Annual review of applied linguistics*, 30:3–23.

Prajyot Naik, Manisha Naik Gaonkar, Veena Thenkanidiyoor, and Aroor Dinesh Dileep. 2020. Kernel based Matching and a Novel training approach for CNN-based QbE-STD. In *2020 international conference on signal processing and communications (SPCOM)*, pages 1–5. IEEE.

Carolina Parada, Abhinav Sethy, and Bhuvana Ramabhadran. 2009. Query-by-Example Spoken Term Detection for OOV terms. In *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 404–409. IEEE.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921. IEEE.

Yurika Permanasari, Erwin H. Harahap, and Erwin Prayoga Ali. 2019. Speech Recognition using Dynamic Time Warping (DTW). *Journal of physics: Conference series*, 1366(1):012091.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2024. Scaling Speech Technology to 1,000+ Languages. *Journal of Machine Learning Research*, 25(97):1–52.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Hiroaki Sakoe and Seibi Chiba. 1978. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49.

Nay San, Martijn Bartelds, Mitchell Browne, Lily Clifford, Fiona Gibson, John Mansfield, David Nash, Jane Simpson, Myfany Turpin, Maria Vollmer, et al. 2021. Leveraging pre-trained representations to improve access to untranscribed speech from endangered languages. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1094–1101. IEEE.

Seed-ASR. 2024. Seed-ASR: Understanding Diverse Speech and Contexts with LLM-based Speech Recognition. Technical report, ByteDance.

My Dung Adeline Tan. 2020. *L'expression du déplacement en chaozhou : les formes introduisant un groupe nominal locatif et l'encodage de la trajectoire*. Ph.D., Institut National des Langues et Civilisations Orientales, Paris.

Guadalupe Valdés. 2000. Spanish for Native Speakers: AATSP Professional Development Series Handbook for Teachers K-16 (Vol. 1). *New York*.

11. Language Resource References

Lim, Ty Eng and Wang, Ilaine and Magistry, Pierre. 2024. *Audio files from WhatTCSay 3*. Zenodo.

Linrong Pan, Chenglong Jiang, Gaoze Hou, and Ying Gao. 2025. Teochew-wild: The first in-the-wild teochew dataset with orthographic annotations. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

Stage-Aware Cross-Lingual Transfer for Faroese ASR: When and Which Languages Matter

Dávid í Lág[1], **Barbara Scalvini**[1], **Carlos Mena**[2], **Jón Guðnason**[3]
[1]Faculty of Science and Technology, University of the Faroe Islands, Faroe Islands
[2]Language Technologies Laboratory, Barcelona Supercomputing Center (BSC), Spain
[3]Department of Engineering, Reykjavik University, Iceland
{davidl, barbaras}@setur.fo, carlos.hernandez@bsc.es, jg@ru.is

Abstract

Automatic speech recognition (ASR) for low-resource languages remains challenging due to limited labeled data. Although multilingual models and the inclusion of related auxiliary languages enable cross-lingual transfer, it is still unclear how introducing cross-lingual information at different training stages-pre-training versus fine-tuning-affects downstream performance. Prior work largely treats transfer as a single-stage optimization problem without disentangling stage effects. We present a stage-aware analysis of cross-lingual transfer for Faroese ASR using related auxiliary languages and Wav2Vec 2.0 XLS-R models. We systematically compare two complementary adaptation pipelines: (i) cross-lingual supervised fine-tuning and (ii) cross-lingual continuous pre-training prior to fine-tuning. Both strategies are evaluated under a unified setup with controlled model architectures, balanced representation of auxiliary languages, and identical evaluation protocols. Results demonstrate that cross-lingual transfer is stage-dependent. Supervised adaptation optimizes in-domain accuracy, while pretraining-level adaptation enhances robustness and reduces Character Error Rate (CER). Auxiliary language effects vary across pipelines, reinforcing the idea that transfer effectiveness depends on when and how cross-lingual information is introduced. Comparisons with large-scale multilingual ASR models highlight trade-offs between model scale and explicit, small-scale domain-aware adaptation. These findings suggest that effective cross-lingual transfer for Faroese low-resource ASR is inherently stage-dependent rather than a single-step design choice.

Keywords: Stage-aware cross-lingual transfer, Low-resource, Domain adaptation, Faroese language technology

1. Introduction

ASR for low-resource languages remains challenging due to limited annotated data, domain mismatch between training and deployment, and constraints on large-scale supervised learning (Baevski et al., 2020b; Conneau et al., 2020; Yadav and Sitaram, 2022; Geng et al., 2025; Müller et al., 2024). Faroese, a low-resource Scandinavian language and spoken by approximately 70,000 speakers, exemplifies these challenges: despite recent efforts Hernández Mena and Simonsen (2022) to create high-quality labeled resources, available data remain insufficient to train ASR systems that generalize beyond clean, read-speech conditions.

Recent advances in multilingual and self-supervised speech models have lowered the barrier for low-resource ASR by enabling cross-lingual transfer from high-resource or related languages (Baevski et al., 2020b; Babu et al., 2021; Radford et al., 2022b; Yadav and Sitaram, 2022). Two broad adaptation strategies are commonly used. Supervised fine-tuning directly adapts a pre-trained model to the ASR objective on the target language, optionally incorporating related languages, whereas continuous self-supervised pre-training reshapes acoustic representations prior to supervised fine-tuning. Although both strategies are well established, they are rarely compared within a unified framework for low-resource languages (Baevski et al., 2020b; Conneau et al., 2020; Babu et al., 2021; Yi et al., 2021).

Most existing work evaluates either supervised multilingual fine-tuning (Cho et al., 2018; Gupta and Boulianne, 2022; Bekarystankyzy et al., 2024; Yi et al., 2021; Williams et al., 2023) or continuous self-supervised pre-training and acoustic adaptation (Baevski et al., 2020b; Conneau et al., 2020; Babu et al., 2021; mag, 2022) in isolation, typically reporting final error rates without distinguishing between the inductive effects of cross-lingual transfer introduced during self-supervised pre-training versus supervised fine-tuning. As a result, it remains unclear whether observed gains stem from improved acoustic invariance, better phonetic alignment, or favorable domain match. This lack of stage-aware analysis forces practitioners to rely on costly trial-and-error when designing multilingual adaptation pipelines.

Moreover, auxiliary language selection adds further uncertainty to multilingual adaptation. In practice, choices are often guided by intuition or prior empirical results rather than principled criteria, making language selection and stage selection intertwined design variables. This lack of systematic guidance increases computational cost and obscures the mechanisms underlying observed trans-

Proceedings of SPEAKABLE @ LREC 2026, pages 150–161
11 May 2026.

fer gains.

To address this gap, this study investigates cross-lingual transfer strategies within the *Wav2Vec 2.0* (Baevski et al., 2020b) model family. We argue that the training stage at which cross-lingual information is introduced critically shapes transfer behavior in low-resource ASR. We therefore compare two controlled, stage-aware pipelines for Faroese ASR: (A) multilingual transfer introduced during supervised fine-tuning, and (B) multilingual transfer introduced during continuous pre-training prior to Faroese supervision. Both pipelines are evaluated against monolingual counterparts under comparable data budgets, architectures, and protocols. While both pipelines use the same underlying datasets, the allocation of data differs across training stages (self-supervised vs. supervised), reflecting their respective training strategies. Evaluation is conducted using standard Faroese test data as well as a newly constructed parliamentary speech benchmark designed to reflect real-world deployment conditions.

Our results show that the two pipelines exhibit distinct and complementary performance profiles. *Pipeline A* effectively consolidates Faroese-specific phonotactics (language-specific sound sequence constraints) and performs well on clean, in-domain speech, but remains sensitive to domain mismatch. In contrast, *Pipeline B* substantially improves robustness to acoustic variability, speaking style, and domain-specific vocabulary, yielding lower error rates on parliamentary speech. These findings suggest that cross-lingual transfer is stage-dependent, with different training stages benefiting from different types of linguistic and acoustic similarity.

To contextualize these findings, we further compare our stage-aware pipelines to two recent large-scale multilingual ASR models trained under different paradigms, using *Whisper* (Radford et al., 2022b) and *Omnilingual* (team et al., 2025) as baselines. This comparison highlights the complementary roles of size and explicit language adaptation, reinforcing the importance of training-stage and language-choice considerations in low-resource ASR.

The contributions of this paper are threefold: (1) a unified comparison of cross-lingual transfer introduced during supervised fine-tuning and during continuous self-supervised pre-training for Faroese ASR; (2) the creation of a new challenging out-of-domain Faroese parliamentary benchmark for robust evaluation; and (3) the release of newly trained Wav2Vec 2.0 models for Faroese, together with the full training and evaluation pipeline to support reproducibility and future research. While our experiments focus on Faroese, the experimental setup can serve as a blue print for low resource ASR optimization for other under-resourced languages with higher resource related languages.

The paper is organized as follows. Section 2 reviews related work. Section 3 describes the datasets and benchmark construction. Section 4 presents the stage-aware transfer pipelines. Section 5 reports empirical findings, followed by analysis in Section 6. Finally, we summarize our findings (Section 7) and discuss limitations of this study (Section 8).

2. Background and Related Work

Modern ASR systems remain highly dependent on large labeled datasets, which limits their applicability to low-resource languages. This has driven research into multilingual training, transfer learning, and self-supervised learning, where models pre-trained on large amounts of unlabeled speech learn transferable representations that can be adapted with limited supervision (Besacier et al., 2014; Thomas et al., 2012). Large multilingual corpora such as Common Voice Ardila et al. (2020) have further supported this paradigm, which underlies contemporary ASR models including Wav2Vec 2.0, Whisper, and Omnilingual.

2.1. Multilingual end-to-end models: Wav2Vec 2.0, Whisper, and Omnilingual

Introduced in 2020, Wav2Vec 2.0 is a transformer-based (Vaswani et al., 2017), end-to-end self-supervised learning framework for ASR, significantly reducing reliance on large annotated data sets (Baevski et al., 2020b). Wav2Vec 2.0 employs quantized speech representations, capturing essential acoustic features across languages (Baevski et al., 2020a), facilitating multilingual learning and leveraging high-resource languages to improve ASR performance for low-resource languages (Getman et al., 2024; Williams et al., 2023). XLSR-53, a variant pretrained on 53 languages (Conneau et al., 2020), particularly demonstrates Wav2Vec 2.0's strong transfer learning capabilities through high-dimensional embeddings from large-scale multilingual training (Bekarystankyzy et al., 2024; Cho et al., 2018; Gupta and Boulianne, 2022). More recently, the XLS-R variant extended multilingual pretraining to 128 languages using a significantly larger data set (Babu et al., 2021) including 64 hours of Faroese audio from YouTube.

Whisper, released in 2022, is a multilingual ASR model trained on 680,000 hours of labeled data covering 99 languages (Radford et al., 2022a). Faroese is not included in its training data. Its encoder–decoder Transformer architecture supports both speech recognition and speech trans-

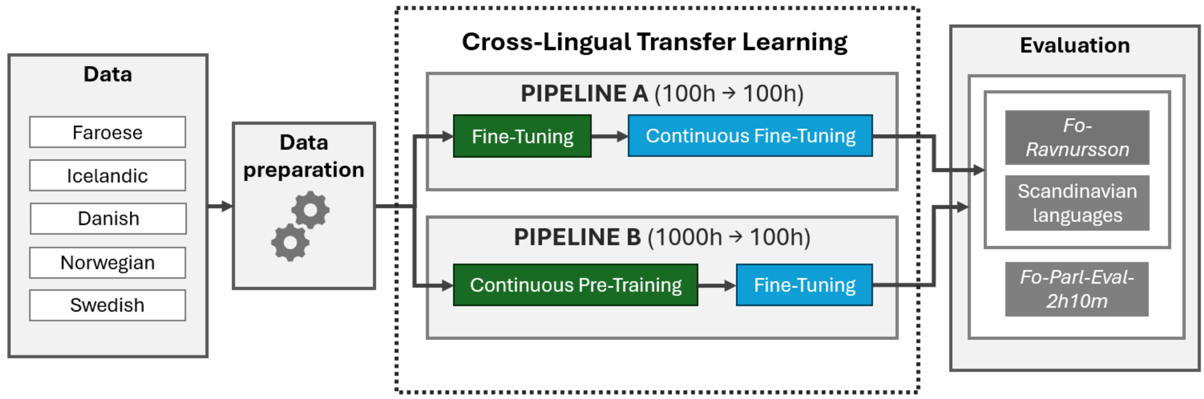


Figure 1: Stage-aware cross-lingual transfer framework for Faroese ASR. The diagram shows four components: (1) multilingual datasets, (2) unified data preparation, (3) two alternative adaptation pipelines, and (4) evaluation on in-domain and out-of-domain Faroese test sets. Pipeline A performs supervised cross-lingual fine-tuning followed by Faroese-only continued fine-tuning, whereas Pipeline B introduces cross-lingual information earlier through continuous pre-training before Faroese fine-tuning. Green blocks indicate stages where auxiliary or multilingual data are used; blue blocks denote Faroese-only supervised training.

lation. Whisper is trained end-to-end using supervised learning on diverse real-world data, improving robustness to noise, accents, and speaking styles.

Recent work has extended multilingual ASR beyond fixed language inventories through the Omnilingual ASR framework, released in November 2025 (team et al., 2025). The framework scales Wav2Vec 2.0-style self-supervised learning to models of up to 7B parameters, pretrained on over 4.3 million hours of speech spanning more than 1,600 languages (team et al., 2025).

While these models demonstrate that large-scale multilingual exposure is effective, they offer limited insights into how different languages affect each other, and when cross-lingual information is most beneficial to a specific low-resource language, therefore requiring an extra, empirical optimization step for low-resource adaptation.

2.2. ASR for Faroese

Faroese ASR development accelerated with the Ravnur project, which produced the Faroese *BLARK* and the 100-hour Ravnur corpus covering multiple dialects and age groups (Simonsen et al., 2022). Using this data, fine-tuned multilingual models such as Wav2Vec 2.0 achieved the first competitive Faroese ASR system (7.6% WER) (Hernandez Mena et al., 2023). Subsequent work introduced grapheme-to-phoneme modeling and expanded linguistic resources for standardized processing (Lamhauge et al., 2023). A recent representation-space analysis of Wav2Vec 2.0 XLSR-53 further examined Faroese in relation to 102 languages and found that Swedish and Norwegian emerge as its closest Scandinavian neighbors in different encoder layers, based on Euclidean distances in the representation space(í Lág et al., 2024).

Despite its close genealogical relationship to Icelandic, Faroese differs substantially in phonology and orthographic conventions. It exhibits rich inflectional morphology, productive compounding, vowel quantity contrasts, and complex consonant clusters (Thráinsson et al., 2012; Eghdam and coauthors incl. user, 2023; Petersen and Voeltzel, 2025). These characteristics increase lexical variability and complicate end-to-end modeling under limited supervision. In addition, extensive historical contact with Danish has resulted in lexical borrowing and code-mixing, particularly in formal domains such as parliamentary speech (Lamhauge et al., 2023; Hernandez Mena et al., 2023).

3. Data

3.1. Data sets

This study draws on a curated set of publicly available speech corpora covering Faroese and four closely related North Germanic languages (Icelandic, Danish, Norwegian, and Swedish). The data are sourced from publicly available collections hosted on the *Huggingface* platform and selected based on recent release and annotation quality.

The corpora are organized according to the two pipelines. Pipeline A uses labeled Faroese speech from *Ravnursson* and mixed-domain labeled data for the four auxiliary languages. Pipeline B uses unlabeled parliamentary speech for all languages during continuous pre-training, including Faroese *FPSC* parliament data í Lág (2025) and *Ravnurs-*

son for subsequent fine-tuning.

Data for *Pipeline A*. Faroese: *Ravnursson* Hernández Mena and Simonsen (2022), (url, 2022b); Icelandic: *Samrómur* O'Brien et al. (2024), (url, 2020); Danish: *CoRal* Nielsen et al. (2024), (url, 2024); Norwegian: *Norwegian Parliamentary Speech Corpus (NPSC)* Solberg and Ortiz (2022), (url, 2022a); Swedish: *RixVox* Rekathati (2023), (url, 2023).

Data for *Pipeline B*. Faroese: *Faroese Parliament Speech Corpus (FPSC)* í Lág (2025), using a 1000-hour subset of parliamentary debates from 2020–2025 í Lág (2025) and *Ravnursson* Hernández Mena and Simonsen (2022); Icelandic: *Althingi Parliamentary Speech* (2005–2016) Helgadóttir et al. (2021); Danish: *FT Speech – Danish Parliament Speech Corpus* (2010–2019) Kirkedal et al. (2020); Norwegian: *NST - Norwegian ASR Database*, parliamentary speech from 2017–2018 spr (2023); Swedish: *RixVox* parliamentary recordings (2003–2023) Rekathati (2023), (url, 2023).

Evaluation data: *Ravnursson* test split and the new *FO-Parl-Eval-2h10m* benchmark All models are evaluated on the official *Ravnursson* test split, corresponding to the dataset used for supervised fine-tuning. In addition, we introduce a new Faroese evaluation benchmark, *FO-Parl-Eval-2h10m* í Lag (2025), comprising 2 hours and 10 minutes of curated parliamentary speech extracted from the *Faroese Parliament Speech Corpus (FPSC)* í Lag (2025). The dataset consists of 344 speech segments ranging from 5 to 29 seconds and reflects real-world parliamentary speech conditions. It is designed to be representative in terms of speaker gender, age groups, and dialectal coverage. As parliamentary speech involves recurring speakers, the evaluation is speaker-dependent, meaning that speakers may appear in both training and evaluation data. This setup reflects realistic deployment conditions but should be interpreted accordingly when assessing generalization. To prevent data leakage at the segment level, all segments included in *FO-Parl-Eval-2h10m* are excluded from any data used in the continuous pretraining stage in Pipeline B.

3.2. Preparation of Data Sets for Experiments

Data preparation is structured around the two experimental pipelines. Data are partitioned and preprocessed differently in each pipeline depending on their role in self-supervised adaptation or supervised training, ensuring stage-appropriate use and comparability across pipelines.

Preparation of Data for Pipeline A. For each auxiliary language (Icelandic, Danish, Norwegian, and Swedish), we first collect all available audio–text data from the respective corpora. From this pool, we randomly sample speech segments until a total duration of 100 hours is reached. This results in a balanced dataset of 100 hours per language.

Formally, let D_l denote the full set of available audio–text pairs for language l, where each sample consists of an audio signal a and its transcription t. We construct a subset $D_l^{100h} \subset D_l$ such that the total duration satisfies:

$$\sum_{(a,t)\in D_l^{100h}} \text{Duration}(a) \approx 100\,\text{h}. \tag{1}$$

The four language-specific subsets are then combined to form a multilingual dataset:

$$D_{\text{multi}} = \bigcup_l D_l^{100h} \tag{2}$$

resulting in a total of 400 hours of speech. Each sample is annotated with its corresponding language identifier. Finally, the combined dataset is split into training, validation, and test partitions for supervised fine-tuning Lág (2025).

Preparation of Data for Pipeline B. From the 1,600-hour *Faroese Parliament Speech Corpus (FPSC)* í Lág (2025), 1000 hours of speech were randomly selected. Non-speech and extremely low-volume segments were removed, and all recordings were normalized in volume. The resulting data were segmented into fixed-length audio chunks of 20 seconds, yielding approximately 180,000 audio files.

For the four auxiliary Scandinavian languages, 500 hours per language were randomly sampled from their respective parliamentary speech corpora and processed using the same pipeline. To ensure consistent cross-lingual exposure during training, the data for Faroese and each auxiliary language were mixed such that consecutive speech segments alternated between Faroese and the auxiliary language. Each language was segmented into 90,000 audio files of 20 seconds. For the combined dataset with all five languages, 200 hours per language were randomly selected from the respective subsets and systematically interleaved to ensure balanced mixing across languages.

4. Methods

4.1. Experiments

An overview of the two adaptation pipelines and their shared components is shown in Figure 1. The pipelines differ in the stage at which cross-lingual information is introduced: Pipeline A incorporates it during supervised fine-tuning, whereas Pipeline B introduces it through an intermediate continuous self-supervised pre-training stage prior to fine-tuning.

Pipeline A consists of two supervised fine-tuning stages—first on the auxiliary language, and subsequently on Faroese. In contrast, Pipeline B inserts a continuous multilingual pre-training stage between the original XLS-R pre-training and the final supervised fine-tuning step. Although XLS-R is initially pretrained on 128 languages, all intermediate adaptation stages in this study operate on controlled language subsets. Specifically, these stages use either a single Scandinavian auxiliary language or a balanced mixture of Scandinavian languages, while the final supervised ASR fine-tuning is performed exclusively on Faroese.

All models are evaluated after a final stage of supervised fine-tuning on the same 100-hour labeled Faroese dataset, ensuring a consistent target-language adaptation across all configurations. Evaluation is performed on the *Ravnursson* test set and the *FO-Parl-Eval-2h10m* benchmark.

For Pipeline A, Wav2Vec 2.0 XLS-R models are evaluated in the 300M, 1B, and 2B parameter variants. Due to computational constraints, Pipeline B is evaluated only using the 1B variant. As architectural baselines, Whisper and Omnilingual ASR models are included for comparison. Whisper is directly fine-tuned on Faroese under the same target-language conditions. Omnilingual models are used solely for inference as external baselines and are not adapted. They are not evaluated on the *Ravnursson* test split due to potential overlap with their training data (Meta AI Research, 2025), and to ensure a fair comparison with models evaluated on disjoint data.

All code, configuration files, and data processing pipelines required to reproduce the experiments are publicly available on GitHub[1]

Pipeline A: Supervised Fine-Tuning–Based Adaptation Pipeline A introduces cross-lingual transfer exclusively through labeled supervision. No additional self-supervised training is performed beyond the original multilingual pretraining.

[1] https://github.com/davidilag/Stage-Aware-Cross-Lingual-Transfer-for-Faroese-ASR-2026

The pipeline consists of two supervised stages per of the six experiments:

1. *FT (Fine-Tuning stage):* The pretrained XLS-R model is fine-tuned on 100 hours of labeled speech from one or more auxiliary Scandinavian languages.

2. *CFT (Continued Fine-Tuning stage):* The model is further fine-tuned on 100 hours of labeled Faroese only.

We compare the results from these experiments with a monolingual baseline, where both FT and CFT stages are performed with Faroese only.

Configurations

We define six experiments (E1–E6). E1 serves as the monolingual baseline. Experiments E2–E5 each incorporate a single Scandinavian auxiliary language during the fine-tuning (FT) phase, while E6 uses a balanced mixture of Scandinavian languages during fine-tuning.

- *FT-E1 (Monolingual baseline):*
  FT(FO $= 100\text{h}$) $\rightarrow$ *CFT*(FO $= 100\text{h}$)
- *FT-E2–E5 (Single auxiliary language):*
  FT($X = 100\text{h},\ X \in \{IS, DK, NO, SW\}$) $\rightarrow$ *CFT*(FO $= 100\text{h}$)
- *FT-E6 (Multilingual mixture):*
  FT(25h per $X,\ X \in \{$IS, DK, NO, SW$\}$) $\rightarrow$ *CFT*(FO $= 100\text{h}$)

All models are made publicly available (í Lág, 2025b,c,a,d,e,f).

Pipeline B: Continuous Self-Supervised Pre–Training–Based Adaptation Pipeline B introduces cross-lingual transfer at the representation level via continuous pre-training before supervised fine-tuning.

The pipeline consists of:

1. *CPT (Continuous Pre-Training stage):* The XLS-R model is further trained using the original Wav2Vec 2.0 self-supervised objective on 1000 hours unlabeled speech. The 1000 hours of unlabeled speech contain one or more auxiliary Scandinavian languages.

2. *FT (Fine-Tuning stage):* The model is then fine-tuned on 100 hours of labeled Faroese.

Configurations

We define six experiments (E1–E6). E1 serves as the monolingual baseline. Experiments E2–E5 each incorporate 500 hours from a single Scandinavian auxiliary language in the CPT phase, accompanied by 500 of Faroese. E6 uses a balanced mixture (200 hours each) of Scandinavian languages, including Faroese, in the CPT phase.

- *CPT-E1 (Monolingual baseline):* *CPT*(FO $= 1000$h) $\rightarrow$ *FT*(FO $= 100$h)
- *CPT-E2–E5 (Balanced bilingual):* *CPT*(FO $=$ 500h $+ X =$ 500h, $X \in$ {IS, DK, NO, SW}) $\rightarrow$ *FT*(FO $= 100$h)
- *CPT-E6 (Multilingual mixture):* *CPT*(200h per X, $X \in$ {FO, IS, DK, NO, SW}) $\rightarrow$ *FT*(FO $= 100$h)

All models are made publicly available (í Lág, 2025g,j,i,k,l,h)

4.2. Whisper and Omnilingual Models

To provide a comparison with state-of-the-art ASR systems, we include Whisper and Omnilingual as reference models. Whisper model variants 39M, 74M, 244M, 1.55B are fine-tuned exclusively on Faroese speech and evaluated on the *Ravnursson* test set. Omnilingual 7B models are already trained on *Ravnursson* and are therefore evaluated in inference-only mode without additional adaptation on *FO-Parl-Eval-2h10m*.

4.3. Formal Definition of Transfer Setups

We formalize the transfer setups using two generic adaptation operators applied to a pretrained multilingual encoder θ_0.

Let D_s^{lab} and D_s^{unlab} denote labeled and unlabeled data from one or more source languages, and D_t^{lab} and D_t^{unlab} labeled and unlabeled data from the target language (Faroese).

Adaptation operators. We define:

$$\mathcal{F}_{\text{ASR}}(\theta; D) = \arg\min_{\theta} \mathcal{L}_{\text{ASR}}(D; \theta), \tag{3}$$

$$\mathcal{P}_{\text{SSL}}(\theta; D) = \arg\min_{\theta} \mathcal{L}_{\text{SSL}}(D; \theta), \tag{4}$$

where $\mathcal{L}_{\text{ASR}}$ denotes the supervised speech recognition training objective and $\mathcal{L}_{\text{SSL}}$ denotes a self-supervised representation learning objective applied to unlabeled speech data.

Transfer setups. Using these operators, the training pipelines are expressed as:

$$\theta_{\text{FT}} = \mathcal{F}_{\text{ASR}}(\theta_0; D_t^{\text{lab}}) \tag{5}$$

$$\theta_{\text{CFT}} = \mathcal{F}_{\text{ASR}}(\theta_0; D_s^{\text{lab}} \cup D_t^{\text{lab}}) \tag{6}$$

$$\theta_{\text{CPT}} = \mathcal{P}_{\text{SSL}}(\theta_0; D_s^{\text{unlab}} \cup D_t^{\text{unlab}}) \tag{7}$$

$$\theta_{\text{CPT}\rightarrow\text{FT}} = \mathcal{F}_{\text{ASR}}\Big(\mathcal{P}_{\text{SSL}}(\theta_0; D_s^{\text{unlab}} \cup D_t^{\text{unlab}}); D_t^{\text{lab}}\Big) \tag{8}$$

Pipeline / Exp.	*Ravnursson* WER	*Ravnursson* CER	*FO-Parl* WER	*FO-Parl* CER
Pipeline A				
FT-E1 (FO)	7.49	2.13	37.78	17.70
FT-E2 (FO-IS)	8.29	2.36	42.00	20.20
FT-E3 (FO-DK)	7.75	2.22	39.39	18.31
FT-E4 (FO-NO)	**6.91**	**2.04**	36.92	**17.16**
FT-E5 (FO-SW)	7.24	2.06	36.68	17.64
FT-E6 (All)	7.38	2.13	**36.61**	17.35
Pipeline B				
CPT-E1 (FO)	**6.59**	**1.85**	**26.46**	**10.68**
CPT-E2 (FO-IS)	6.80	1.95	27.64	10.96
CPT-E3 (FO-DK)	6.96	1.97	26.90	10.84
CPT-E4 (FO-NO)	7.08	2.00	26.68	10.86
CPT-E5 (FO-SW)	7.29	2.06	31.06	14.72
CPT-E6 (All)	7.47	2.11	29.80	12.72

Table 1: Evaluation results for two training pipelines on the *Ravnursson* and *FO-Parl-Eval-2h10m* test sets. Highest performing models are listed in bold text and lowest in red text.

This formulation isolates training stage from data selection, allowing the two pipelines to be interpreted as alternative compositions of the same adaptation operators, differing only in when cross-lingual and target-specific information is introduced.

5. Results

Table 1 summarizes performance under the two pipelines on two Faroese test sets. Performance is consistently lower on *FO-Parl-Eval-2h10m* than on the *Ravnursson*, indicating that the parliamentary benchmark is substantially more challenging and better reflects difficult real-world conditions. Across all language configurations, Pipeline B yields clear gains over Pipeline A on *FO-Parl-Eval-2h10m*, with the largest absolute improvements observed for Faroese-only training.

A second observation is that the optimal auxiliary language depends on the training strategy. In Pipeline B, single-language Scandinavian auxiliaries (notably Icelandic, Danish, and Norwegian) typically outperform the balanced five-language mixture, whereas in Pipeline A, gains are smaller and less consistent, with Norwegian and Swedish providing the strongest improvements among the single-language variants.

As shown in Figure 3, continuous fine-tuning (Pipeline A) improves performance for Faroese with all auxiliary languages except Icelandic. Norwegian, as auxiliary language, yields the strongest improvement for Faroese, followed by Swedish, the Scandinavian mixture, and Danish. In contrast, Icelandic does not provide a measurable benefit over direct fine-tuning.

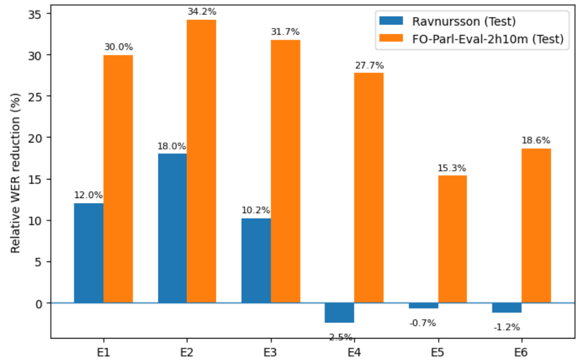


Figure 2: Relative WER reduction (%) of Pipeline B over Pipeline A by experiment (E1–E6) on *Ravnursson* and *FO-Parl-Eval-2h10m* evaluation data sets.

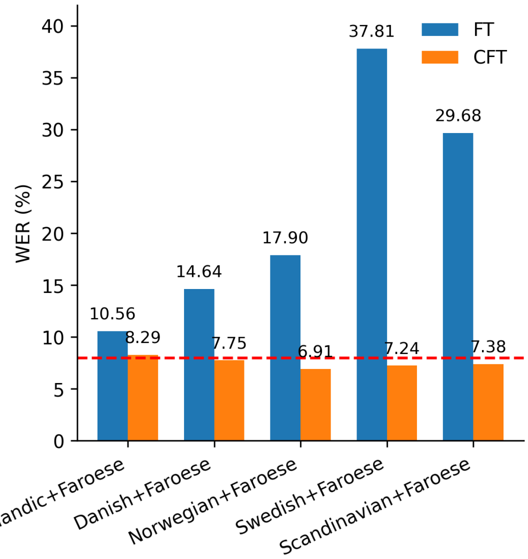


Figure 3: WER after multilingual fine-tuning and continuous fine-tuning (Pipeline A) using different Scandinavian source-language configurations. The dashed red line indicates a reference WER level (Faroese only).

Finally, model capacity and training duration matter for Faroese-only fine-tuning. Table 2 shows that extending training from 30 to 60 epochs improves all XLS-R sizes, with the 1B variant providing the strongest Faroese-only baseline on *Ravnursson*. Table 3 shows clear scaling for Whisper, where larger models substantially reduce error rates and establish a competitive supervised baseline.

Figure 2 summarizes the relative error reduction achieved by Pipeline B over Pipeline A across experiments *E1–E6*. Gains are consistently larger on the parliamentary benchmark, confirming that CPT primarily improves robustness under more challenging conditions, while improvements on the *Ravnursson* test set are smaller and language dependent.

Model Size	30 epochs		60 epochs	
	WER (%)	CER (%)	WER (%)	CER (%)
XLS-R 300M	**7.46**	**2.08**	6.78	1.95
XLS-R 1B	7.49	2.13	**5.85**	**1.73**
XLS-R 2B	7.94	2.28	6.58	1.91

Table 2: Wav2Vec 2.0 XLS-R fine-tuned exclusively on Faroese for 30 and 60 epochs (FT-E1), evaluated on *Ravnursson*.

Model Size	WER (%)	CER (%)
Tiny (39M)	35.92	13.47
Base (74M)	22.86	7.40
Small (244M)	15.03	4.75
Large (1.55B)	**6.51**	**2.11**

Table 3: Whisper fine-tuned exclusively on 100h Faroes, evaluated on *Ravnursson*.

In table 4 we compare our best performing Wav2Vec 2.0 model with two Omnilingual 7B models evaluated on *FO-Parl-Eval-2h10m*. The Omnilingual LLM variant achieves the lowest WER, while the Pipeline B model attains comparable performance with substantially lower CER. In contrast, the Omnilingual CTC variant degrades markedly. These results show that targeted language adaptation in a continual pre-training setting can still provide domain-specific robustness over large, massively multilingual, out of the box models.

Model	Variant	WER (%)	CER (%)
Wav2Vec2	CPT→FT	26.46	10.68
Omnilingual 7B	LLM	23.06	11.71
Omnilingual 7B	CTC	35.99	15.51

Table 4: Performance on *FO-Parl-Eval-2h10m* data set. Wav2Vec 2.0 model is the best perming model in the Pipeline B. Omnilingual models are evaluated in inference-only mode since it already has been fine-tuned on *Ravnursson*.

6. Discussion

Our comparative analysis demonstrates that cross-lingual transfer to Faroese behaves differently across adaptation stages and auxiliary language configurations, with distinct performance patterns observed for supervised fine-tuning and continuous pre-training. Performance does not follow a uniform pattern across languages; rather, the effectiveness of a given auxiliary language depends on whether cross-lingual exposure is introduced during continuous self-supervised pre-training (Pipeline A) or during supervised fine-tuning (Pipeline B).

Within Pipeline A, Norwegian and Swedish consistently provide the strongest gains when used as auxiliary languages, despite not exhibiting the best standalone ASR performance. We speculate the reason for this might be phonetic proximity as primary driver of transfer effectiveness during fine-tuning. Faroese and Norwegian, in particular, share a strong similarity in acoustic–phonetic structure—including consonant realization, vowel quantity, stress patterns, palatalization, and prosody—possibly facilitating parameter adaptation in later supervised fine-tuning stages, where the model specializes to Faroese phonotactics and grapheme-to-phoneme mappings. In contrast, Icelandic, although genealogically closest to Faroese, contributes limited improvement in this pipeline, suggesting that phylogenetic relatedness alone is insufficient to enhance transfer during fine-tuning. Norwegian and Swedish also proved to be the closest to Faroese within the Wav2Vec 2.0 representation space (í Lág et al., 2024), suggesting perceived proximity by the model may facilitate transfer.

In Pipeline B, transfer dynamics differ substantially. Continuous pre-training operates at the representation level and appears less sensitive to fine-grained phonetic similarity. Instead, gains reflect broader structural and lexical alignment. Icelandic and Danish yield improvements during CPT, likely due to morphosyntactic overlap and stylistic similarity in parliamentary speech. The strong effect of Danish on the parliamentary evaluation set supports this interpretation, as formal Faroese discourse contains a high density of Danish loanwords.

Taken together, the two pipelines suggest complementary mechanisms: Pipeline B enhances robustness to real world language usage, while Pipeline A sharpens phonetic and pronunciation-specific modeling. The auxiliary language that is optimal in one pipeline is not necessarily optimal in the other, reinforcing that cross-lingual transfer must be treated as a stage-aware and language-specific optimization problem.

Comparison with fine-tuned Whisper and out-of-the-box Omnilingual yields additional perspective. Whisper, fine-tuned on Faroese data, demonstrates strong word-level robustness due to large-scale supervised multilingual training. On *FO-Parl-Eval-2h10m*, the Omnilingual 7B LLM variant achieves the best result, outperforming our best Pipeline B configuration, while the same Pipeline B model attains lower CER, indicating more stable character-level modeling. These differences suggest complementary error profiles: large multilingual models improve word-level prediction, whereas continuous pre-training remains advantageous for sub-word precision and language-specific orthographic consistency.

7. Conclusions

This work demonstrates that cross-lingual transfer for low-resource Faroese ASR is inherently stage-dependent, and that the timing of cross-lingual information injection fundamentally shapes downstream behavior. Through a unified comparison of two adaptation pipelines, we show that the stage at which auxiliary languages are introduced determines whether transfer learning primarily benefits from phonetic similarity or from syntactic/lexical overlap.

A key contribution of this study is the controlled, side-by-side evaluation of both pipelines under closely matched language data representation, shared model architectures, and consistent evaluation protocols. This isolates the effect of adaptation technique and auxiliary language composition. Pipeline A seemingly consolidates Faroese-specific phonotactics and performs strongly in-domain, whereas Pipeline B consistently improves spelling (lowest CER) and robustness under parliamentary speech conditions. The two pipelines are therefore complementary rather than interchangeable.

Beyond Faroese, this work introduces new insights on stage-aware adaptation for structuring low-resource, cross-lingual ASR. By reframing multilingual transfer as a design choice over training strategies and data composition, the study contributes a methodological lens for analyzing adaptation strategies in low-resource ASR within self-supervised and large-scale multilingual settings.

8. Limitations

This study has several limitations. First, due to computational constraints, we focus on smaller model variants, and continuous pre-training (CPT) is conducted for only one model size. While this limits conclusions about scalability, it reflects realistic deployment conditions in low-resource settings, where computational resources are limited.

Second, our analysis focuses on a single target language, Faroese, with specific sociolinguistic characteristics. Although this provides a focused case study, the generalizability of our findings to other low-resource languages remains to be validated.

Finally, the continuous pre-training stage relies on data from the same domain as one evaluation set (*FO-Parl-Eval-2h10m*). This may advantage CPT-adapted models on that dataset. However, given the severe data constraints typical of low-resource languages, avoiding domain overlap is often impractical, making this limitation representative of real-world conditions rather than an isolated methodological choice.

9. Bibliographical References

2020. Samrómur icelandic speech corpus. https://huggingface.co/datasets/language-and-voice-lab/samromur_asr.

2022. Magic dust for cross-lingual adaptation of monolingual wav2vec-2.0. In *ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

2022a. Norwegian parliamentary speech corpus (npsc) by nbailab. https://huggingface.co/datasets/NbAiLab/NPSC.

2022b. Ravnursson corpus. https://huggingface.co/datasets/carlosdanielhernandezmena/ravnursson_asr.

2023. Rixvox, swedish parliament speech from period 2003-2023. https://huggingface.co/datasets/KBLab/rixvox.

2024. Coral: Danish conversational and read-aloud dataset. https://huggingface.co/datasets/alexandrainst/coral.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick Von Platen, Yatharth Saraf, Juan Pino, et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Alexei Baevski, Alexis Conneau, and Michael Auli. 2020a. Wav2vec 2.0: Learning the structure of speech from raw audio. *Meta AI*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020b. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Akbayan Bekarystankyzy, Orken Mamyrbayev, Mateus Mendes, Anar Fazylzhanova, and Muhammad Assam. 2024. Multilingual end-to-end asr for low-resource turkic languages with common alphabets. *Scientific Reports*, 14(1):13835.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. In *Speech Communication*, volume 56, pages 85–100.

Jaejin Cho, Murali Karthick Baskar, Ruizhi Li, Matthew Wiesner, Sri Harish Mallidi, Nelson Yalta, Martin Karafiat, Shinji Watanabe, and Takaaki Hori. 2018. Multilingual sequence-to-sequence speech recognition: architecture, transfer learning, and language modeling. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 521–527. IEEE.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Iben Eghdam and co-authors incl. user. 2023. Standardising pronunciation for a grapheme-to-phoneme converter for faroese. In *Proceedings of the 24th Annual Conference of the International Speech Communication Association (INTERSPEECH 2023)*, page to appear. Preprint version consulted.

Yutian Geng, Hilary Dempster, Robert Jimerson, Martin Haspelmath, Luke Zettlemoyer, and Shinji Watanabe. 2025. Evaluating speech foundation models for automatic speech recognition in the low-resource kanyen'kéha language. In *Proc. Interspeech*. Despite recent progress in SFMs, low-resource Indigenous ASR remains challenging due to limited annotated data and extensive vocabulary variation.

Yaroslav Getman, Tamás Grósz, Katri Hiovain-Asikainen, and Mikko Kurimo. 2024. Exploring adaptation techniques of large speech foundation models for low-resource asr: a case study on northern sámi. *Interspeech 2024*.

Vishwa Gupta and Gilles Boulianne. 2022. Progress in multilingual speech recognition for low resource languages kurmanji kurdish, cree and inuktut. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6420–6428.

Carlos Hernandez Mena, Annika Simonsen, and Jon Gudnason. 2023. Asr language resources for faroese. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 32–41.

Dávid í Lág, Barbara Scalvini, and Jon Gudnason. 2024. Mapping faroese in the multilingual representation space: Insights for ASR model optimization. In *The Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies*.

Sandra Lamhauge, Iben Debess, Carlos Hernández Mena, Annika Simonsen, and Jon Gudnason. 2023. Standardising pronunciation for a grapheme-to-phoneme converter for Faroese. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 308–317, Tórshavn, Faroe Islands. University of Tartu Library.

Meta AI Research. 2025. Omnilingual ASR: lang_ids.py. https://github.com/facebookresearch/omnilingual-asr/blob/main/src/omnilingual_asr/models/wav2vec2_llama/lang_ids.py. GitHub repository. Accessed: 2026-02-16.

Sonja Müller, Daniel Fuchs, and Laurette Pretorius. 2024. Exploring asr fine-tuning on limited domain-specific data for low-resource languages. In *Southern African Linguistics and Applied Language Studies*. Shows that modern ASR trained on government/political data performs poorly on out-of-domain broadcast news, underscoring domain mismatch under low-resource conditions.

Hjalmar P. Petersen and Laurence Voeltzel. 2025. *Faroese Phonetics and Phonology*, volume 34 of *Phonology and Phonetics*. De Gruyter Mouton, Berlin.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022a. Robust speech recognition via large-scale weak supervision. https://arxiv.org/abs/2212.04356.

Alec Radford et al. 2022b. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*. Introduces Whisper, a multilingual model trained on 680k hours of weakly supervised data, showing strong zero-/few-shot transfer to many low-resource languages.

Annika Simonsen, Sandra Saxov Lamhauge, Iben Nyholm Debess, and Peter Juel Henrichsen. 2022. Creating a basic language resource kit for faroese. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4637–4643.

Omnilingual ASR team, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, Kehan Lyu, Sagar Miglani, Vineel Pratap, Kaushik Ram Sadagopan, Safiyyah Saleem, Arina Turkatenko, Albert Ventayol-Boada, Zheng-Xin Yong, Yu-An Chung, Jean Maillard, Rashel Moritz, Alexandre Mourachko, Mary Williamson, and Shireen Yates. 2025. Omnilingual asr: Open-source multilingual speech recognition for 1600+ languages.

Samuel Thomas, Michael L. Seltzer, Kenneth Church, and Hynek Hermansky. 2012. Deep neural network features and semi-supervised training for low-resource speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6704–6708.

Höskuldur Thráinsson, Hjalmar P. Petersen, Jógvan í Lon Jacobsen, and Zakaris Svabo Hansen. 2012. *Faroese: An Overview and Reference Grammar*, 2 edition. Fróðskapur, Tórshavn.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

A. Williams, A. Demarco, and C. Borg. 2023. The applicability of wav2vec2 and whisper for low-resource maltese asr. In *Proceedings of SIGUL 2023*.

Hemant Yadav and Sunayana Sitaram. 2022. A survey of multilingual models for automatic speech recognition. *Language Resources and Evaluation*. Surveys multilingual and cross-lingual ASR, emphasizing lack of labeled data for most languages and the role of transfer/self-supervised learning in low-resource settings.

C. Yi, J. Wang, C. Ning, S. Zhou, and B. Xu. 2021. Transfer ability of monolingual wav2vec2.0 for low-resource speech recognition. In *Proceedings of the International Joint Conference on Neural Network*, pages 1–6.

Dávid í Lág. 2025a. Wav2vec2 xls-r 1b: Fine-tuning on 100h danish followed by continuous fine-tuning on 100h faroese (ft→cft, e3). https://huggingface.co/davidilag/wav2vec2-xls-r-1b-E3-faroese-100h-30-epochs_20250124_v2.

Dávid í Lág. 2025b. Wav2vec2 xls-r 1b: Fine-tuning on 100h faroese (ft, e1). https://huggingface.co/davidilag/wav2vec2-xls-r-1b-faroese-100h-60-epochs_20250108_v2.

Dávid í Lág. 2025c. Wav2vec2 xls-r 1b: Fine-tuning on 100h icelandic followed by continuous fine-tuning on 100h faroes (ft→cft, e2). https://huggingface.co/davidilag/wav2vec2-xls-r-1b-E2-faroese-100h-30-epochs_20250124_v3.

Dávid í Lág. 2025d. Wav2vec2 xls-r 1b: Fine-tuning on 100h norwegian followed by continuous fine-tuning on 100h faroese (ft→cft, e4). https://huggingface.co/davidilag/wav2vec2-xls-r-1b-E4-faroese-100h-30-epochs_20250208_v4.

Dávid í Lág. 2025e. Wav2vec2 xls-r 1b: Fine-tuning on 100h swedish followed by continuous fine-tuning on 100h faroese (ft→cft, e5). https://huggingface.co/davidilag/wav2vec2-xls-r-1b-E5-faroese-100h-30-epochs_20250124.

Dávid í Lág. 2025f. Wav2vec2 xls-r 1b: Fine-tuning on 25h each of danish, icelandic, norwegian, and swedish followed by continuous fine-tuning on 100h faroese (ft→cft, e6). https://huggingface.co/davidilag/wav2vec2-xls-r-1b-E6-faroese-100h-30-epochs_20250209.

Dávid í Lág. 2025g. Wav2vec2 xls-r 300m: Continuous pre-training on 1000h faroese followed by fine-tuning on 100h faroese (cpt→ft, e1). https://huggingface.co/davidilag/wav2vec2-xls-r-300m-cpt-1000h_faroese-cp_best-faroese-100h-60-epochs_run8_2025-09-24.

Dávid í Lág. 2025h. Wav2vec2 xls-r 300m: Continuous pre-training on 200h each of faroese, danish, icelandic, norwegian, and swedish followed by fine-tuning on 100h faroese (cpt→ft, e6). https://huggingface.co/davidilag/wav2vec2-xls-r-300m-cpt-200h-FO-IS-NO-DK-SE-cp-best-faroese-100h-30-epochs_run9_2025-09-11.

Dávid í Lág. 2025i. Wav2vec2 xls-r 300m: Continuous pre-training on 500h faroese and 500h danish followed by fine-tuning on 100h faroese (cpt→ft, e3). https://huggingface.co/davidilag/wav2vec2-xls-r-300m-pt-500h-FO-500h-DK-cp-best-ft-faroese-100h-30-epochs_run9_2025-09-11.

Dávid í Lág. 2025j. Wav2vec2 xls-r 300m: Continuous pre-training on 500h faroese and 500h icelandic followed by fine-tuning on 100h faroese (cpt→ft, e2). https://huggingface.co/davidilag/wav2vec2-xls-r-300m-pt-500h-FO-500h-IS-cp-best-faroese-100h-30-epochs_run9_2025-09-10.

Dávid í Lág. 2025k. Wav2vec2 xls-r 300m: Continuous pre-training on 500h faroese and 500h norwegian followed by fine-tuning on 100h faroese (cpt→ft, e4). https://huggingface.co/davidilag/wav2vec2-xls-r-300m-pt-500h-FO-500h-NO-cp-best-faroese-100h-30-epochs_run9_2025-09-11.

Dávid í Lág. 2025l. Wav2vec2 xls-r 300m: Continuous pre-training on 500h faroese and 500h swedish followed by fine-tuning on 100h faroese (cpt→ft, e5). https://huggingface.co/davidilag/wav2vec2-xls-r-300m-pt-500h-FO-500h-SE-cp-best-faroese-100h-30-epochs_run9_2025-09-11.

10. Language Resource References

2023. *NST Norwegian ASR Database (16 kHz) – Reorganized*. Originally developed by Nordisk Språkteknologi; reorganized by the National Library of Norway (Språkbanken). Last updated 2023-12-19.

Rosana Ardila and Megan Branson and Kelly Davis and Michael Henretty and Michael Kohler and Josh Meyer and Reuben Morais and Lindsay Saunders and Francis M. Tyers and Gregor Weber. 2020. *Common Voice: A Massively-Multilingual Speech Corpus*.

Helgadóttir, Inga Rún and Kjaran, Róbert and Nikulásdóttir, Anna Björk and Gudnason, Jón. 2021. *Althingi Parliamentary Speech*. Reykjavík University. LDC Catalog No. LDC2021S01.

Hernández Mena, Carlos Daniel and Simonsen, Annika. 2022. *Ravnursson Faroese Speech and Transcripts*.

Kirkedal, Andreas and Stepanović, Marija and Plank, Barbara. 2020. *FT Speech: Danish Parliament Speech Corpus*.

Dávid í Lág. 2025. *Scandinavian ASR Dataset: 100 Hours per Language (Danish, Icelandic, Norwegian, Swedish)*. Hugging Face. Public dataset for multilingual ASR research.

Nielsen, Dan Saattrup and Lehmann, Sif Bernstorff and Madsen, Simon Leminen and Pedersen, Anders Jess and van Zee, Anna Katrine and Blach, Torben. 2024. *CoRal: A Diverse Danish ASR Dataset Covering Dialects, Accents, Genders, and Age Groups*.

O'Brien, Luke and Gunnarsson, Thorsteinn Dadi and Magnúsdóttir, Eydís Huld and Gudnason, Jon. 2024. *Samrómur L2 24.10*. Reykjavík University.

Rekathati, Faton. 2023. *The KBLab Blog: RixVox: A Swedish Speech Corpus with 5500 Hours of Speech from Parliamentary Debates*.

Solberg, Per Erik and Ortiz, Pablo. 2022. *The Norwegian Parliamentary Speech Corpus*.

í Lag, Dávid. 2025. *FO-Parl-Eval-2h10m: Faroese Parliamentary Speech Evaluation Dataset*. University of the Faroe Islands. 2h10m manually transcribed Faroese parliamentary speech evaluation dataset for ASR benchmarking.

í Lág, Dávid. 2025. *FPSC: Faroese Parliament Speech Corpus*. University of the Faroe Islands. Public dataset of Faroese parliamentary speech with metadata and weakly supervised transcripts, 1,600 hours.

Doing More with Less: Determining Optimal Pre-training Model for Irish Automatic Speech Recognition through Multi-step Fine-tuning

Caoilfhionn Ní Dheoráin, Ruth Holmes, Nicholas Evans, Thomas Laurent, Anthony Ventresque, Ellen Rushe
School of Computer Science and Statistics - Trinity College Dublin, SFI Lero,
School of Computing - Dublin City University
{nidheorc, ruth.holmes, nevans, tlaurent, anthony.ventresque}@tcd.ie, ellen.rushe@dcu.ie

Abstract

In recent years, there has been an upsurge in research on automatic speech recognition (ASR) for low-resource languages. Particularly, transfer learning using multi-lingual models has become a popular remedy for the lack of available datasets for target languages. However, given the complexities associated with each individual language, we argue it is unlikely that a single multi-lingual pre-training model will provide equal performance gains across all languages. We also recognise the important, and insufficiently studied influence that the specific pre-training dataset has on the performance of the model. In this paper, using the Irish language as a case study, we propose a more directed, incremental form of pre-training which we term *multi-step fine-tuning*. This method accounts for the complex relationships between the language and dataset features of the source pre-training and target datasets. We show multi-step fine-tuning improves performance over simple multi-lingual fine-tuning alone, and we investigate factors leading to certain pre-trained models achieving better results through linguistic and dataset similarity measures. This research also investigates the uniformity of the performance gains across different demographics. We show that the optimal pre-training strategy can differ between demographics suggesting that more careful pre-training dataset selection is necessary to ensure equitable outcomes in practice.

Keywords: Automatic speech recognition, low-resource language, language similarity, Irish

1. Introduction

Although around 1.9 million people in Ireland report some level of proficiency in the Irish language, its everyday use remains limited. According to the most recent census in 2022 (Central Statistics Office, 2023), only about 72,000 people speak the language on a daily basis, despite recent increases in the overall number of speakers. Though over 39% of the population indicated some ability to speak Irish, English remains the dominant language used in most social and professional settings. Consequently, technological advancements – such as those in Automatic Speech Recognition (ASR) – have been sparce, with the exception of the seminal work of the Abair project (Chasaide et al., 2017).

However, as large public datasets at the scale of higher resource languages are not currently available, developments in ASR for Irish, as with many lower resource languages, remains challenging. This lack of fundamental technologies, specifically language technologies, has created a digital divide between Irish and English resources while also contributing to the potential for *digital extinction* of Irish. The results of this divide is speakers of the language reverting to using English, with only 1.5% of the population using the language outside of the education system (Lynn, 2023). This is in line with previous research classifying the Irish language as "definitely endangered" (Chiaráin et al., 2022).

Enormous strides have been made in ASR over the last decade with the advent of large, parallelisable deep learning models such as Whisper (Radford et al., 2023) and Wav2vec (Schneider et al., 2019), but much of these advancements have centred on ASR for higher resource languages such as English. This gap is well recognised (Li et al., 2022), leading to a variety of approaches to address the deficits of ASR for low-resource languages. For example, transfer learning is a commonly proposed approach to train models in scenarios where high quality transcribed data is unavailable, a common challenge in the field of ASR.

In this paper, we propose a more guided approach to transfer learning in which a pre-trained multi-lingual ASR model is first fine-tuned on one language, before being further fine-tuned on the target language – in this case Irish. We propose a language similarity-based approach to the selection of data used in transfer learning, and a more careful, multi-faceted analysis of downstream effects the data we use has on performance. Our objective is to disentangle the dataset characteristics that lead to an increase in performance from the aspects of learned representations that are still unknown. With the available open source data for Irish as a case study, we:

1. Demonstrate how language and dataset selection affect transfer learning performance.
2. Disentangle the characteristics of a model's

Proceedings of SPEAKABLE @ LREC 2026, pages 162–173
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

training dataset (e.g., dataset size, similarity with the target language) on the efficacy of transfer learning.
3. Determine whether these characteristics can be used to find the optimal pre-training dataset.
4. Measure the effects of different multi-step fine-tuned models on the performance across a variety of demographics and dialects of Irish.

The remainder of the paper is structured as follows: Section 2 provides background and discusses related work on transfer learning and representation learning for ASR and low-resource ASR. Section 3 details the proposed transfer learning method and the linguistic and dataset characteristics we focus on. Section 4 states the research questions that were investigated and describes the experimental setup used to evaluate them. Section 5 presents the results along with a detailed analysis of performance disparities across demographic lines. Section 6 concludes the paper.

2. Background & Related Work

A foundational concept of transfer learning for low-resource learning is the notion of learning generalised representations that transfer across languages. We detail the approaches to and challenges surrounding learning these representations (Section 2.1), the potential role of linguistic similarity to guide this process (Section 2.2), and the role of users in determining whether this transfer of representations might be successful (Section 2.3).

2.1. Building Generalisable Representations

There have been several attempts to learn invariant speech representations for ASR. These are loosely considered to be latent features that are universal to all languages or common across several. One such approach is to create robust or "generalised" speech representations which can be transferred to several speech tasks. Kawakami et al. (2020) aimed to learn robust multilingual representations of speech which transferred to phonologically diverse languages and, indeed, found that such developments led to a marked improvement in WER for several low-resource languages. One issue with approaches such as these, however, is that these "invariant" features are intrinsically opaque and scarcely explored in any detail. Though transferable representations are likely instrumental to accurate low resource speech recognition, the true universality of these features is rarely evaluated beyond a single score for a collection of low-resource languages. To this end, a more focused evaluation of pre-trained bilingual and trilingual models was performed by Lehečka et al. (2024). In contrast to Kawakami et al. (2020), the authors found that monolingual models outperformed bilingual and trilingual Wav2Vec models on oral history archives (even when controlling for dataset size). A similar result was found by Babu et al. (2022) where monolingual English models outperformed a similarly sized multilingual XLS-R model, a phenomenon the authors termed *capacity dilution*. The XLS-R model only reduced the WER upon significantly increasing the number of parameters.

Given the mixed results obtained using multilingual models, it remains challenging for practitioners to reliably build effective ASR models using them. Though it is domain invariant features that are likely to be learned using these models, very little attempt is made to test whether *definable* invariant features are being learned. It is therefore challenging to determine what is actually being transferred from a source model to a target, and whether these features are even desirable to transfer.

2.2. Linguistic Similarity

Linguistic similarity is explored less in the literature with more focus on the use of large uncurated datasets. However, the notion of capacity dilution described by Babu et al. (2022) suggests that, with a fixed capacity architecture, some language-specific features are likely to be lost in favour of models with increased language coverage. Additionally, large datasets are often very imbalanced with a disproportionate percentage of examples being from higher resourced languages (Ardila et al., 2020). To address these issues, work has been done to determine similarities between source and target languages which centers around the use of language identification or phonetic classifiers.

Early work on this topic by Zhang et al. (2014) used the posterior scores over source languages for target-language utterances to determine similarity. The bottleneck features of the source language with the highest posterior score were then used for pre-training. This work found that the closest languages appeared to provide the most benefit to training, stating that there was *"no data like similar data"*. Thomas et al. (2016) took a similar approach by discriminating between phonemes of all source languages, combining the resultant phoneme scores for each language into a global language score. These scores were then used to create a language similarity matrix on which spectral clustering was applied to create language clusters. Multilingual feature frontends were then trained using the language groups identified within each cluster. Qian and Zhou (2022) extended this idea by extracting hidden representations of a language identification model and obtaining the similarity of a given utterance's embedding to the average embedding of the target language. The similarity was then used to weight the loss from a given utterance in order to favour some utterances within a multilingual dataset over others depending on their similarity during training. This technique also led to improvements over standard multilingual pre-training. Li et al. (2019) also saw improvements when corpus-level embeddings were used to select related corpora for training, a method that outperformed fine-tuned multilingual models. Though seemingly effective, techniques using language identification or those based on corpus-level embeddings require training of a discriminative model to obtain similarity scores and rely on the efficacy of the model. They also lack interpretability. They do not provide information as to *why* two languages have been found to be similar given that the similarity is entirely data-driven and learned features cannot be definitively interpreted.

Datasets using the same language with dialectal differences have also been explored. Yi et al. (2020) evaluated the efficacy of supervised pre-training using the Libri 100h (Panayotov et al., 2015) (English dataset) against supervised pre-training using HKUST (Liu et al., 2006) (Mandarin dataset) on the target test set, CALLHOME-MA (also a Mandarin dataset). The authors found that pre-training on HKUST, a "target-similar dataset" Yi et al. (2020) was more effective. They also found that using an abundance of multilingual data for self-supervised pre-training (Libri 1000h (Panayotov et al., 2015)) did lead to superior performance when evaluated on the same CALLHOME-MA dataset over the supervised model trained on HKUST. However this analysis did not further explore the potential reasons for the improved performance or detail the characteristics of the dataset. We are taking a more interpretable approach to finding dataset similarity, aiming to develop a method that identifies dataset characteristics needed to enhance performance.

2.3. Building Models that Serve People

A factor that is also rarely considered in the work described above is the use cases or demographic coverage of ASR systems. There are however a few exceptions, for example, Kawakami et al. (2020) specifically considered the use case of speakers in their model development and evaluation, Jimerson et al. (2023) investigated which model architectures work best for various low-resource languages, and Mitra et al. (2016); Le Ferrand et al. (2020); Littell et al. (2016) among others have dedicated work to specific endangered languages to overcome challenges they face. However, in contrast, a majority of approaches are not designed for specific languages and assume that the use of large, uncurated, datasets will ensure coverage across demographics, this still remains to be seen. For instance, Markl (2022) found that even English models, targeting one of the most well-resourced languages in ASR, perform worse for marginalised communities, including those who speak English as a second language or use specific dialects. Koenecke et al. (2020) also found that common ASR services demonstrated higher error rates for individuals using AAVE. Reitmaier et al. (2022) warned that low-resource ASR is in danger of being treated as an *"intellectual challenge"* with training datasets neither being collected with sufficient input from the communities that speak the languages, nor curated to adequately serve them. This makes interpretable and multifaceted model evaluation important, as understanding the shortcomings of existing benchmarks and models is instrumental in developing more equitable technology.

3. Methodology

This section explains the process of multi-step fine-tuning (Section 3.1), then defines the dataset-dependent (Section 3.2) and independent (Section 3.3) metrics used to determine similarity.

3.1. Multi-step fine-tuning

When performing transfer learning, a large-scale multilingual or mono-lingual dataset is first used to train an ASR model. For language-specific ASR models, models are either trained "from scratch" or, more commonly, fine-tuned from a multilingual model. Our objective is to determine whether certain language-specific models can lead to increased performance when further fine-tuned on a target language. That is, where a language-specific model has been built by fine-tuning a multilingual model on a single language dataset, we seek to fine-tune the model a second time on the target language. For the purposes of this work, we will term this strategy as *multi-step fine-tuning* and we will refer to the languages used in the first stage of fine-tuning as *source fine-tuning languages*. The language used during the final stage of fine-tuning is referred to as the *target language*. We seek to understand whether this strategy is more effective than fine-tuning a multilingual model on Irish alone, whether an increase in performance is uniform across different source fine-tuning languages, and whether the same multi-step fine-tuned model leads to the same performance gains across dialects and demographics.

3.2. Dataset Size

Dataset size is consistently associated with improved performance for deep learning architectures (Lehečka et al., 2024; Kawakami et al., 2020; Yi et al., 2020; Yusuyin et al., 2025). We evaluate if the size of the source fine-tuning dataset is associated with the performance on the target language. While ASR datasets are typically described in terms of speech duration, we use number of samples since Common Voice clips have relatively consistent lengths.

3.3. Language Similarity/Proximity

We hypothesis that dataset-independent language features between the source fine-tuning languages and the target language also influence the model performance. We used the following two methods to calculate this similarity:

3.3.1. Genetic Proximity

Genetic Proximity is computed using a "Genetic Proximity Calculator" tool provided by eLinguistics.net (Elinguistics, 2020) and is independent of the datasets used to train models. This tool calculates the genetic proximity between two languages based on a cognate score (a metric to quantify the similarity between words across different languages). It is derived by comparing the consonants of 18 words that are commonly used in comparative linguistics studies. Consonants specifically are compared as they tend not to evolve as quickly as vowels (Vincent and Johannes, 2020). This tool outputs a score between 0 and 100 calculated using the cognate scoring and statistical context, indicating if the languages are similar or unrelated.

3.3.2. Averaged Lang2vec Similarity

Lang2vec is a Python tool used to query the URIEL database that represents languages using typological, phylogenetic, and geographical features such as genetic, geographic, syntactic, inventory, phonological, and featural vectors (Littell et al., 2017). It provides pre-calculated cosine distances based on these features that can be used to measure language similarity, meaning this metric is also independent of the datasets used to train models. However, analysis performed by Toossi et al. (2024) on the reproducibility of URIEL's language distances indicated that "31.24% of the languages in URIEL have no linguistic feature information", yet it still provides the distances for these languages. This problem inevitably impacts the reliability of this measure for low-resource languages which tend to have more missing values. To address these missing feature vector values, we propose the following method of calculating cosine similarity between the source fine-tuning language and the target language instead of using the pre-calculated cosine distances:

- Find all complete overlapping feature sets between the target and source fine-tuning languages.
- Calculate the cosine similarity for each of these complete feature sets.
- The averaged lang2vec similarity score is the mean cosine similarity for each feature set.

This strategy of similarity score averaging means that all feature sets are given equal weighting and the scores are not influenced by the size of the sets.

4. Experimental Setup

This section includes the specific models and dataset used to carry out the experiments. We choose Irish to the be target language and Dutch, French, German, Persian, Portuguese and English to be the source fine-tuning languages. The research questions we explore are:

RQ.1 Is multi-step fine-tuning more effective than directly fine-tuning a multilingual model?

RQ.2 Is the performance increase provided by multi-step fine-tuning uniform across source fine-tuning languages?

RQ.3 Is higher proximity between the source fine-tuning language and the target language associated with better model performance?

RQ.4 Is the size of the source fine-tuning dataset associated with better model performance?

RQ.5 Is multi-step fine-tuned model performances uniform across dialects and demographics?

4.1. Data

With Irish as the target language for our experiments, we used the open source Common Voice dataset version 15.0 (Mozilla Corporation, 2021) for Irish fine-tuning with its training (536 clips), validation (516 clips), and testing[1] (517 clips) splits. Special characters were removed, words were converted to lowercase and a vocabulary was built using tokenisation. Audio was sampled at 16kHz and zero-padded to match the longest segment.

4.2. Model

The base model in all cases was the Wav2Vec2 cross-lingual speech representation large-scale model trained on 53 languages (XLSR-53) (Conneau et al., 2020). The "feature encoder" of each base model was frozen in the initial fine-tuning stage. Then, this was done again in our experiments using the Irish data. This means that the transformer encoder or "context network" is

[1] Conneau et al. (2020) states all Common Voice languages were used during unsupervised pre-training of XLSR-53 however it is unclear which training/test splits were used for training in the original paper. Despite this, the Irish portion of the dataset has since expanded compared to what was available at the time of this pre-training.

fine-tuned with CTC and the lower level "feature encoder" is frozen to preserve speech features learned from pre-training. Each model configuration was fine-tuned in three independent runs, and the reported results represent the average performance across these runs.

4.2.1. Baseline Single-Step Fine-Tuning

The baseline for experiments was single-step fine-tuning using Irish. Hyperparameters were selected using Bayesian hyperparameter optimisation (Yang and Shami, 2020) with Word Error Rate (WER) as the objective. The best configuration was achieved using a learning rate of 0.0003, a batch size of 8, and 45 training epochs. We note that the objective here is not to create the optimal model for Irish ASR but to find reasonable hyperparameters that can be used uniformly across models in order to compare them.

4.2.2. Multi-Step Fine-Tuning

For our source fine-tuned models, we use XLSR-53 models that are openly available and already fine-tuned for our choice of source fine-tuning languages provided by Grosman (2021). These models had already completed the first stage in the multi-step fine-tuning method using the train and validation splits of Common Voice 6.1 (along with additional audio clips from CSS10 (Park and Mulc, 2019) for Dutch only). To create the multi-step fine-tuned models, they are fine-tuned again on Irish using the same pre-processing steps, data, and hyperparameters as the baseline described in Section 4.2.1.

4.3. Averaged Lang2vec Similarity

As discussed in Section 3.3.2, some source fine-tuning languages chosen for our experiments didn't contain the same complete feature sets as Irish, the target language. French, German, and Persian contained complete sets of the same features. English, Portuguese, and Dutch were missing values that were present for Irish, therefore these feature sets were not included when calculating the averaged lang2vec similarity for those respective languages.

4.4. Demographic and Dialects

We sought to determine whether models performed the same for different speaker demographics and dialects. To do this, we filtered the unseen Irish test dataset to isolate data from each demographic group annotated within the Common Voice dataset (e.g, females, males, teens, 20s, etc.).

Table 1: WER(%) of the different Models

Model	WER	# training samples
English	56.7	580501
French	57.6	314745
Portuguese	59.0	11106
Persian	61.4	12806
Dutch	61.4	14398
Irish (baseline)	66.2	536
German	94.9	262113

5. Results

Section 5.1 details the results of multi-step fine-tuning by comparing the performance of the different models. The results concerning the effect of language and dataset characteristics are split across three sections. First, Section 5.2 analyses the degree to which performance is associated with dataset training size. Second, Section 5.3 focuses on the relation between performance and the dataset-independent language similarity metrics. Finally, Section 5.4 provides an analysis of the performance difference along demographic lines.

5.1. Multi-step Fine-tuning Comparison

Table 1 shows the models' WER performance and source fine-tuning dataset size. Interestingly, we see the benefit provided by multi-step fine-tuning is dataset-specific. In most cases there is an improvement, with the exception of German, where we see a large decrease in performance with a relative percentage increase in WER of $\sim 43\%$. English provides the most improvement with a relative percentage decrease in WER of $\sim 14.4\%$, followed by French and Portuguese with a decrease of $\sim 13\%$ and $\sim 10.9\%$ respectively. This motivates our conclusion to **RQ.1** and **RQ.2**, that multi-step fine-tuning provides a benefit and that the specific source fine-tuning language used impacts performance.

5.2. Effect of Dataset Size

Figure 1 plots the number of training samples from each source fine-tuning dataset against their models' WER performance. This shows that the number of training samples used in the first step of fine-tuning is not clearly associated with increased performance in WER. A Pearson correlation coefficient of 0.008 between WER and the size of the datasets indicates no strong linear relationship. While the English model is fine-tuned on the largest dataset and achieves the lowest WER rate, the French model is trained on a dataset almost half

Table 2: Ranked Similarity Scores of the Different Datasets in Comparison to the Irish Dataset

(a) Genetic proximity between languages

Dataset	Genetic Proximity
French	57.7
Portuguese	59.7
Persian	60.6
German	76.5
English	78.5
Dutch	80.8

(b) Averaged Lang2vec similarity

Dataset	Averaged lang2vec
English	0.6842
Portuguese	0.6558
Dutch	0.6076
Persian	0.5938
German	0.5923
French	0.5922

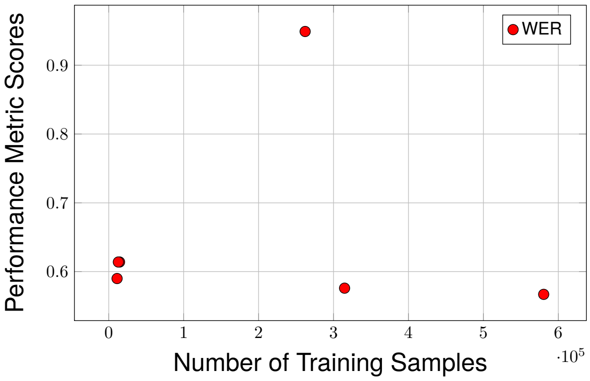


Figure 1: Relationship between the Number of Training Samples and Performance Metrics

the size of English, but there is only a relative percentage difference of $\sim 1.6\%$ in WER. Looking at the model first fine-tuned on Portuguese, it uses a dataset $\sim 1.9\%$ the size of English, and there is only a relative percentage increase of $\sim 4\%$ in WER. The German dataset is the third largest dataset and it achieves the worst performance. This appears to show that the number of examples alone cannot explain the increase in performance against the baseline, answering **RQ.4**, and that the dataset size *in conjunction with* other factors should be considered.

5.3. Language Similarity/Proximity Comparison

When using more generalised similarity metrics such as averaged Lang2vec and Genetic Proximity, we observe some strong correlations between performance and similarity. Specifically, Genetic Proximity of languages in Table 2a has a positive Pearson correlation co- efficient of 0.347. This could indicate phonetic similarities between languages plays an important role in boosting performance, but analysis across additional languages is necessary to make this claim more robust. Interestingly, the correlation between the averaged Lang2vec similarity metric shown in Table 2b and the performance of the models is negatively associated with a correlation of -0.4253. This could be due to features irrelevant to speech being considered. This motivates future further analysis on feature selection.

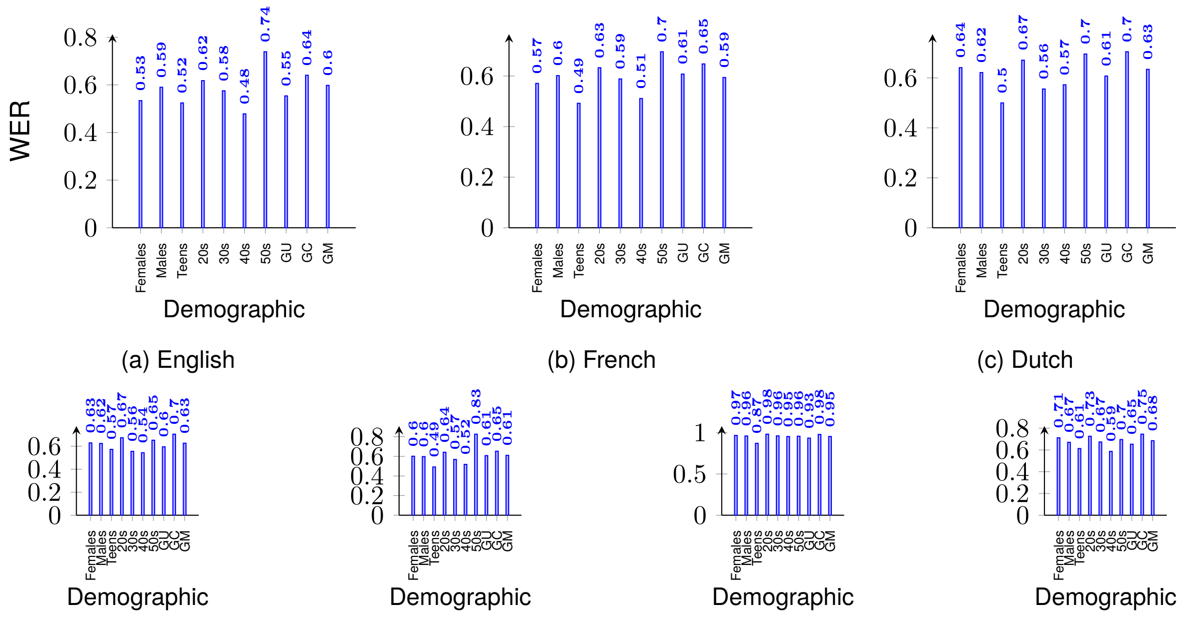
5.4. Performance Variation across Demographic Lines

While WER offers a general performance overview, we observed that it varied across demographics. We compared the WER by gender (Section 5.4.1), age (Section 5.4.2), and dialect (Section 5.4.3).

5.4.1. Gender

The annotated gender categories in the unseen Irish test data of common voice were: `Female`, `Male` and `Other`. When all the models were evaluated on this dataset, the models first fine-tuned on Portuguese (Figure 2e), German (Figure 2f), Persian (Figure 2d) and Dutch (Figure 2c) as well as the baseline model only fine-tuned on Irish (Figure 2g), performed worse for the `Female` category compared to `Male`. While, on face value, this performance disparity seems due to the lack of `Female` representation in datasets, it should be noted that the annotations are incomplete. There were many utterances in all datasets without gender labels. Of all gender labels in the metadata of the training and validation splits, only 11% were labelled `Female` in the Irish dataset, 10.7% in the German dataset, 9.4% in the Dutch dataset, 19.8% in the Persian dataset and as little as 5.8% in the Portuguese dataset. In all cases the `Other` gender category was the most under-represented. With no `Other` gender category labels in the Irish test dataset, we were unable to see the effect this under-representation had on the performance.

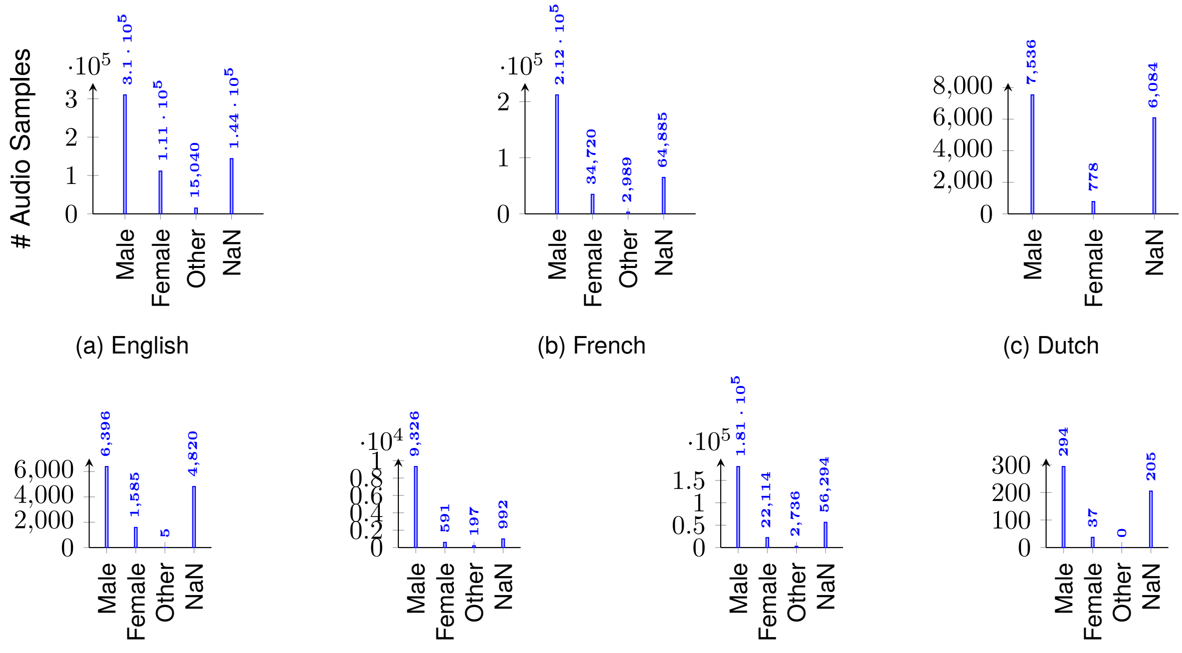
Interestingly, for models with English (Figure 2a) and French (Figure 2b) as the source fine-tuning language, audio labelled `Female` is in general more accurately detected than those labelled `Male`. Of the existing gender annotations, the French dataset has only 13.9% `Female` samples (Figure 3b) and the English dataset has only 25.5%(Figure 3a).



(a) English (b) French (c) Dutch

(d) Persian (e) Portuguese (f) German (g) Irish

Figure 2: WER of each Model by Demographic



(a) English (b) French (c) Dutch

(d) Persian (e) Portuguese (f) German (g) Irish

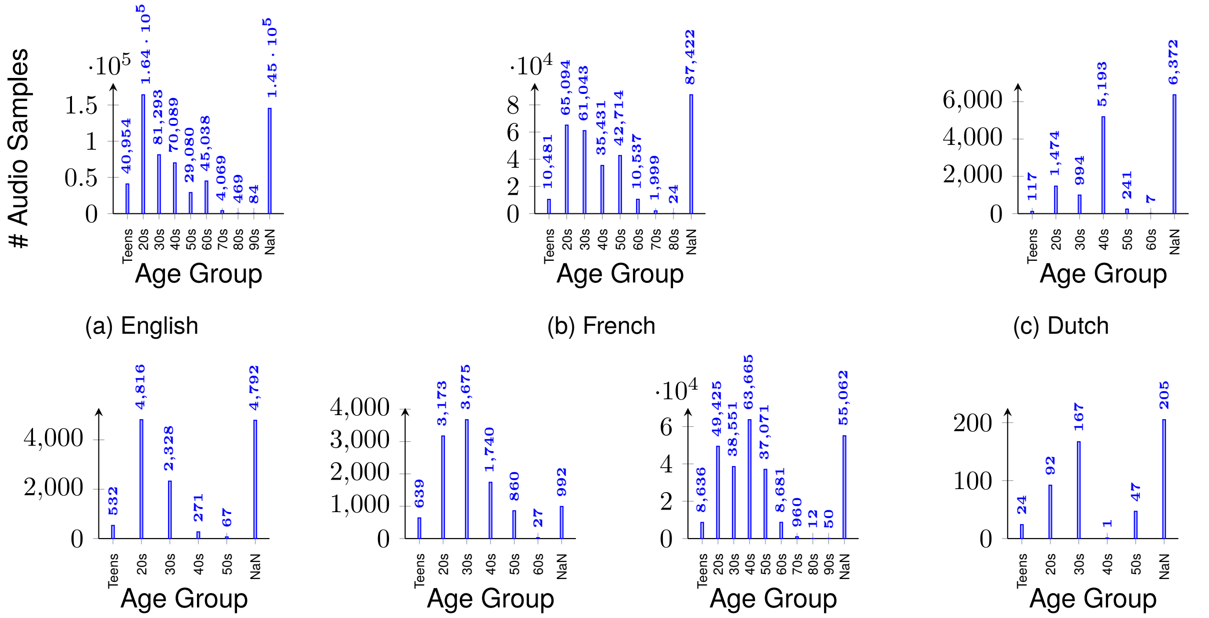
Figure 3: Gender Distribution of Training Datasets

Due to the missing gender annotations, it is challenging to deduce whether performance disparity between genders is due to dataset imbalance, however it is clear that such a disparity exists.

5.4.2. Age

In Figure 2 we observed a notable variation in performance across age-groups. This aligns with findings by Werner et al. (2019) showing ASR systems being impacted by the age of speakers. Moore (2011) also discusses variations in pronunciation unique to younger speakers, which seems to be reflected in our results since in general audio labelled `Teenager` performs well across all models.

Interestingly, in the French, Portuguese, Dutch, and German datasets (Figure 4), teenagers are the smallest represented age group, *based on the*



(a) English (b) French (c) Dutch

(d) Persian (e) Portuguese (f) German (g) Irish

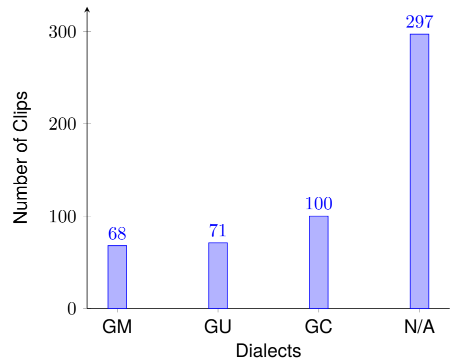
Figure 4: Age Distribution of Training Datasets

annotations, out of all age groups in the test set.

5.4.3. Dialect

We now move to discuss the performance variation across the three dialects labelled within the Irish common voice dataset: Gaeilge Uladh (GU), Gaeilge Chonnacht (GC) and Gaeilge na Mumhan (GM). Apart from the model source fine-tuned on French, audio labelled GU had the lowest WER. The GU dialect has the second most labelled audio associated with it, but GM only falls slightly behind by three audio clips (Figure 5). However, with a significant number of missing dialect labels, it remains unclear if the number of training utterances is a contributing factor in this bias towards GU. The performance difference is, perhaps, better explained by the fact that GU is more distinct from the other two dialects, which models have apparent difficulty disambiguating between. Lonergan et al. (2023a) observed similar behaviour for dialect classification.

Assuming that the distribution of the labelled dialects accurately reflects the true distribution within the data, we can see that the balance of the dialect classes do not necessarily lead to proportionally balanced WER. For example, GC is the most represented dialect but consistently performs worst. This could align with findings by Lonergan et al. (2023a) that balanced corpora do not necessarily lead to balanced performance. Further investigation is needed to determine the sources of this bias.



GM: Gaeilge na Mumhan, GU: Gaeilge Uladh, GC: Gaeilge Chonnacht, N/A: Unknown

Figure 5: Dialect Distribution in Irish Training Dataset

6. Conclusion

This work has demonstrated the advantages and remaining challenges for ASR in Irish using open-source datasets. We demonstrated that, though multi-step fine-tuning can provide performance gains over single-step fine tuning in many cases, these gains are not uniform across different source fine-tuning languages, and more specifically are often not uniform across demographic lines. Therefore, a one-size-fits-all approach is not effective, highlighting the need for detailed analysis of the model's performance instead of a single evaluation metric.

While the usefulness of a dataset often gets overlooked in Machine Learning, we attempted to understand what dataset performs better for a certain low-resource language. We defined this usefulness using both dataset-dependent and independent similarity measures. From our case study of six source languages, we show that certain measures, in particular Genetic Proximity, can be indicative of the optimal pre-training model. While increasing this study to more languages could establish these measures as robust predictors, from our experiments it seems these metrics alone can't show which model will work best, meaning that to tackle this problem we need to consider more factors.

Interestingly, through our experiments on dataset-dependent features, we found that dataset size is not strongly associated with performance gains, indicating that blindly increasing dataset size is unlikely to result in improved performance. Furthermore, the percentage increase in data points does not translate to proportionate performance gains.

We also found that there are often gaps in demographic information. Given that speech data can often identify the speaker, collecting it may be too high-risk, especially for speakers of marginalised groups. This motivates a privacy preserving alternative for measuring demographic coverage.

7. Ethics/Broader Impact Statement

This work relies exclusively on open-source, publicly available datasets for the purposes of reproducibility. Our work focuses on effective model performance in a low-resource setting in order to make these type of speech recognition systems more accessible for different communities. We investigated our method for one specific language (Irish) in order to promote research into the nuances and the unique characteristics of each individual language, something that often gets overlooked in the development of speech recognition models. We evaluate our models' performance across different demographics in order to assess the biases that exists within these systems and to highlight that a single aggregate metric does not take into account the variation in model performance for different speakers.

In our limitations section, we outline that a large amount of speaker metadata is missing from the datasets used in experiments. Though additional metadata labels would undoubtedly facilitate a more rigorous analysis of performance differences between, and at the intersection of, different demographic groups, it is also important for those curating datasets and researchers evaluating ASR to study and understand *why* participants do not feel comfortable disclosing personal information (age, gender etc.) while contributing to datasets. More specifically, it is incumbent on the community to develop ways of evaluating ASR that both protect data subjects and their privacy, while also rigorously testing model errors through an intersectional lens. Based on the lack of speaker metadata in the datasets studied, we endeavor to research this aspect of evaluation further in future work.

8. Limitations

Only open-sourced data from the Common Voice dataset was used in these experiments. Using more datasets could improve the reliability of our experiments. We also only considered six source fine-tuning languages and the base model in all cases was Wav2Vec. Scaling this experiment to more language datasets and base models (such as OpenAI's newer Whisper model (Radford et al., 2023)) could have increased the robustness of our findings and allowed us to detect other factors that lead to performance gains. The source fine-tuned models provided by (Grosman, 2021) do not perfectly match our training conditions. This be another factor contributing to the impact the source-language has on the performance on the model. Another limitation of using the Common Voice was that some language datasets were not available. These included Scottish Gaelic and Manx, which are in the same the Celtic language family as Irish. As described throughout the paper, lack of dataset availability is a common issue within ASR research for low-resource languages such as these.

A large portion of the datasets in Common Voice are missing annotations. This had implications for our study as we endeavored to investigate the link between performance and demographic representation in the source fine-tuning dataset. Certain demographic groups in the datasets were missing more labels than others. For example, of the gender labels, `Female` and `Other` were available for considerably fewer utterances than `Male` . This either indicates that those who do not identify as male are reluctant to disclose their gender, or that they are under-represented in the dataset in the first place. Similarly to the gender labels, age and dialect labels were also missing. It was difficult for us to draw conclusions about the performance across different demographics given this incomplete information. A broader sociolinguistic analysis of utterances is likely necessary to determine the reasons for disparate performance across examples. Furthermore, we also note that we have not analysed the intersection between demographics, which are themselves likely to reveal differences in performance. It is clear that more detailed analysis of pre-training datasets and models is necessary to disentangle the sources of performance differences.

9. Bibliographical References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 4218–4222. European Language Resources Association.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2022. Xls-r: Self-supervised cross-lingual speech representation learning at scale. In *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pages 2278–2282. ISCA.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Central Statistics Office. 2023. Press statement census 2022 results profile 8 - the irish language and education.

Ailbhe Ní Chasaide, Neasa Ní Chiaráin, Christoph Wendler, Harald Berthelsen, Andy Murphy, and Christer Gobl. 2017. The abair initiative: Bringing spoken irish into the digital space. In *INTERSPEECH*, pages 2113–2117.

Neasa Ní Chiaráin, Oisín Nolan, Madeleine Comtois, Neimhin Robinson Gunning, Harald Berthelsen, and Ailbhe Ní Chasaide. 2022. Using speech and nlp resources to build an icall platform for a minority language, the story of an scéalaí, the irish experience to date. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 109–118.

Ruth Holmes, Ellen Rushe, Mathieu De Coster, Maxim Bonnaerens, Shin'ichi Satoh, Akihiro Sugimoto, and Anthony Ventresque. 2023. From scarcity to understanding: Transfer learning for the extremely low resource irish sign language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2008–2017.

Robert Jimerson, Zoey Liu, and Emily Prud'Hommeaux. 2023. An (unhelpful) guide to selecting the best asr architecture for your under-resourced language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1008–1016.

Kazuya Kawakami, Luyu Wang, Chris Dyer, Phil Blunsom, and Aaron van den Oord. 2020. Learning robust and multilingual speech representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1182–1192. Association for Computational Linguistics.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the national academy of sciences*, 117(14):7684–7689.

Eric Le Ferrand, Steven Bird, and Laurent Besacier. 2020. Enabling interactive transcription in an indigenous community. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3422–3428.

Jan Lehečka, Josef V Psutka, Luboš Šmídl, Pavel Ircing, and Josef Psutka. 2024. A comparative analysis of bilingual and trilingual wav2vec models for automatic speech recognition in multilingual oral history archives. In *Interspeech 2024*, pages 1285–1289.

Jinyu Li et al. 2022. Recent advances in end-to-end automatic speech recognition. *APSIPA Transactions on Signal and Information Processing*, 11(1).

Xinjian Li, Siddharth Dalmia, Alan W. Black, and Florian Metze. 2019. Multilingual speech recognition with corpus relatedness sampling. In *20th Annual Conference of the International Speech Communication Association, Interspeech 2019, Graz, Austria, September 15-19, 2019*, pages 2120–2124. ISCA.

Patrick Littell, Kartik Goyal, David R Mortensen, Alexa N Little, Chris Dyer, and Lori Levin. 2016. Named entity recognition for linguistic rapid response in low-resource languages: Sorani kurdish and tajik. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 998–1006.

Liam Lonergan, Mengjie Qian, Harald Berthelsen, Andy Murphy, Christoph Wendler, Neasa Ní Chiaráin, Christer Gobl, and Ailbhe Ní Chasaide.

2022. Automatic speech recognition for irish: the abair-éist system. In *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, pages 47–51.

Liam Lonergan, Mengjie Qian, Neasa Ní Chiaráin, Christer Gobl, and Ailbhe Ní Chasaide. 2023a. Towards spoken dialect identification of irish. In *Proceedings of the 2nd annual meeting of the Special Interest Group of Under-resourced Languages, a Workshop at Interspeech 2023*.

Liam Lonergan, Mengjie Qian, Neasa Ní Chiaráin, Christer Gobl, and Ailbhe Ní Chasaide. 2023b. Towards dialect-inclusive recognition in a low-resource language: Are balanced corpora the answer? In *Proc. Interspeech 2023*, pages 5082–5086.

Teresa Lynn. 2023. Language report irish. In *European Language Equality: A Strategic Agenda for Digital Language Equality*, pages 163–166. Springer.

Nina Markl. 2022. Language variation and algorithmic bias: understanding algorithmic bias in british english automatic speech recognition. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 521–534.

Vikramjit Mitra, Andreas Kathol, Jonathan D Amith, and Rey Castillo García. 2016. Automatic speech transcription for low-resource languages-the case of yoloxóchitl mixtec (mexico). In *INTERSPEECH*, pages 3076–3080.

Robert Moore. 2011. "if i actually talked like that, i'd pull a gun on myself": accent, avoidance, and moral panic in irish english. *Anthropological Quarterly*, pages 41–64.

Yanmin Qian and Zhikai Zhou. 2022. Optimizing data usage for low-resource speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:394–403.

Thomas Reitmaier, Electra Wallington, Dani Kalarikalayil Raju, Ondrej Klejch, Jennifer Pearson, Matt Jones, Peter Bell, and Simon Robinson. 2022. Opportunities and challenges of automatic speech recognition systems for low-resource language speakers. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–17.

Samuel Thomas, Kartik Audhkhasi, Jia Cui, Brian Kingsbury, and Bhuvana Ramabhadran. 2016. Multilingual data selection for low resource speech recognition. In *Interspeech*, pages 3853–3857.

Hasti Toossi, Guo Qing Huai, Jinyu Liu, Eric Khiu, A Seza Doğruöz, and En-Shiun Annie Lee. 2024. A reproducibility study on quantifying language similarity: The impact of missing values in the uriel knowledge base. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop, NAACL 2024, Mexico City, Mexico, June 18, 2024*, pages 233–241. Association for Computational Linguistics.

Beaufils Vincent and Tomin Johannes. 2020. Stochastic approach to worldwide language classification: the signals and the noise towards long-range exploration.

Lauren Werner, Gaojian Huang, and Brandon J Pitts. 2019. Automated speech recognition systems and older adults: a literature review and synthesis. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 63, pages 42–46. SAGE Publications Sage CA: Los Angeles, CA.

Peter Wu, Jiatong Shi, Yifan Zhong, Shinji Watanabe, and Alan W Black. 2021. Cross-lingual transfer for speech processing using acoustic language similarity. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1050–1057. IEEE.

Li Yang and Abdallah Shami. 2020. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415:295–316.

Cheng Yi, Jianzhong Wang, Ning Cheng, Shiyu Zhou, and Bo Xu. 2020. Applying wav2vec2. 0 to speech recognition in various low-resource languages. *arXiv preprint arXiv:2012.12121*.

Saierdaer Yusuyin, Te Ma, Hao Huang, Wenbo Zhao, and Zhijian Ou. 2025. Whistle: Data-efficient multilingual and crosslingual speech recognition via weakly phonetic supervision. *IEEE Transactions on Audio, Speech and Language Processing*.

Yu Zhang, Ekapol Chuangsuwanich, and James Glass. 2014. Language id-based training of multilingual stacked bottleneck features. In *Proc. Interspeech*, pages 1–5. Citeseer.

10. Language Resource References

Conneau, Alexis and Baevski, Alexei and Collobert, Ronan and Mohamed, Abdelrahman and Auli, Michael. 2020. *Unsupervised cross-lingual representation learning for speech recognition*. Model available on HuggingFace: https://huggingface.co/facebook/wav2vec2-large-xlsr-53.

Elinguistics. 2020. *Elinguistics*. Accessed: 2025-02-10.

Grosman, Jonatas. 2021. *Fine-tuned XLSR-53 large models for speech recognition*.

Littell, Patrick and Mortensen, David R and Lin, Ke and Kairis, Katherine and Turner, Carlisle and Levin, Lori. 2017. *URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors*.

Liu, Yi and Fung, Pascale and Yang, Yongsheng and Cieri, Christopher and Huang, Shudong and Graff, David. 2006. *Hkust/mts: A very large scale mandarin telephone speech corpus*. Springer.

Mozilla Corporation. 2021. *Mozilla Common Voice*. Accessed: 29th September 2023.

Panayotov, Vassil and Chen, Guoguo and Povey, Daniel and Khudanpur, Sanjeev. 2015. *Librispeech: an asr corpus based on public domain audio books*. IEEE.

Park, Kyubyong and Mulc, Thomas. 2019. *CSS10: A Collection of Single Speaker Speech Datasets for 10 Languages*.

Radford, Alec and Kim, Jong Wook and Xu, Tao and Brockman, Greg and McLeavey, Christine and Sutskever, Ilya. 2023. *Robust speech recognition via large-scale weak supervision*. PMLR.

Schneider, Steffen and Baevski, Alexei and Collobert, Ronan and Auli, Michael. 2019. *wav2vec: Unsupervised pre-training for speech recognition*.

Blank-Aware Decoding for Transcript-Free Phoneme Alignment in Low-Resource Languages and Dialects

Domenico De Cristofaro[1,2,3], Barbara Plank[3,4], Alessandro Vietti[1,2]
Free University of Bozen[1], ALPS[2], MaiNLP[3], LMU Munich[4]
ddecristofaro@unibz.it, b.plank@lmu.de, avietti@unibz.it,

Abstract

We present a blank-aware decoding approach for transcript-free phoneme alignment with CTC-based speech foundation models, designed to improve annotation bootstrapping in low-resource languages. While CTC models provide frame-level phoneme posteriors without requiring transcripts, greedy decoding produces blank-dominated and temporally unstable segmentations that are difficult to correct manually. Our approach introduces two training-free blank-resolution strategies operating directly on CTC logits: (i) confidence-ratio substitution, which promotes competitive non-blank hypotheses relative to the blank symbol, and (ii) recursive context adjustment, which enforces local contextual consistency within blank spans. Experiments on English (TIMIT) and on Sardinian and Tyrolean dialect corpora show consistent improvements in boundary F1 prediction, phoneme duration regularity, and segmentation stability over greedy CTC decoding. Although absolute boundary deviations remain higher than transcript-conditioned aligners, the resulting alignments are structurally coherent and suitable for manual correction. A post-hoc phoneme-class analysis further reveals systematic asymmetries in blank resolution, highlighting complementary roles of local acoustic evidence and contextual cues, and outlining prominising venues for future improvements.

1. Introduction and Motivation

Developing speech corpora for low-resource and under-documented languages is a demanding process. Beyond the challenges of data collection and audio recording, the most critical bottleneck lies in annotation, especially phonetic ones. Many low-resource varieties lack a standardized orthography, requiring phonetic rather than orthographic transcription (Le Ferrand et al., 2025). Such annotation demands trained phoneticians, who are even more scarcely available for minority and dialectal languages, in contrast to standard varieties. Moreover, for speech analysis, corpus development, and model interpretability studies (Pasad et al., 2024; Choi et al., 2024), annotations must be time-aligned at either word or phoneme level. Producing temporally precise phonetic boundaries manually is labor-intensive and represents a major obstacle to scaling documentation efforts.

Automatic forced alignment systems can alleviate this burden when transcripts are available. Classical HMM-based aligners and modern neural aligners achieve high boundary accuracy under transcript-conditioned decoding (Young et al., 2006; Schiel, 1999; McAuliffe et al., 2017). However, in genuinely low-resource settings, transcripts may be unavailable, unreliable, or themselves be costly to produce. Transcript-free alignment therefore represents an appealing alternative for bootstrapping phonetic annotation (Draxler, 2022).

Connectionist Temporal Classification (CTC)–based speech foundation models provide frame-level posterior distributions over phoneme vocabularies without requiring alignment supervision during training (Graves et al., 2006). This makes them promising candidates for transcript-free phoneme alignment. In practice, however, greedy CTC decoding yields unstable and fragmented boundaries. The dominant cause is the pervasive presence of the *blank* symbol in the decoding path: most frames are assigned to blank, with sparse non-blank emissions separated by long blank spans (see Figure 1). As a result, raw transcript-free CTC outputs lack strong temporal anchoring and are poorly suited for direct use in annotation. This dominance of blank is not incidental but a structural consequence of the CTC objective. By marginalizing over all possible frame-to-label alignments, CTC allows the blank symbol to absorb temporal uncertainty, functioning as a buffer between phoneme predictions. The learned posteriors therefore become "peaky" concentrating mass on blank except at a few frames (Huang et al., 2024; Zeyer et al., 2021). While transcript-conditioned forced alignment constrains this uncertainty via Viterbi decoding against a known phoneme sequence (Yang et al., 2023), the transcript-free case lacks such structural guidance, leading to boundary instability.

Crucially, we observe that blank frames are not informationally empty. Although blank often dominates the top-1 decoding path, the posterior distribution over non-blank symbols typically exhibits structured alternatives. The second-highest probability frequently aligns with either the preceding or the following phoneme prediction, or with a phoneme acoustically consistent with the surrounding context. This indicates that *blank spans encode latent phonetic structure rather than random*

Proceedings of SPEAKABLE @ LREC 2026, pages 174–182
11 May 2026.

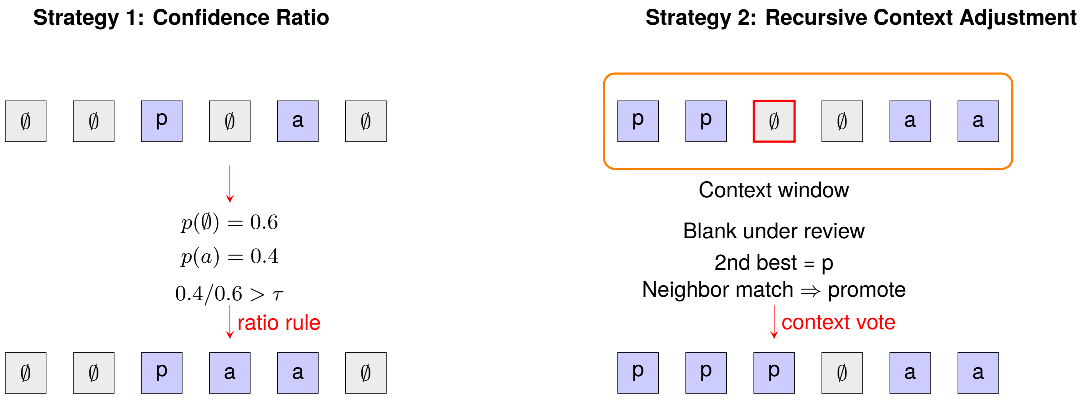


Figure 1: Two blank-handling strategies: (1) Confidence-ratio substitution; (2) Recursive context adjustment.

noise: the model maintains contextual continuity signals beneath blank dominance. Transcript-free CTC outputs are therefore not arbitrary but structurally underdetermined. We argue that improving transcript-free alignment requires recovering this latent structure instead of suppressing blanks indiscriminately. To this end, we introduce two complementary, training-free blank-resolution strategies that operate directly on CTC posteriors. The first, *confidence-ratio substitution*, promotes a non-blank hypothesis when it is sufficiently competitive relative to the blank. The second, *recursive context adjustment*, resolves blank segments by enforcing consistency with neighboring predictions within a local window. Both strategies exploit structured competition in the posterior distribution without modifying the acoustic model. We evaluate these approaches on TIMIT (Garofolo et al., 1993) for English as a controlled benchmark, on a Sardinian (Chizzoni and Vietti, 2025) dataset and an in-house Tyrolean dialect corpus representing realistic low-resource conditions. A post-hoc analysis further reveals systematic phoneme-class asymmetries: vowels exhibit strong context-driven blank resolution, while obstruents remain locally anchored. These findings help understand when and why blank-handling strategies are effective and demonstrate that transcript-free CTC alignment can serve as a practical bootstrapping tool for low-resourced languages phonetic corpus development.

All code and evaluation scripts is available at Github.

2. Related Work

Classical forced alignment is typically performed with HMM pipelines such as HTK (Young et al., 2006), MAUS (Schiel, 1999), and Montreal Forced Aligner (MFA) (McAuliffe et al., 2017). With accurate transcripts these systems deliver strong boundary accuracy, but they are difficult to deploy in low–resource scenarios where transcriptions are noisy or unavailable. Neural approaches increasingly rely on end-to-end ASR models trained with Connectionist Temporal Classification (CTC) (Graves et al., 2006). Torchaudio's alignment API (Yang et al., 2023) performs Viterbi alignment on CTC posteriors conditioned on transcripts, but it does not address the transcript-free case. Parallel work in zero-resource learning targets unsupervised phoneme segmentation: the ZeroSpeech benchmarks (Dunbar et al., 2017; Zhang et al., 2020) have fostered methods based on clustering (Ondel et al., 2016), self-supervised representations (Baevski et al., 2020), and predictive coding (Liu et al., 2022). These techniques discover units but lack phoneme-level temporal precision. Closer to our setting are decoding heuristics from ASR—confidence thresholding (Hwang and Sung, 2016), top-k rescoring (Watanabe et al., 2017), and blank suppression (Liu et al., 2021). While effective for recognition, they have not been systematically studied for transcript-free alignment. Our work adapts these ideas to alignment, proposing training-free, confidence- and context-aware decoding directly on CTC logits. In addition, we provide a post-hoc perturbation and reliability analysis across broad phoneme classes, clarifying when local acoustic evidence versus contextual cues resolve blanks, an angle missing from prior alignment studies.

3. Proposed Methods

Let $\mathbf{z}_t \in \mathbb{R}^C$ be the CTC logits at frame $t = 1, \dots, T$ over the phone vocabulary $\mathcal{V}$ augmented with the blank symbol $\varnothing$. Posteriors refer to the frame-level probability distribution over phoneme labels obtained by applying a softmax to the CTC logits. Let $(c_t^{(k)}, p_t^{(k)})_{k=1}^{K}$ denote the top-K symbols and probabilities at frame t in descending order ($K \in \{2, 3, 4\}$ in our code). The blank ID is the tokenizer pad ID, and we write $c = \varnothing$ when c equals that ID. A frame-wise label sequence $\hat{y}_{1:T}$ is converted to segments by merging consecutive identical labels.

3.1. Confidence–ratio blank substitution

CTC paths are dominated by blanks ($c_t^{(1)} = \varnothing$). We replace a blank at t with a non-blank candidate only if its posterior is sufficiently competitive relative to the blank, as measured by their ratio — the *confidence ratio* (CR):

$$\hat{y}_t = \begin{cases} c_t^{(k^*)}, & \text{if } c_t^{(1)} = \varnothing \\ & k^* = \min\left\{k \in \{2,\dots,K\} : \dfrac{p_t^{(k)}}{p_t^{(1)}} > \tau\right\} \\ c_t^{(1)}, & \text{otherwise.} \end{cases} \tag{1}$$

where $\tau \in (0,1)$ is a ratio threshold. The ratio condition ensures that a candidate phone is competitive with the blank. This heuristic thus keeps blanks by default and only replaces them when a competing phone is strong relative to the blank.

3.2. Recursive Context Adjustment

3.2.1. Proto-segmentation

We first compress the top-1 sequence into *proto segments* $S_i = (t_i^s, t_i^e)$, $i = 1,\dots,M$, each carrying its top-K list $\Pi_i = \Big((c_i^{(1)}, p_i^{(1)}), (c_i^{(2)}, p_i^{(2)}), (c_i^{(3)}, p_i^{(3)})\Big)$. Leading blanks are dropped until the first non-blank appears.

3.2.2. Neighborhood set

For a given position i, $\mathcal{N}_i$ denotes the set of first-choice candidates $c_j^{(1)}$ occurring within a window of w ($w = 2$) positions to the left and the right of i, excluding the current position and restricted to valid indices $j \in [1, M]$

$$\mathcal{N}_i = \{\, c_j^{(1)} \;\mid j \in \{i-w,\dots,i-1,\; i+1,\dots,i+w\} \cap [1,M]\,\}. \tag{2}$$

3.2.3. Blank resolution rule

For any *blank* proto segment ($c_i^{(1)} = \varnothing$), we look for an alternative among its top-K candidates that is context-consistent:

$$k^* = \arg \max_{k\in\{2,\dots,K\},\; c_i^{(k)} \in N_i} p_i^{(k)}, \quad \tilde{c}_i = c_i^{(k^*)} \tag{3}$$

If such a candidate exists, we *promote* it to the top of Π_i (i.e., swap with $c_i^{(1)}$); otherwise S_i remains blank. Because promoting S_i can change the neighborhoods of $S_{i\pm d}$, we sweep over $i = 1 \dots M$ and apply (3) repeatedly until no segment changes:

$$\Pi^{(r+1)} \leftarrow \mathrm{AdjustOnce}\big(\Pi^{(r)}\big) \quad \text{until} \quad \Pi^{(r+1)} = \Pi^{(r)}.$$

3.3. Post-Processing

Termination is guaranteed in finite steps because each successful update replaces a blank top-1 with a non-blank and no rule reintroduces blanks. Finally, any residual blanks are discarded and consecutive identical labels are merged into phone segments. Importantly, boundary timestamps are preserved as predicted by the decoding path; no additional boundary snapping or temporal re-alignment is performed during merging. Finally, optional leading and trailing silence segments are added using a simple energy-based detection criterion, based on low short-time signal energy, in order to avoid artificial padding at utterance boundaries. These steps preserve the decoding semantics while producing well-formed segments for evaluation.

4. Experimental settings

4.1. Hyperparameter Selection

Both blank-resolution strategies introduce a small number of decoding hyperparameters: the threshold τ for confidence-ratio substitution, and the number of alternative candidates K considered during recursive adjustment. All hyperparameters are selected via grid search without retraining, operating directly on the CTC logits. For confidence-ratio substitution, we vary the ratio threshold $\tau \in \{0.05, 0.1, 0.2, 0.3\}$. For recursive context adjustment, we explore the number of considered alternatives $K \in \{2,3,4\}$, corresponding to the top-K non-blank candidates available at each blank segment.

Hyperparameters are tuned separately for each dataset and speech condition. Our objective is not to learn globally transferable decoding parameters, but to optimize segmentation stability for each practical annotation scenario. Since the strategies are decoding-only and do not modify model weights, tuning does not affect the underlying acoustic representations. The selection follows a multi-objective criterion focused on segmentation quality. For each configuration we compute (i) average boundary deviation (ABD), (ii) phoneme duration error (PDUR), and (iii) boundary F1. Before aggregation, each metric is normalized to $[0,1]$ using min-max scaling across configurations, so that no single metric dominates due to scale differences (ABD is in milliseconds while F1 is bounded in $[0,1]$). All three metrics are treated as equally important, reflecting our practical prior that temporal accuracy, duration regularity, and segment coherence are all relevant to annotation bootstrapping; no principled basis for differential weighting was available a priori. Metrics are averaged across files within each dataset (and, for Tyrolean, within each speech condition), and configurations are ranked using the resulting aggregate score. Lower aggregate scores indicate better overall trade-offs, corresponding to lower ABD and PDUR values and higher boundary F1. Although phoneme error rate (PER) was not used as a selection criterion, we report it for completeness (Table 1). On both English (TIMIT) and Sardinian, confidence-ratio substitution leaves PER unchanged relative to greedy decoding, while recursive adjustment slightly increases PER (e.g., 0.29→0.33 on TIMIT). At the same time, recursive adjustment substantially improves boundary F1 and phoneme duration regularity (Table 3). This pattern indicates that segmentation stability and phoneme identity accuracy are partially decoupled in transcript-free CTC decoding. The proposed strategies primarily regularize temporal structure rather than improving phoneme classification, which justifies excluding PER from hyperparameter optimization.

Dataset	Strategy	PER (%)
TIMIT	Base	29
	CR	29
	Rec	33
Sardinian	Base	47
	CR	47
	Rec	48

Table 1: Phoneme Error Rate (PER) for each decoding strategy on TIMIT and Sardinian.

PER is not reported for Tyrolean due to inventory differences and limited phoneme normalization across dialect-specific variants, which would make cross-condition comparisons unreliable. Nevertheless, segmentation trends on Tyrolean mirror those observed on TIMIT and Sardinian.

Dataset	Condition	CR (τ)	Rec (K)
TIMIT	read	0.2	4
Sardinian	spont.	0.2	4
Tyrolean	read	0.05	3–4
Tyrolean	spont.	0.1–0.2	**2**

Table 2: Selected hyperparameters for confidence-ratio substitution (CR) and recursive context adjustment (Rec), obtained via grid search. Tyrolean results are reported separately for read speech and spontaneous monologues.

Notably, spontaneous Tyrolean monologues favor a smaller context size ($K = 2$), in contrast to other datasets where larger K values perform better. This suggests that in highly variable spontaneous speech, broader contextual voting may introduce instability, and more conservative blank resolution is preferable.

4.2. Datasets

We evaluate on three speech corpora. The test set of TIMIT (Garofolo et al., 1993) with a total of 34.35 minutes is used as a controlled benchmark, providing manually time-aligned phoneme boundaries with 61 labels mapped to the standard 39-phone set. In addition, to assess performance in low-resource settings, we use two dialectal corpora unseen during model training. The Sardinian corpus contains extracts of longer monologues with a total of 40 minutes of spontaneous Campidanese speech from four speakers, manually annotated by native-speaker phoneticians (Chizzoni and Vietti, 2025). The Tyrolean corpus comprises approximately 95 minutes of speech, including both spontaneous monologues and read speech, collected and annotated by the authors, as part of an ongoing documentation effort and not yet publicly released. Together, the Sardinian and Tyrolean data represent realistic low-resource conditions with spontaneous speech, dialectal variation, and heterogeneous phoneme inventories.

4.3. Model and Features

All experiments use the `wav2vec2-xlsr-53-espeak-cv-ft` model, a multilingual phoneme-level CTC system trained with espeak-generated labels. The underlying XLSR-53 backbone is pretrained in a self-supervised manner on speech from 53 languages (Baevski et al., 2020), learning acoustic representations directly from raw audio without supervision. The model is subsequently fine-tuned to predict over a cross-lingual phoneme vocabulary of roughly 360 IPA symbols plus the blank token, reflecting the union of phonemes across the training languages (Xu et al., 2021). We chose this model because its self-supervised pretraining yields acoustically grounded representations that are not conditioned on transcript alignment during representation learning. Although the final CTC layer is trained with G2P-derived phoneme labels, the underlying encoder retains rich acoustic structure, making it particularly suitable for transcript-free boundary inference. From this model, we extract frame-level posterior distributions at 20 ms resolution and apply our decoding strategies for boundary assignment.

5. Results and Discussion

Table 3 reports alignment results on the three corpora: TIMIT, Sardinian, and Tyrolean. For corpus

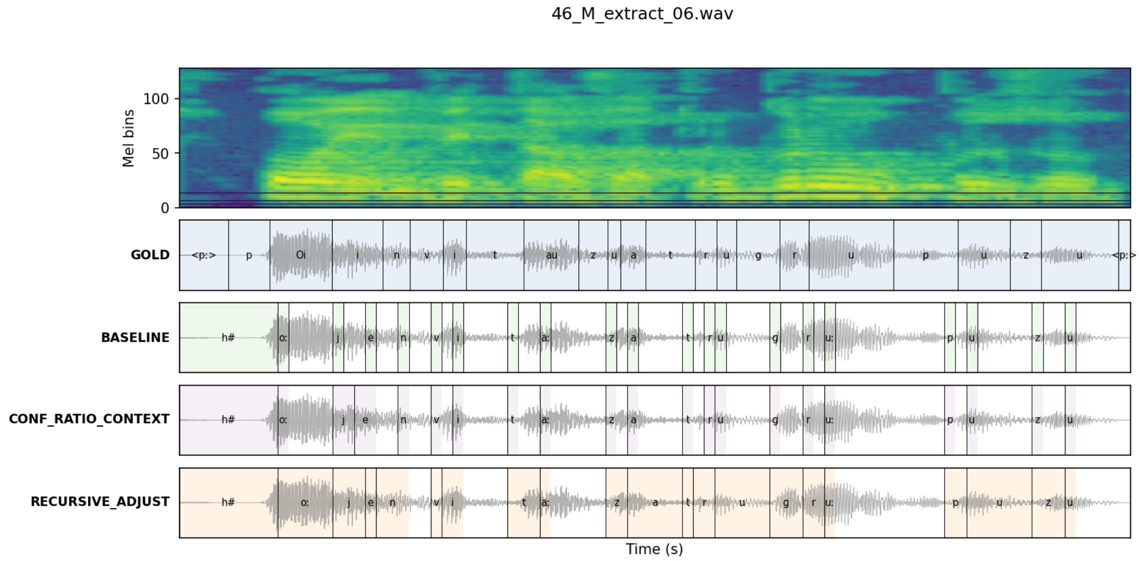


Figure 2: Example phoneme alignment on a Sardinian speech segment ("poi invitaus àterus grupus"; "and then they invited another group"). From top to bottom: gold annotation, greedy CTC baseline, confidence-ratio substitution, and recursive context adjustment.

Dataset	Strat.	ABD	PDUR	Prec	F1
TIMIT	Base	100.73	63.33	0.20	0.20
	CR	**100.57**	62.04	0.21	0.21
	Rec	114.30	**55.09**	**0.23**	**0.23**
Sard.	Base	137.38	71.38	0.17	0.17
	CR	**137.29**	70.12	0.17	0.17
	Rec	142.35	**62.87**	**0.23**	**0.23**
Tyr.-sent	Base	98.52	57.63	0.26	0.26
	CR	**98.05**	**54.63**	0.28	0.28
	Rec	127.36	72.04	**0.34**	**0.34**
Tyr.-mon	Base	1025.87	69.97	0.25	0.24
	CR	1025.79	**66.53**	0.27	0.27
	Rec	**856.49**	88.65	**0.32**	**0.32**

Table 3: Average alignment scores across datasets. CR = confidence-ratio substitution; Rec = recursive context adjustment.

bootstrapping, an alignment is considered usable when phoneme segments are temporally coherent and require local correction rather than full re-segmentation (Draxler, 2022). Therefore, we evaluate alignment quality using three complementary segmentation metrics: average boundary deviation (ABD, ms), phoneme duration error (PDUR, %), and boundary F1. On TIMIT, the greedy CTC baseline exhibits large boundary deviations (ABD ≈ 100 ms) and high duration error (PDUR ≈ 63 ms), confirming that raw CTC decoding paths are poorly suited for temporal alignment. Both proposed blank-resolution strategies improve segmentation quality without supervision. Confidence-ratio substitution yields a small but consistent reduction in ABD, while maintaining comparable duration error and boundary F1. Recursive context adjustment produces the strongest improvement in duration consistency, reducing PDUR from 63 to 55 ms, and also yields the highest boundary F1. This comes at the cost of increased ABD, revealing an explicit trade-off between boundary anchoring and segment-level regularization. Under the joint ABD–PDUR–F1 criterion used for hyperparameter selection, recursive adjustment provides the most favorable trade-off under the chosen multi-metric criterion on TIMIT.

The trade-off between ABD and PDUR reflects structurally distinct aspects of segmentation quality. ABD measures the absolute temporal distance between predicted and gold boundaries: even a single well-anchored boundary per phoneme can keep ABD low. PDUR, by contrast, captures duration regularity across the full segmentation. Greedy CTC decoding tends to produce at least one acoustically grounded boundary per phoneme, typically at the onset or offset of the clearest acoustic event, while blank spans inflate perceived duration. Recursive context adjustment collapses these blank spans, improving duration regularity at the cost of displacing the originally well-anchored boundary, which increases ABD. Confidence-ratio substitution, being more conservative, preserves those anchored boundaries while making smaller local adjustments to duration.

Results on the Sardinian corpus show the same qualitative trends in a more challenging low-resource setting with spontaneous speech. Absolute boundary deviations are higher than on TIMIT (~ 140 ms vs. ~ 100 ms), reflecting increased temporal variability and phonetic diversity. Confidence-ratio substitution slightly reduces ABD while leaving PDUR and boundary F1 largely unchanged. Recursive context adjustment substantially improves duration regularity (PDUR $71 \rightarrow 63$ ms) and boundary F1, at the cost of increased ABD. Recursive adjustment again provides the most favorable overall trade-off for downstream annotation bootstrapping.

For Tyrolean, results are reported separately for read speech (*sent*) and spontaneous monologues (*mon*), revealing a strong interaction between speaking style and decoding strategy. On read speech, confidence-ratio substitution achieves the best joint trade-off across ABD, PDUR, and boundary F1, yielding the lowest boundary deviation and duration error. Recursive adjustment increases both ABD and PDUR, but substantially improves boundary F1, indicating stronger structural consistency at the expense of temporal precision.

In spontaneous monologues, baseline alignment quality degrades substantially (ABD > 1000 ms), highlighting the difficulty of transcript-free alignment in highly variable speech. In this setting, recursive context adjustment is particularly effective, reducing ABD by more than 15% $(1026 \rightarrow 856$ ms) and yielding the highest boundary F1. Although PDUR increases, the joint ABD–PDUR–F1 score favors recursive adjustment, indicating that strong structural regularization is beneficial for spontaneous, low-resource speech. Figure 2 illustrates these contrasting behaviors on Sardinian example. Which strategy is preferred depends on the joint behavior of ABD, PDUR, and boundary F1, and varies systematically with speaking style. Also both strategies could help for different bootstrapping set up.

5.1. Post-hoc Edit Analysis and Correction Time

To assess the practical usability of the proposed alignment strategy, we qualitatively examined the relationship between the number of phoneme edit operations and the human correction time required to obtain the final reference transcription and alignment for four files of spontaneous Tyrolean dialect. Importantly, correction time reflects the full annotation effort: it includes both phoneme identity corrections (substitutions, insertions, deletions) and temporal boundary adjustments. Files with a larger number of edit operations generally required longer correction times, indicating that token-level errors contribute substantially to annotation workload. However, the relationship is not strictly proportional. In several cases, boundary refinements required careful acoustic inspection even when phoneme identities were largely correct, increasing correction time independently of the raw edit count. Conversely, clusters of local phoneme substitutions could often be corrected relatively quickly once the surrounding context was clear. These observations are consistent with the hypothesis that total correction time reflects a combined cost of segmental errors and temporal misalignment, though the small sample size prevents stronger conclusions. As such, phoneme error rate alone does not fully capture the annotation burden associated with transcript-free alignment.

File	Ref	S	D	I	Edits	PER
0008mon001	1448	121	80	61	262	18.09
0008mon002	892	97	32	63	192	21.52
0008mon003	882	85	30	58	173	19.61
0008mon004	1048	145	47	64	256	24.43
Total	4270	448	189	246	883	20.68

Table 4: Phoneme-level edit statistics for Recursive Adjustment compared to the corrected reference transcription. S = substitutions, D = deletions, I = insertions.

File	Correction Time	Audio Length	RTF
0008mon001	1:22:42	2.4	34.4
0008mon002	0:49:45	1.4	35.5
0008mon003	0:57:46	1.3	44.4
0008mon004	1:13:32	1.5	49.02
Total	4:23:45	6.6	40

Table 5: Human correction time for boundary and transcription refinement. RTF (Real-Time Factor) is computed as correction time divided by audio duration.

Tables 4 and 5 provide complementary perspectives on alignment usability. While phoneme-level PER ranges between 18% and 24%, the corresponding human correction time varies more substantially, with real-time factors (RTF) between 34 and 49, and an overall RTF of approximately 40. Here, RTF is defined as the ratio between human correction time and audio duration; for example, an RTF of 40 indicates that correcting one minute of audio requires approximately 40 minutes of manual work.

Notably, files with comparable PER values exhibit markedly different correction times. For instance, 0008mon003 does not have the highest PER, yet shows one of the largest RTF values. This mismatch indicates that transcription edits alone do not fully explain annotation effort. Correction time includes both phoneme identity changes and boundary refinements, the latter often requiring careful acoustic inspection even when segmental labels are correct. These findings suggest that PER provides only a partial proxy for annotation workload, and that temporal instability contributes significantly to human correction cost in transcript-free alignment, coroborating similar findings (Martin et al., 2024). For reference, prior work reports real-time factors between 10 and 50 for automatic segmentation followed by manual boundary correction alone (Draxler, 2022). In our case, correction time in-

Class	n	Prev (%)	Next (%)
Stop	8051	3.1	6.8
Fricative	9109	2.5	5.9
Affricate	454	0.8	4.6
Vowel	20123	9.8	22.9
Nasal	3271	2.1	6.5
Liquid	3583	2.9	6.9
Glide	1003	1.4	2.2

Table 6: Post-hoc agreement of the second-best candidate with the gold previous or next phoneme, by broad phoneme class (TIMIT).

Class	Context-only $\Delta p2$	Local occlusion $\Delta p2$
Stop	0.03	0.07
Fricative	0.02	0.05
Affricate	0.01	0.04
Vowel	0.09	0.15
Nasal	0.03	0.08
Liquid	0.04	0.09
Glide	0.02	0.05

Table 7: Median change in the second-best posterior ($\Delta p2$) under context-only vs. local occlusion perturbations within a $\pm$40 ms window around blank frames. Vowels show stronger context effects, while obstruents are more locally anchored.

cludes both boundary refinement and phoneme identity correction, making the task strictly more demanding. Therefore, the observed RTF values ($\approx 34 - 49$, overall ≈ 40) are consistent with this range, though direct comparison is limited since our correction task encompasses both boundary refinement and phoneme identity correction. We emphasize that this analysis is preliminary and based on a small sample of four recordings; it is intended as an initial indication of annotation effort rather than a definitive evaluation.

5.2. Post-hoc Analysis of Blank Frames

To our knowledge, prior work has not systematically quantified blank spans as a function of phoneme class. This provides new evidence that vowels are context-driven, while obstruents remain locally anchored. We restrict the following post-hoc blank analysis to TIMIT, where gold phoneme labels are phonetically consistent and directly comparable across speakers. For Sardinian, we observed the same qualitative trends, but leave a detailed phoneme-class breakdown for future work given its larger inventory and sparser annotation. Table 6 shows how often the second-best candidate in blank frames agrees with the gold previous or next phoneme. The agreement is highest for vowels (Prev: 9.8%, Next: 22.9%), reflecting that blank spaces in vowel regions are strongly influenced by neighboring phones. Stops, fricatives, and nasals show a much weaker agreement ($< 7\%$), consistent with their short duration and sharper acoustic cues. This asymmetry is consistent with recursive adjustment benefiting vowels most, since it explicitly exploits neighborhood consistency, while having limited effect for obstruents whose boundaries are more locally anchored.

Table 7 compares the median change in the second-best posterior ($\Delta p2$) under two perturbations: (i) *context-only*, keeping only a $\pm$40 ms window around the blank, and (ii) *local occlusion*, masking the same window. For vowels, context perturbations produce much larger shifts than occlusions ($\Delta p2 = 0.15$ vs. 0.09), showing that vowel blanks are primarily resolved by broader temporal context. For obstruents such as stops and fricatives, occlusion induces slightly stronger effects than context, reflecting their sharper, locally anchored cues. Overall, these results suggest that blanks are often resolved by context for sonorants, but by local acoustics for obstruents. This asymmetry further justifies our design strategy: confidence-ratio substitution captures locally competitive cases, while recursive adjustment exploits contextual agreement, especially effective for vowels.

6. Conclusion

We introduced two training-free blank-resolution strategies for transcript-free phoneme alignment with CTC-based speech foundation models. Across TIMIT, Sardinian, and Tyrolean, both heuristics improve segmentation quality over greedy CTC decoding, with recursive context adjustment yielding the strongest overall trade-offs when jointly optimizing ABD, PDUR, and boundary F1.

Our results highlight a key property of CTC-based models: they robustly encode phoneme *identity*, yet lack strong temporal anchoring. Blank-aware decoding partially mitigates this by stabilizing blank spans and enforcing structural consistency, producing more coherent segmentations suitable for manual correction. However, absolute boundary precision remains limited (ABD ~100 ms vs. ~20 ms in transcript-conditioned aligners), underscoring the inherent difficulty of transcript-free alignment.

Post-hoc analysis further reveals phoneme-class asymmetries: vowels benefit most from context-driven blank resolution, whereas obstruents remain more locally anchored. This suggests that a single global decoding rule is suboptimal, and motivates future work on adaptive blank-resolution strategies informed by phoneme class or posterior sensitivity.

Beyond metric improvements, the main contribution of this work is *practical*: our approach provides a lightweight bootstrapping mechanism for corpus development when transcripts are unavailable or unreliable. Preliminary correction-time analysis on four spontaneous Tyrolean recordings suggests that temporal instability contributes substantially to human workload, though a larger-scale user study would be needed to quantify annotation time savings robustly. Future work should explore integrating adaptive blank resolution into semi-supervised or joint training pipelines for low-resource languages and dialects.

7. Limitations

While the proposed strategies substantially improve segmentation stability over greedy CTC decoding, transcript-free alignment remains less temporally precise than transcript-conditioned forced aligners, with absolute boundary deviations remaining around 100 ms. This reflects a fundamental limitation of transcript-free alignment rather than of the proposed methods. All experiments are conducted using a single multilingual CTC-based phoneme recognizer, allowing us to isolate decoding effects; future work should assess generalization across architectures. Moreover, the strategies are heuristic and decoding-only, and cannot correct systematic biases learned during acoustic model training. Finally, optimal behavior varies with speaking style, suggesting that adaptive or class-aware blank resolution may be preferable to a single global rule. Finally, our analysis of human correction time is based on four spontaneous monologues, representing a limited sample size. Although these recordings reflect realistic zero-shot, low-resource and spontaneous conditions and therefore provide ecologically valid evidence of annotation effort, the results should be interpreted as preliminary. Correction time may vary across speakers, speaking styles, recording quality, and annotator expertise. A larger-scale user study would be required to obtain statistically robust estimates of annotation time savings and to quantify inter-annotator variability.

8. Acknowledgements

Funded by the European Social Fund Plus Project code ESF2_f3_0003 “Excellence Scholarships for PhD students on topics of strategic relevance for South Tyrol”

9. Bibliographical References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460.

Kwanghee Choi, Ankita Pasad, Tomohiko Nakamura, Satoru Fukayama, Karen Livescu, and Shinji Watanabe. 2024. Self-supervised speech representations are more phonetic than semantic.

Christoph Draxler. 2022. Automatic transcription of spoken language using publicly available web services. In Jacopo Saturno and Lorenzo Spreafico, editors, *Fare linguistica applicata con le digital humanities*, volume 14 of *Studi AItLA*, pages 27–47. AItLA.

Ewan Dunbar, Xuan-Nga Cao, Jorge Benjumea, Julien Karadayi, Marianne Bernard, Laurent Besacier, Thomas Schatz, Maarten Versteegh, and Emmanuel Dupoux. 2017. The zero resource speech challenge 2017. In *Proc. ASRU*, pages 323–330.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proc. ICML*, pages 369–376.

Ruizhe Huang, Xiaohui Zhang, Zhaoheng Ni, Li Sun, Moto Hira, Jeff Hwang, Vimal Manohar, Vineel Pratap, Matthew Wiesner, Shinji Watanabe, Daniel Povey, and Sanjeev Khudanpur. 2024. Less peaky and more accurate ctc forced alignment by label priors.

Kyu J. Hwang and Wonyong Sung. 2016. Character-level incremental speech recognition with recurrent neural networks. In *Proc. Interspeech*, pages 2420–2424.

Eric Le Ferrand, Bo Jiang, Joshua Hartshorne, and Emily Prud'hommeaux. 2025. That doesn't sound right: Evaluating speech transcription quality in field linguistics corpora. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 627–635, Vienna, Austria. Association for Computational Linguistics.

Andy T. Liu, Shu-wen Yang, Po-Han Chi, and Hung-yi Huang. 2022. Discrete and efficient audio representation learning with self-supervised learning. In *Proc. ICASSP*, pages 6367–6371.

Lirong Liu, Yajie Zhou, Haoran Lin, Wenwen Zhao, and He Bu. 2021. Investigating ctc alignment stability for end-to-end speech recognition. In *Proc. Interspeech*, pages 2341–2345.

Vincent P. Martin, C. Beaumard, Jean-Luc Rouas, and Yaru Wu. 2024. Is automatic phoneme recognition suitable for speech analysis? temporal and performance evaluation of an automatic speech recognition model in spontaneous french. *Speech Prosody 2024*.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Proc. Interspeech*, pages 498–502.

Lukas Ondel, Martin Karafiát, and Lukáš Burget. 2016. Variational inference for acoustic unit discovery. In *Proc. SLTU*, pages 208–215.

Ankita Pasad, Chung-Ming Chien, Shane Settle, and Karen Livescu. 2024. What do self-supervised speech models know about words?

Florian Schiel. 1999. Automatic phonetic transcription of non-prompted speech. In *Proc. of the ICPhS*, pages 607–610.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R. Hershey, and Tsubasa Hayashi. 2017. Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253.

Qiantong Xu, Alexei Baevski, and Michael Auli. 2021. Simple and effective zero-shot cross-lingual phoneme recognition.

Jitong Yang, Shuo Chen, Yu Zhang, Sahar Ghannay, Jing Gao, Tara N Sainath, and Abdelrahman Mohamed. 2023. Torchaudio: Building blocks for reproducible and composable speech processing. *arXiv preprint arXiv:2306.12404*.

S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. 2006. *HTK: Hidden Markov Toolkit (Version 3.4)*. Cambridge University Engineering Department.

Albert Zeyer, Ralf Schluter, and Hermann Ney. 2021. Why does ctc result in peaky behavior? *ArXiv*, abs/2105.14849.

Yu-An Zhang, Ming Chen, Zheng Liu, and Xun Wang. 2020. Unsupervised learning for tts alignment. In *Proc. Interspeech*, pages 1803–1807.

10. Language Resource References

Chizzoni, Ilaria and Vietti, Alessandro. 2025. *Lost in Transcription: Towards Linguistically Informed Evaluation Metrics*. Associazione Italiana di Linguistica Computazionale.

Garofolo, J. S. and Lamel, L. F. and Fisher, W. M. and Fiscus, J. G. and Pallett, D. S. and Dahlgren, N. L. 1993. *DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM*. NIST.

On the Role of Encoder Depth: Pruning Whisper and LoRA Fine-Tuning in SLAM-ASR

Ganesh Pavan Kartikeya Bharadwaj Kolluri, Michael Kampouridis, Ravi Shekhar
School of Computer Science and Electronic Engineering, University of Essex
{karthik.kolluri, mkampo, r.shekhar}@essex.ac.uk

Abstract

Automatic speech recognition (ASR) has advanced rapidly in recent years, driven by large-scale pretrained models and end-to-end architectures such as SLAM-ASR. A key component of SLAM-ASR systems is the Whisper speech encoder, which provides robust acoustic representations. While model pruning has been explored for the full Whisper encoder–decoder architecture, its impact within the SLAM-ASR setting remains under-investigated. In this work, we analyze the effects of layer pruning in the Whisper encoder when used as the acoustic backbone of SLAM-ASR. We further examine the extent to which LoRA-based fine-tuning can recover performance degradation caused by pruning. Experiments conducted across three Whisper variants (Small, Medium, Large-v2), three languages representing distinct resource levels (Danish, Dutch, English), and over 200 training runs demonstrate that pruning two encoder layers causes only 2–4% WER degradation, and that combining this pruning with LoRA adaptation consistently outperforms the unpruned baseline while reducing total parameters by 7–14%. Moreover, our error analysis reveals that LoRA primarily compensates through the language model's linguistic priors, reducing total word errors by 11-21% for Dutch and English, with substitutions and deletions showing the largest reductions. However, for low-resource Danish, the reduction is smaller (4-7%), and LoRA introduces increased insertion errors, indicating that compensation effectiveness depends on the LLM's pre-existing language proficiency and available training data.

Keywords: Automatic Speech Recognition, Pruning, SpeechLLM, SLAM-ASR

1. Introduction

Recent advances in multimodal models have enabled automatic speech recognition systems (ASR) through architectures that connect pre-trained speech encoders to large language models (LLMs). SLAM-ASR (Ma et al., 2024; Zhang et al., 2026b,a) is the one such instance, which is a simpler approach and yet achieved comparative ASR performance: a frozen speech encoder extracts acoustic representations, and a small trainable projector maps these into the embedding space of a frozen LLM, which then generates transcriptions autoregressively. While this modular design is attractive for its simplicity, deploying such systems remains computationally expensive, as the full encoder must process every input audio regardless of whether all its layers contribute meaningfully to ASR performance.

Model compression through layer pruning has been widely studied for ASR. Kim et al. (2023) combined language-specific adapters with magnitude pruning to compress Whisper by up to 50%, still with competitive CER performance, while Gu et al. (2024) achieved 80% compression through one-shot unstructured pruning with minimal WER degradation. On the decoder side, Gandhi et al. (2023) distilled Whisper's decoder from 32 layers to 2 using knowledge distillation with large-scale pseudo-labeling, achieving 6 times faster inference within 1% WER of the original. However, these studies focused on traditional encoder-decoder ASR, where both the encoder and decoder are jointly trained. In the SLAM-ASR case, the encoder remains the same; the decoder is replaced by a frozen LLM that was never exposed to speech data and a lightweight projector that bridges the two components together. The effects of encoder layer pruning on this architecture remain unexplored. Layer-wise analyses suggest that upper encoder layers encode increasingly abstract linguistic features (Pasad et al., 2021), capabilities the LLM already possesses, raising the possibility that these layers are partially redundant. Separately, Low-Rank Adaptation (LoRA) (Hu et al., 2022) has been applied to LLM-based ASR for multilingual adaptation (Song et al., 2024; Fang et al., 2025; Concina et al., 2025) and low-resource improvement (Ma et al., 2024; Nagano et al., 2025; Fong et al., 2025; Burdisso et al., 2026), but no prior work has examined whether LoRA can compensate for performance lost through structured encoder layer pruning, or how data resource availability regulates these interactions.

This paper addresses these gaps through a systematic study of encoder layer pruning and LoRA compensation in SLAM-ASR [1]. Specifically, we focus on two research questions:

RQ1: *How does progressive encoder layer pruning affect ASR performance, and how does data resource availability modulate this effect?*

RQ2: *Can LoRA adaptation compensate for*

[1] The code is available at https://github.com/KarthikKolluriKB/SLAM-ASR-Encoder-Pruning-LoRA

Proceedings of SPEAKABLE @ LREC 2026, pages 183–193
11 May 2026.

pruning induced degradation?

To answer these questions, we evaluated across three languages representing distinct resource levels: Danish (4.2 hours, low-resource), Dutch (50 hours, medium-resource), and English (100 hours, high-resource), using three Whisper encoder variants (Small, Medium, Large-v2) (Radford et al., 2023) paired with Qwen2.5-3B LLM (Qwen et al., 2025). All three languages are supported by both the Whisper encoders and Qwen2.5-3B, ensuring that observed differences reflect data availability rather than missing language coverage. Our experiments cover pruning depths from the full encoder down to a single layer, with and without LoRA, spanning over 200 individual training runs.

Our main findings are as follows: First, removing the top one to two encoder layers results in only marginal WER degradation (with 2-4%) across all encoder scales and languages, supporting the hypothesis that upper encoder layers are partially redundant when an LLM handles linguistic processing. Beyond this arrangement, medium and high-resource languages degrade smoothly, while low-resource Danish exhibits erratic, non-monotonic behaviour. Second, combining modest pruning (two layers) with LoRA consistently outperforms the unpruned baseline while reducing total parameters by 7–14%, demonstrating that fewer encoder parameters paired with lightweight LLM adaptation can match or exceed full baseline model performance. Third, LoRA's compensatory benefit is not uniform across resource levels: it provides robust improvement for Dutch and English across all pruning depths, but is less effective for Danish, suggesting an interaction between data availability and adaptation capacity that needs to be further investigated. Additionally, error analysis reveals that LoRA compensates primarily through the LLM's linguistic priors, reducing word-level errors by 11–21% for Dutch and English, with substitution and deletion corrections showing the largest reductions, confirming a decoding-side compensation mechanism.

2. Related Works

This section reviews prior research in two areas relevant to our study: (1) layer pruning strategies for speech encoders, and (2) parameter-efficient fine-tuning for LLM-based speech systems.

2.1. Layer Pruning in Speech Encoders

Model compression through layer pruning has been extensively studied for ASR. Kim et al. (2023) introduced PEPSI (Parameter-Efficient Pruning and Adaptation for Speech Foundational Models), an adapt-and-prune framework for Whisper that combines language-specific adapters with iterative magnitude pruning. Their experiments on Common Voice showed that pruning can reduce model size by up to 50% while maintaining competitive CER across Korean and Malayalam. However, PEPSI employs *unstructured* pruning that removes individual weights rather than entire layers, resulting in sparse networks that require specialized hardware for efficient inference.

More recently, Gu et al. (2024) proposed SparseWAV, a one-shot unstructured pruning method for large speech foundation models, achieving up to 80% compression with minimal WER degradation. Irigoyen et al. (2025) revealed that certain encoder components actually *improve* when pruned, acting as implicit regularizers. Decoder self-attention at 50% sparsity achieved 2.38% absolute WER reduction on LibriSpeech test-other.

A parallel line of study focuses on *decoder* compression. Distil-Whisper (Gandhi et al., 2023) uses knowledge distillation (Hinton et al., 2015) with large-scale pseudo-labeling to compress the decoder from 32 layers to 2 while keeping the encoder frozen, achieving 6$\times$ faster inference within 1% WER of the original model. Similarly, BaldWhisper (Sy et al., 2025) targets low-resource deployment by merging decoder layer pairs in Whisper-base rather than removing them, achieving 2.15$\times$ faster inference with 48% size reduction while maintaining over 90% of baseline performance on Bambara speech data.

However, these studies focus on traditional encoder-decoder ASR, where Whisper's own decoder directly generates transcriptions. SLAM-ASR (Ma et al., 2024) presents a fundamentally different setting: the decoder is replaced by an LLM, and a lightweight projector that maps encoder outputs to the LLM embedding space. The effects of encoder layer pruning on this SLAM-ASR architecture remain unexplored. Here, the projector must learn to align potentially degraded encoder representations with a fixed, independently pre-trained LLM. Unlike traditional ASR, where encoder and decoder are jointly trained, SLAM-ASR offers no such flexibility: the LLM is frozen and was never exposed to speech, placing the entire burden of representation quality on the encoder.

2.2. Parameter-Efficient Fine-Tuning for LLM-based Speech Systems

Parameter-efficient fine-tuning (PEFT) methods (Houlsby et al., 2019) have become essential for adapting pre-trained large language models to specific downstream tasks without updating the entire model's parameters. Low-Rank Adaptation (LoRA) (Hu et al., 2022) has emerged as the dominant approach, decomposing weight updates into low-rank matrices that typically account for less than 1% of total parameters.

In LLM-based ASR, LoRA has been applied across multiple components. Song et al. (2024) proposed LoRA-Whisper, which incorporates the LoRA matrix into Whisper for multilingual ASR to effectively mitigate language interference, achieving 18.5% relative WER reduction for multilingual ASR and 23.0% for language expansion while using only 5% of trainable parameters compared to full fine-tuning. Building on this idea, Mu et al. (2025) proposed HDMoLE, which combines multiple LoRA experts through hierarchical routing and dynamic thresholds for multi-accent LLM-based ASR. Their method achieves comparable CER to full fine-tuning on multi-accent and standard Mandarin datasets while using only 9.6% of the trainable parameters.

Within LLM-based ASR models, the SLAM-ASR framework (Ma et al., 2024) showed that applying LoRA to LLM attention layers improves low-resource ASR, especially when combined with partial encoder fine-tuning (Tang et al., 2023; Wu et al., 2023). However, Kumar et al. (2025) observed that SLAM-ASR generalizes poorly across domains, highlighting the need for effective adaptation strategies in real-world deployment.

Despite this progress, no prior work has examined whether LoRA can compensate for information lost through structured encoder layer removal. While PEPSI (Kim et al., 2023) combines LoRA with unstructured weight pruning, the effects of removing entire encoder layers and whether LoRA can recover from such architectural changes in SLAM-ASR systems remain unexplored. It is also unknown whether LoRA's compensatory effectiveness depends on the amount of available training data.

3. Methodology

This study investigates two strategies for improving the parameter efficiency of SLAM-ASR systems: encoder layer pruning and LoRA adaptation of the LLM. Specifically, we ask whether upper encoder layers can be removed without significant performance loss, given that the LLM is already proficient in linguistic tasks, and whether LoRA can compensate for any degradation introduced by pruning. We evaluated these strategies across three languages representing distinct resource levels to examine how data availability interacts with both pruning robustness and LoRA effectiveness.

We follow the SLAM-ASR framework (Ma et al., 2024), which connects a frozen pre-trained speech encoder to a frozen large language model through a small trainable projector module. Only projector parameters are updated during training, and both

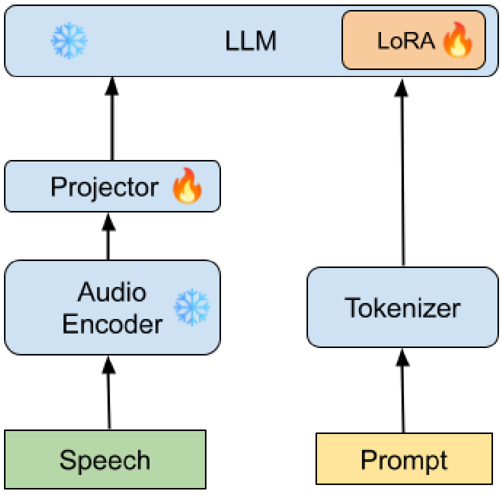


Figure 1: Overview of the SLAM-ASR architecture.

the encoder and LLM remain frozen. This modular design is crucial for our study because each component has a definite role; we can selectively modify the encoder through pruning and study how the system responds without disrupting other components. Figure 1 illustrates this architecture.

For encoder pruning, these previous studies on layer-wise analysis have shown that lower encoder layers primarily capture low-level acoustic features, while upper layers encode increasingly abstract linguistic information (Pasad et al., 2021). We hypothesize that in the SLAM-ASR architecture, these upper layers may be partially redundant, and that removing them offers a favourable efficiency–performance trade-off by delegating linguistic processing to the LLM. To verify this, we adopt a top-down pruning strategy: starting from the full encoder, we sequentially remove layers from the top, producing increasingly compressed configurations. For each configuration, we retrain only the projector from scratch, keeping both the pruned encoder and LLM frozen.

To investigate whether the LLM can be adapted to better handle degraded encoder representations, we apply Low-Rank Adaptation (LoRA) (Hu et al., 2022) to all attention projection matrices (Q, K, V, O) in the LLM. We hypothesize that, rather than recovering missing encoder information directly, LoRA enables the LLM to leverage its pre-trained linguistic knowledge: vocabulary, grammar, and contextual plausibility to adapt to the lost acoustic details. This creates a parameter trade-off: we remove millions of encoder parameters while adding far fewer LoRA parameters. We evaluate all combinations of pruning depths with and without LoRA, using Word Error Rate (WER) as the primary evaluation metric.

4. Experimental Setup

4.1. Datasets

We evaluate our approach on the Common Voice 22 corpus (Ardila et al., 2020), a crowdsourced multilingual speech dataset. To study how encoder pruning and LoRA compensation interact with training data availability, we select three languages that span an order-of-magnitude range in training data: Danish (4.2 hours), Dutch (54 hours), and English (100 hours). This selection is intentional; these three resource levels allow us to test whether pruning degradation patterns are consistent across data resource availability (RQ1) and whether LoRA compensation depends on sufficient fine-tuning data (RQ2). Dataset statistics are summarised in Table 1.

Table 1: Dataset statistics for each language from Common Voice 22.

Language	Split	Samples	Hours
Danish	Train	3,592	4.18
	Dev	2,511	3.21
	Test	2,684	3.45
Dutch	Train	43,458	54.15
	Dev	12,032	16.03
	Test	12,033	16.38
English	Train	58,140	100.00
	Dev	16,398	27.22
	Test	16,391	27.02

Preprocessing. Audio files are resampled to 16 kHz and converted to 80-channel log-Mel spectrograms following the Whisper preprocessing pipeline. We filter utterances to retain only those between 0.5 and 30 seconds in duration. Transcriptions are lowercased with punctuation removed, preserving apostrophes for contractions, following standard ASR preprocessing conventions.

4.2. Implementation Details

Model Architecture. We use Whisper (Radford et al., 2023), a transformer-based (Vaswani et al., 2017) speech encoder, as it is a widely adopted multilingual encoder available in multiple model sizes, making it well-suited for studying how pruning dynamics vary with encoder capacity. We pair it with Qwen2.5-3B (Qwen et al., 2025) as the LLM backbone, selected for its multilingual pre-training coverage, which is essential given that our experiments span three distinct languages. We experiment with three Whisper variants to study how pruning dynamics vary with model capacity. For all three Whisper variants: Small (12 layers), Medium (24 layers), and Large-v2 (32 layers), we evaluate configurations down to 2 layers in increments of 2.

Projector Architecture. We employ a two-layer MLP with concatenation-based temporal downsampling. The projector concatenates 5 consecutive frames (reducing the sequence length by a factor of 5), then applies two linear transformations with ReLU activation and a dropout rate of 0.1. The hidden dimension is 2048. We apply LayerNorm after the final projection to match the scale of LLM text embeddings.

LoRA Configuration. We apply Low-Rank Adaptation (Hu et al., 2022) to the query, key, value, and output projection matrices of all LLM attention layers. To prevent overfitting on smaller datasets, we use separate configurations for each resource level: $r{=}8$, $\alpha{=}16$ for Danish (low-resource) and $r{=}16$, $\alpha{=}32$ for Dutch/English (medium/high-resource). In preliminary experiments, using r=16 for Danish led to overfitting, motivating the reduced rank. All LoRA modules use a dropout of 0.1, adding approximately 0.8M (r=8) and 1.5M (r=16) trainable parameters, respectively.

4.3. Training Details

All models are trained using AdamW (Loshchilov and Hutter, 2019) with learning rate $1{\times}10^{-4}$, weight decay 0.01, and gradient clipping at 1.0. We employ a cosine learning rate schedule with linear warmup over the first 5% of training steps. Training uses mixed-precision (bfloat16) on a single NVIDIA A6000 GPU (48GB). Both the Whisper encoder and the Qwen2.5-3B base weights remain frozen; only the projector weights (and LoRA adapters when enabled) are updated.

4.4. Evaluation Details

We report Word Error Rate (WER) on the held-out test sets of each language. During evaluations, we use beam search with a beam size of 2. Text normalisation is applied to both predictions and references before evaluation: all text is lowercased, and punctuation is removed. For Danish, special characters (æ, ø, å) are preserved during normalisation to avoid penalising correct transcriptions. WER is computed using the `jiwer` library [2].

5. Results and Discussion

This section presents findings organized around two research questions: (1) How does progressive encoder layer pruning affect ASR performance, and how does data availability modulate this effect?

[2] https://github.com/jitsi/jiwer

(Section 5.1); and (2) Can LoRA adaptation compensate for pruning-induced degradation? (Section 5.2). We further provide an error analysis examining per-utterance trade-offs and the mechanism underlying LoRA compensation (Section 5.3)

5.1. RQ1: Effect of Encoder Layer Pruning

This section examines the impact of progressive removal of encoder layers on SLAM-ASR performance. By evaluating across three resource levels, we aim to understand how pruning depth interacts with data availability; specifically, whether resource-constrained settings exhibit different degradation patterns compared to data-rich conditions.

Our results broadly support the hypothesis that upper encoder layers are partially redundant in this architecture, given the LLM's existing linguistic capacity, as hypothesized in Section 3. Across all three Whisper variants (Small, Medium, Large-v2), removing the top one to two encoder layers results in only marginal WER degradation, with absolute increases remaining within 2–4% regardless of language (Table 2). We refer to this range as the safe pruning zone: the number of layers that can be removed before WER degrades beyond a practically meaningful threshold.

Beyond this zone, degradation diverges according to data resource availability. Dutch and English (medium and large resources) degrade smoothly and approximately monotonically as layers are removed. Danish (low resource) exhibits a qualitatively different pattern: rather than increasing monotonically, WER fluctuates erratically, spiking at certain depths and dropping at others. Notably, increasing the encoder capacity does not resolve this instability. Whisper-Large-v2, despite having nearly three times the layers of Whisper-Small, exhibits the same number of spikes for Danish, while Whisper-Medium shows the smoothest degradation curve among the three. Since Danish is the only language exhibiting this behaviour and also the language with the least training data, we hypothesize that limited data availability may be a contributing factor. However, further investigation, including controlled experiments with comparable amounts of training data across other languages, is needed to draw definitive conclusions.

5.2. RQ2: LoRA Compensation for Pruning

The previous section showed that shallow pruning is feasible, but deeper removal degrades performance. Here, we examine whether applying LoRA to the LLM can compensate for this degradation, and whether the resulting pruned+LoRA models can match or exceed unpruned baselines across data resource levels.

Our experimentation suggests that the answer is yes, within a shallow pruning regime, and that the compensation benefits generalize across encoder scales, though not equally across data resource levels. Table 3 summarizes what we term the sweet-spot configurations: the best-performing pruned+LoRA variant for each encoder scale, alongside the corresponding unpruned baseline and its LoRA-enhanced counterpart. Across all three Whisper variants, removing two encoder layers and applying LoRA to the LLM consistently outperforms the full unpruned baseline while reducing total parameter count by 7–14%. For Whisper-Small, the 10L+LoRA configuration achieves a 14.4% parameter reduction (from 88.15M to 75.5M) and improves WER on all three languages relative to the 12L baseline. The same pattern holds for Whisper-Medium (22L+LoRA, 7.6% reduction) and Whisper-Large-v2 (30L+LoRA, 5.9%), confirming that the finding is encoder, not scale-dependent.

Importantly, this improvement is not specific to pruned models. LoRA also improves the unpruned baseline encoder across all three languages (Table 2), indicating that it provides a general alignment benefit between the encoder and LLM, rather than simply patching information lost through layer removal. For Dutch and English, LoRA reduces WER across all pruning depths tested, from the full encoder to six layers. Even at aggressive pruning depths where base models produce severely degraded output, LoRA offers substantial recovery. For example, LoRA reduces Whisper-Large-v2 Danish WER from 130.99 (20L base) to 59.06 (20L+LoRA). However, for Danish, LoRA's compensation benefit is inconsistent at deeper pruning levels, which strengthens the finding from RQ1 that low-resource data conditions interact differently with both pruning and LoRA adaptation.

Since LoRA also improves the unpruned baseline (Table 2), a natural question is whether LoRA specifically compensates for pruning damage or simply provides a general performance boost. If LoRA acts as a general boost, then it should help the full model and the pruned model by roughly the same amount. However, if LoRA specifically compensates for the information lost through pruning, then it should help the pruned model more, because there is more to recover. To test this, we compute the WER reduction from LoRA in both settings: the difference between the full baseline and full baseline+LoRA, and the difference between the two-layer pruned and two-layer pruned+LoRA (Table 5). In seven of nine conditions, LoRA provides a larger WER reduction for the pruned model than for the full model. For example, in Whisper-Small Dutch, LoRA reduces the full model's WER by 2.58 but the pruned model's WER by 4.18, indicating

Table 2: Test set WER (%) across encoder layer pruning configurations for three Whisper scales. *Red.%*: reduction of parameters in percentage. *Base*: projector-only training; *+LoRA*: LoRA additionally applied to the LLM. Blue : full baseline WER. **Green** : 2 layer pruned+LoRA that outperforms the full baseline.

		Danish		Dutch		English	
Layers	Red.%	Base	+LoRA	Base	+LoRA	Base	+LoRA
Whisper-Small (88.15M base params)							
12 (full)	0%	47.41	41.94	17.98	15.40	21.84	17.87
10	16.1%	50.49	**47.23**	21.16	**16.98**	25.81	**20.64**
8	32.2%	71.22	51.20	27.25	21.04	33.15	24.40
6	48.2%	115.68	67.02	32.14	24.45	40.81	31.85
4	64.3%	82.56	72.60	45.62	31.69	74.32	41.73
2	80.4%	123.21	115.48	62.28	48.97	110.85	95.30
Whisper-Medium (307.24M base params)							
24 (full)	0%	39.00	38.88	12.74	10.92	16.33	14.13
22	8.2%	39.73	**37.91**	14.22	**11.84**	17.89	**15.82**
20	16.4%	43.18	41.71	15.73	13.44	18.29	16.54
18	24.6%	50.59	44.74	18.25	14.66	20.44	18.47
16	32.8%	49.72	49.20	20.58	15.76	23.15	20.22
14	41.0%	55.55	51.95	22.64	16.91	25.85	21.79
12	49.2%	59.15	59.21	27.61	20.04	29.44	25.38
10	57.4%	63.21	62.02	27.85	24.50	33.79	27.70
8	65.6%	69.62	66.03	38.16	27.89	43.79	34.03
6	73.8%	77.08	79.41	44.00	34.27	97.35	72.30
4	82.0%	105.16	102.63	53.40	43.46	111.71	77.48
2	90.2%	148.37	90.25	66.83	56.23	110.12	78.01
Whisper-Large-v2 (636.83M base params)							
32 (full)	0%	36.09	34.07	12.05	10.32	13.36	11.58
30	6.2%	38.36	**35.81**	13.61	**11.44**	14.87	**12.16**
28	12.4%	40.08	39.12	15.75	12.67	16.26	15.62
26	18.5%	48.28	44.75	17.15	14.03	18.32	17.08
24	24.7%	52.02	47.63	18.96	15.23	20.75	20.13
22	30.9%	52.44	56.24	21.31	16.71	22.31	21.45
20	37.1%	130.99	59.06	26.00	18.73	36.34	32.92
18	43.3%	73.87	61.64	28.92	22.06	48.38	39.86
16	49.4%	67.48	64.40	38.82	24.87	54.12	46.28
14	55.6%	77.97	73.77	45.21	29.81	62.45	59.82
12	61.8%	115.51	84.08	59.81	39.03	74.42	65.36
10	68.0%	104.79	87.67	59.60	43.42	84.06	74.17
8	74.2%	160.15	90.33	63.61	48.18	88.33	77.34
6	80.3%	120.43	99.58	90.60	51.14	98.28	87.84
4	86.5%	118.30	122.81	72.41	62.04	104.34	92.24
2	92.7%	123.52	132.03	83.75	74.39	108.47	96.20

that the degraded encoder representations leave more room for LoRA to recover. The two exceptions (Small Danish and Medium English) involve either low-resource instability or a negligible difference (2.20 vs 2.07). These results confirm that LoRA does not merely improve performance uniformly; it provides greater recovery where pruning has caused more damage. These findings motivate a deeper investigation into the specific error types that LoRA corrects, which we examine next.

5.3. Error Analysis

The aggregate WER results from Sections 5.1 and 5.2 summarise performance as a single number per test set. However, this can obscure important variation: a low average WER could result from uniform small improvements across all utterances, or from large improvements on some utterances offset by regressions on others. To investigate this, we analyse the results at three levels of granularity: we first measure how many individual utterances are affected by pruning and LoRA (Section 5.3.1), then

Table 3: Sweet-spot configurations (bold) per encoder scale, showing WER (%) on the test set. Net Δ is relative to the full baseline. Reported params use the $r{=}16$ LoRA config (Dutch/English); Danish uses $r{=}8$ (~0.8M vs. ~1.5M overhead).

Encoder	Configuration	Params	Net Δ	DA	NL	EN
Small	12L Baseline	88.15M	–	47.41	17.98	21.84
	12L + LoRA	89.65M	+1.5M	41.94	15.40	17.87
	10L + LoRA	**75.5M**	**–12.7M**	**47.23**	**16.98**	**20.64**
Medium	24L Baseline	307.24M	–	39.00	12.74	16.33
	24L + LoRA	309.0M	+1.8M	38.88	10.92	14.13
	22L + LoRA	**284.0M**	**–23.2M**	**37.91**	**11.84**	**15.82**
Large-v2	32L Baseline	636.83M	–	36.09	12.05	13.36
	32L + LoRA	638.3M	+1.5M	34.07	10.32	11.58
	30L + LoRA	**599.0M**	**–37.8M**	**35.81**	**11.44**	**12.16**

Table 4: Utterance-level degradation after removing two layers, with and without LoRA. Δ Recov.: reduction in percentage of degraded utterances after applying LoRA.

		% Utterances Degraded		
Encoder	**Lang**	**Base−2L**	**+LoRA**	**Δ Recov.**
Small	DA	42.7	38.1	−4.6
	NL	33.3	23.5	−9.8
	EN	35.9	23.9	−12.0
	DA	41.9	35.7	−6.2
Medium	NL	24.6	24.3	−0.3
	EN	25.3	22.7	−2.6
	DA	33.6	25.8	−7.8
Large-v2	NL	24.1	18.2	−5.9
	EN	22.8	15.4	−7.4

Table 5: WER reduction from applying LoRA to full baseline and two-layer pruned models, measured in percentage points. Bold indicates the larger reduction. In seven of nine conditions, LoRA provides a larger reduction for the pruned model.

		WER Reduction	
Encoder	**Lang**	**Full+LoRA**	**Pruned+LoRA**
Small	DA	**5.47**	3.26
	NL	2.58	**4.18**
	EN	3.97	**5.17**
Medium	DA	0.12	**1.82**
	NL	1.82	**2.38**
	EN	**2.20**	2.07
Large-v2	DA	2.02	**2.55**
	NL	1.73	**2.17**
	EN	1.78	**2.71**

examine what types of errors LoRA corrects (Section 5.3.2), and finally quantify the word-level error distribution to identify the underlying compensation mechanism (Section 5.3.3).

5.3.1. Utterance-Level Impact of Pruning and LoRA

To assess whether pruning and LoRA affect utterances uniformly, we perform a per-utterance comparison across three settings: the full unpruned baseline, the two-layer pruned model without LoRA, and the two-layer pruned model with LoRA. We use the two-layer pruned depth because it represents the best efficiency–performance trade-off identified in Section 5.2 (10L for Small, 22L for Medium, 30L for Large-v2). For each utterance in the test set, we compute WER under all three settings and compare the two pruned variants against the unpruned baseline: if an utterance's WER increased after pruning, it is counted as degraded; otherwise, it is counted as preserved or improved. Table 4 reports the percentage of utterances degraded by pruning alone (Base−2L), the percentage still degraded after applying LoRA (+LoRA), and the recovery difference (Δ Recov.) across all encoder variants and languages.

Pruning alone (Base-2L) degrades a substantial share of utterances, and this share is consistently higher for Danish (33.6-42.7%) than for Dutch (24.1-33.3%) or English (22.8-35.9%), consistent with the resource-level patterns from Section 5.1. Second, adding LoRA reduces the degradation rate in all nine conditions, meaning some previously degraded utterances are recovered. However, even after LoRA, 15-38% of utterances remain worse than the unpruned baseline, while the majority (62-85%) match or improve upon it. This reveals that the aggregate WER gains reported in Section 5.2 are not the result of uniform improvements; they are driven by strong recoveries on a subset of utterances that outweigh the remaining regressions.

5.3.2. Qualitative Error Patterns

The previous analysis shows how many utterances are affected; we now examine what kinds of errors LoRA actually corrects. With thousands of utterances across nine encoders and language configurations, manual inspection of every case is infeasible. We therefore apply two filters to isolate the most informative cases. First, we select utterances where the pruned model produces severe errors (WER > 0.8), ensuring that encoder degradation is substantial enough to produce identifiable error patterns rather than minor single-word differences. Second, we require that LoRA achieves full recovery (WER $= 0.0$) of the same utterances, which confirms that the correction is attributable to LoRA's adaptation of the LLM rather than to residual information preserved in the pruned encoder. Examples are drawn from all three encoder variants and all three languages. Table 6 presents representative cases grouped by error type.

Five common patterns we have identified: repetitive hallucination, where the pruned model generates looping output; idiom and fixed expression errors, where familiar phrases are distorted beyond recognition; named entity fragmentation; world knowledge substitution, where domain-specific terms are replaced by plausible but incorrect alternatives; and phonetic confusion. Particularly, all five categories involve errors recoverable through linguistic rather than acoustic information.

5.3.3. Mechanism Interpretation

The qualitative patterns above suggest that LoRA compensates through the LLM's linguistic knowledge. To test this quantitatively, we analyse word-level errors that LoRA corrects. Standard ASR error analysis decomposes word errors into three categories: substitutions, insertions, and deletions. For each encoder variant and language, we count these error types for both the pruned model and the pruned+LoRA model under the two-layer pruned configuration. Table 7 reports the percentage change in each error type after applying LoRA. The Tot column reports the percentage change in total word errors (substitutions + insertions + deletions combined); because each error type contributes differently to the total count, Tot reflects the weighted combination rather than a simple average of the three individual percentages.

Since LoRA adapts only the LLM's attention matrices (Q, K, V, O) while the pruned encoder remains frozen, all compensation operates on the decoding side. We would therefore expect error reductions for languages where the LLM has strong linguistic grounding, as the LLM can leverage its vocabulary, grammar, and contextual knowledge to correct degraded encoder representations. The results confirm this (Table 7): across all three encoder scales, LoRA reduces total word errors for English and Dutch in all six conditions, with substitutions and deletions showing consistent reductions. Danish follows a different pattern: while substitutions and deletions decrease, insertion errors increase across all three encoder scales, indicating that when the LLM lacks sufficient linguistic grounding for a language, LoRA can introduce spurious tokens rather than recover missing information. This language-dependent pattern reinforces that LoRA's effectiveness is tied to the LLM's pre-existing proficiency in each language rather than to the acoustic properties of the signal.

6. Conclusion

This paper presented a systematic study of encoder layer pruning and LoRA compensation in SLAM-ASR, evaluated across three Whisper encoder variants (Small, Medium, Large-v2) and three languages representing distinct resource levels (Danish, Dutch, and English). We find that removing the top one to two encoder layers results in only marginal WER degradation (within 2–4%) across all encoder scales and languages, supporting the hypothesis that upper encoder layers are partially redundant when an LLM handles downstream linguistic processing. Beyond this shallow pruning regime, medium and high resource languages degrade smoothly, while low-resource Danish often exhibits non-monotonic instability. Combining two-layer pruning with LoRA consistently outperforms the unpruned baseline while reducing total parameters by 7–14%. Error analysis reveals that LoRA compensates primarily through substitution and deletion corrections, leveraging the LLM's linguistic knowledge rather than repairing acoustic information, though this benefit is less consistent for low-resource Danish. Future work should explore alternative pruning strategies beyond top-down removal, encoder-side LoRA adaptation, validation across different system architectures, and controlled experiments to disentangle the effects of data availability from pre-trained representation quality.

7. Ethical Considerations and Limitations

In this work, we have used publicly available models, architectures, and datasets and have not collected any sensitive/private data. The ultimate goal of our study is to contribute to analyzing the effect of speech encoder pruning on LLM-based ASR.

Table 6: Representative error patterns where the pruned model produces severe errors (WER > 0.8) but LoRA achieves full recovery (WER $= 0.0$). Examples sampled across all three encoder variants.

Pattern	Lang	Reference	Pruned Hypothesis	+LoRA Hypothesis
Hallucination	DA	der kogte gryden over	da kom gud og gik op	der kogte gryden over
Named entity	DA	charlot var altid inde hos fru simonin nu	chrisel indelte altså inden for sin sinde	charlot var altid inde hos fru simonin nu
Semantic drift	DA	lille soldat du skal være vor konge…	livet skal du have og du skal have den dig selv…	lille soldat du skal være vor konge…
Repetition	NL	…heel interessant en hebben ons zeer geholpen…	zeer grotendeels [×28]	…heel interessant en hebben ons zeer geholpen…
Named entity	NL	de veiligheidssituatie is sindsdien rampzalig verslechterd	de veiligheid ziet de waarde van een dienst…	de veiligheidssituatie is sindsdien rampzalig verslechterd
Compound word	NL	de luchtvaart-maatschappijen	de luchtvaart maatschappijen	de luchtvaart-maatschappijen
Idiom	EN	out of sight out of mind	at a site at a light	out of sight out of mind
Phonetic	EN	heaven forbid	have fun for bit	heaven forbid
World knowledge	EN	it is isoelectronic to benzene	it is said to be a traditional venetian dish	it is isoelectronic to benzene

Table 7: Word-level error change (%) after applying LoRA to the two-layer pruned configurations. *Sub*: substitutions; *Ins*: insertions; *Del*: deletions; *Tot*: total word errors. **Bold**: error increase.

Encoder	Lang	Sub	Ins	Del	Tot
Small	DA	–7.7	**+11.9**	–10.7	–5.7
	NL	–17.5	–37.4	–15.3	–20.4
	EN	–19.8	–9.8	–37.4	–20.6
Medium	DA	–4.9	**+18.4**	–22.9	–4.4
	NL	–13.7	–34.5	–11.4	–16.7
	EN	–10.6	**+0.5**	–30.3	–11.6
Large-v2	DA	–8.6	**+10.8**	–11.7	–6.6
	NL	–15.3	–14.9	–13.5	–15.0
	EN	–17.4	–7.1	–35.6	–18.2

Due to the use of all public details, we don't see any immediate ethical issue.

Our work investigates the pruning of the speech encoder in the SLAM-ASR using the Whisper and Qwen models. We have carefully limited our analysis to three languages of different resource levels. While this choice allows us to conduct careful analysis, we acknowledge that expanding the range of models and datasets could provide additional insights. Our evaluation uses a single system architecture (Whisper encoder, ConcatLinear projector, and Qwen2.5-3B); different LLM backbones, projector designs, or encoder families, however, we believe the finding will hold. Our pruning strategy follows a top-down approach motivated by prior evidence of upper-layer redundancy; however, redundancy patterns may vary across layers, and exploring other removal strategies could reveal additional compression opportunities.

8. Acknowledgments

This work was supported by a Knowledge Transfer Partnership (KTP) project (project number 10131983) funded by UKRI through Innovate UK, in collaboration with Hivedome. RS was supported by the ELOQUENCE project (grant number 101070558) funded by the UKRI and the European Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the UKRI, European Union, or European Commission-EU. Neither the European Union nor the granting authority can be held responsible for them.

9. Bibliographical References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the twelfth language resources and evaluation conference*, pages 4218–4222.

Sergio Burdisso, Esaú Villatoro-Tello, Andrés Carofilis, Shashi Kumar, Kadri Hacioglu, Srikanth R. Madikeri, Pradeep Rangappa, Manjunath K. E, Petr Motlícek, Shankar Venkatesan, and Andreas Stolcke. 2026. Text-only adaptation in LLM-based ASR through text denoising. *arXiv preprint arXiv:2601.20900*.

Lorenzo Concina, Jordi Luque, Alessio Brutti, Marco Matassoni, and Yuchen Zhang. 2025. The Eloquence team submission for task 1 of MLC-SLM challenge. In *Workshop on Multilingual Conversational Speech Language Model (MLC-SLM)*, pages 50–53.

Yangui Fang, Jing Peng, Xu Li, Yu Xi, Chengwei Zhang, Guohui Zhong, and Kai Yu. 2025. Low-resource domain adaptation for speech LLMs via text-only fine-tuning. *arXiv preprint arXiv:2506.05671*.

Seraphina Fong, Marco Matassoni, and Alessio Brutti. 2025. Speech LLMs in Low-Resource Scenarios: Data Volume Requirements and the Impact of Pretraining on High-Resource Languages. *arXiv preprint arXiv:2508.05149*.

Sanchit Gandhi, Patrick Von Platen, and Alexander M Rush. 2023. Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling. *arXiv preprint arXiv:2311.00430*.

Tianteng Gu, Bei Liu, Hang Shao, and Yanmin Qian. 2024. Sparsewav: Fast and accurate one-shot unstructured pruning for large speech foundation models. *Interspeech 2024*, pages 4498–4502.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Julian Irigoyen, Arthur Söhler, and Andreas Søeborg Kirkedal. 2025. Pruning as Regularization: Sensitivity-Aware One-Shot Pruning in ASR. *arXiv preprint arXiv:2511.08092*.

Hyeon Soo Kim, Chung Hyeon Cho, Hyejin Won, and Kyung Ho Park. 2023. Adapt and prune strategy for multilingual speech foundational model on low-resourced languages. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 85–94.

Shashi Kumar, Iuliia Thorbecke, Sergio Burdisso, Esaú Villatoro-Tello, Manjunath KE, Kadri Hacioğlu, Pradeep Rangappa, Petr Motlicek, Aravind Ganapathiraju, and Andreas Stolcke. 2025. Performance evaluation of slam-asr: The good, the bad, the ugly, and the way forward. In *2025 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5. IEEE.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Ziyang Ma, Guanrou Yang, Yifan Yang, Zhifu Gao, Jiaming Wang, Zhihao Du, Fan Yu, Qian Chen, Siqi Zheng, Shiliang Zhang, and Xie Chen. 2024. An Embarrassingly Simple Approach for LLM with Strong ASR Capacity. *arXiv preprint arXiv:2402.08846*.

Bingshen Mu, Kun Wei, Qijie Shao, Yong Xu, and Lei Xie. 2025. Hdmole: Mixture of lora experts with hierarchical routing and dynamic thresholds for fine-tuning llm-based asr models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Tohru Nagano, Gakuto Kurata, Samuel Thomas, Hong-Kwang J Kuo, Daniel Bolanos, Hyun Jung, and George Saon. 2025. LLM based text generation for improved low-resource speech recognition models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921. IEEE.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Zheshu Song, Jianheng Zhuo, Yifan Yang, Ziyang Ma, Shixiong Zhang, and Xie Chen. 2024. LoRA-Whisper: Parameter-Efficient and Extensible Multilingual ASR. In *25th Annual Conference of the International Speech Communication Association, Interspeech 2024, Kos, Greece, September 1-5, 2024*. ISCA.

Yaya Sy, Christophe Cerisara, and Irina Illina. 2025. BaldWhisper: Faster Whisper with Head Shearing and Layer Merging. *arXiv preprint arXiv:2510.08599*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jian Wu, Yashesh Gaur, Zhuo Chen, Long Zhou, Yimeng Zhu, Tianrui Wang, Jinyu Li, Shujie Liu, Bo Ren, Linquan Liu, and Yu Wu. 2023. On Decoder-Only Architecture For Speech-to-Text and Large Language Model Integration. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.

Yuchen Zhang, Haralambos Mouratidis, and Ravi Shekhar. 2026a. Speak in Context: Multilingual ASR with Speech–Context Alignment via Contrastive Learning. In *Proceedings of the Fifteenth biennial Language Resources and Evaluation Conference (LREC 2026)*.

Yuchen Zhang, Ravi Shekhar, and Haralambos Mouratidis. 2026b. Language family matters: Evaluating SpeechLLMs across linguistic boundaries. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, pages 487–499, Rabat, Morocco. Association for Computational Linguistics.

TaLK-Corpus: A Regionally Diverse Evaluation Set for Sri Lankan Tamil Speech

Adsajan Thillainathan[1] Nishanthini Kanthakumar[1] Nivethiga Rasan[2] Kengatharaiyer Sarveswaran[1]

[1]Department of Computer Science, University of Jaffna
[2]Department of Linguistics and English, University of Jaffna
{2021sp146, 2020csc026, rnivethiga, sarves}@univ.jfn.ac.lk

Abstract

This paper introduces the TaLK Corpus, the first speech benchmark corpus for Sri Lankan Tamil Automatic Speech Recognition (ASR) covering speech from 22 administrative districts of Sri Lanka. The corpus contains 1 hour and 33 minutes of speech from 22 native speakers (one per district) and includes rich metadata on demographics, location history, recording conditions, and domain information, along with transcriptions in Tamil script and the International Phonetic Alphabet (IPA). Standardised preprocessing (16 kHz mono WAV format) and segmentation using Silero Voice Activity Detection (VAD) resulted in 1,214 utterances. All recordings were manually transcribed by trained linguists, and MD5-based file naming used to ensure data integrity and consistency. TaLK corpus enables district-wise benchmarking of ASR systems and supports dialect-sensitive evaluation. We establish baseline results for multilingual models (Whisper Large-V3 and Facebook's MMS) in zero-shot settings. The evaluation reveals substantial performance disparities across districts, highlighting the impact of regional phonological variation in low-resource Sri Lankan Tamil. Although Whisper Large-V3 outperforms MMS overall, it shows considerable variability, with mean Word Error Rates ranging from 0.672 to 0.903 across districts. These findings demonstrate strong regional effects even within a single model. By releasing TaLK-Corpus under the CC-BY-NC 4.0 licence, we aim to support dialect-robust ASR research and foster inclusive speech technologies for Sri Lankan Tamil-speaking communities.

Keywords: Sri Lankan Tamil, Speech Corpus, Dialectal Variation, Low-Resource Languages, ASR Benchmark Dataset

1. Introduction

Speech processing has become a critical area of research as voice interfaces power an increasing number of applications, from virtual assistants to real-time transcription and accessibility tools (Rakotomalala et al., 2021). In this context, Automatic Speech Recognition (ASR), which is the conversion of spoken language into text, and Text-to-Speech (TTS) have gained prominence. Transformer-based models such as Whisper (Radford et al., 2023), achieve high accuracy in controlled settings for high-resource languages. However, persistent challenges limit their robustness in real-world deployment, including noise and environmental variability, speaker and accent variability, multilingualism and code-switching, data scarcity in low-resource languages, and spontaneous or conversational speech (Ahlawat et al., 2025).

Dialectal processing also presents challenges for ASR. Dialectal variations introduce phonetic mismatches (e.g., vowel shifts, consonant changes), lexical differences, prosodic variations, and morphosyntactic deviations from standard training data (Palivela et al., 2025). Models trained primarily on mainstream varieties exhibit systematic biases, leading to substantially higher Word Error Rates (WER) for regional or minority dialects compared to standard forms. In low-resource languages, these issues are exacerbated by limited dialect-specific data, lack of standardized orthographies, and the absence of diverse training corpora, hindering the development of inclusive, robust systems (Dhasmana et al., 2026).

Benchmark datasets play a central role in advancing Automatic Speech Recognition (ASR) by enabling standardized and reproducible evaluation of model robustness across linguistic and acoustic variability.

Existing Tamil ASR benchmarks, (Mozilla Foundation, 2019; OpenSLR, 2019; Bharathi et al., 2022, 2024), mainly focus on Indian Tamil or demographic diversity without structured geographic coverage in Sri Lanka. While Sri Lankan resources like EmoTa (Thevakumar et al., 2025) exist, they target emotion recognition rather than ASR and do not provide district-level representation, leaving a gap for geographically structured Sri Lankan Tamil speech datasets.

To address this gap, we introduce a regionally di-

"TaLK" is inspired by the IETF BCP 47 language tag ("ta-LK") for Sri Lankan Tamil, which corresponds to "ta_LK" in the Unicode CLDR locale format.

Proceedings of SPEAKABLE @ LREC 2026, pages 194–200
11 May 2026.

verse evaluation set for Sri Lankan Tamil speech, called TaLK-Corpus[2], covering speakers from all 22 districts. The dataset is explicitly designed as a benchmarking resource, enabling reproducible district-wise evaluation of ASR systems. By incorporating Tamil script, Roman transliteration, and broad IPA annotation, TaLK supports conventional WER-based comparison, establishing a standardized benchmark for Sri Lankan Tamil dialectal ASR research.

This paper makes two main contributions. First, we present a carefully manually curated speech corpus for Sri Lankan Tamil, comprising 1 hour and 33 minutes of high-quality, geographically representative audio–text pairs with verbatim transcriptions. Second, we establish a zero-shot performance baseline for this dialect by evaluating two widely used multilingual ASR models—Whisper V3 and Facebook's Massively Multilingual Speech (MMS).

2. Background and Motivation

Tamil is one of the world's oldest living classical languages, belonging to the Dravidian family, with a rich literary tradition spanning over 2,000 years (Newbigin, 2019). It is spoken by more than 86 million people (Zeidan, 2020), primarily in Tamil Nadu (India), Sri Lanka, Singapore, Malaysia, and diaspora communities, and holds official status in several regions. Despite its cultural and demographic significance, Tamil is a low-resource language in modern NLP and speech technologies (Sarveswaran et al., 2021), suffering from limited large-scale annotated digital corpora compared to high-resource languages like English. Key linguistic characteristics include agglutinative morphology, a rich phonemic inventory (notably retroflex consonants and vowel length distinctions) (Jain and Bhowmick, 2025), syllable-timed prosody (Thinakaran et al., 2025), diglossia (distinct literary and colloquial varieties), and frequent code-switching (Prasanna and Arora, 2024), especially with English.

Sri Lankan Tamil speech exhibits notable regional variation due to historical settlement patterns, prolonged contact with Sinhala and English, and population movements (Unjum et al., 2026; Yasmini, 2017). These factors lead to differences in pronunciation, accent, vocabulary, and prosody across districts and even within them. While Tamil ASR has advanced through corpora dominated by Indian Tamil (e.g., Common Voice Tamil, IISc-MILE) (Mozilla Foundation, 2019; OpenSLR, 2019), Sri Lankan-specific resources remain scarce and often task-specific, such as EmoTa for emotion recognition (Thevakumar et al., 2025), limiting the development of dialect-robust models for Sri Lankan contexts.

More importantly, there is no benchmark dataset available for Sri Lankan Tamil to evaluate the performance of ASR systems on this language variety. To address this gap, we present Version 1 of the TaLK speech corpus, a district-level Sri Lankan Tamil evaluation corpus comprising 22 speakers (one per district), with over one hour of total speech and rich metadata, annotated in the Tamil script.

3. Related Work

In Tamil ASR, several datasets have supported model development, including resources from Mozilla Common Voice Tamil (Mozilla Foundation, 2019) and OpenSLR (OpenSLR, 2019). While these corpora provide valuable training and testing material, they primarily represent Indian Tamil and largely consist of read or crowd-sourced speech, with limited structured metadata for dialect-aware benchmarking

The LT-EDI shared tasks on Speech Recognition for vulnerable Individuals in Tamil introduced evaluation datasets targeting elderly and transgender speakers in naturalistic settings (Bharathi et al., 2022, 2024; Nishanth et al., 2025). These initiatives represent an important step toward inclusive ASR benchmarking in Tamil by focusing on demographic diversity from the Indian region. Further, their design centers on speaker-group variability rather than geographically structured dialect variation, and they do not provide district-level dialectal coverage.

In the Sri Lankan context, EmoTa: A Tamil Emotional Speech Dataset (Thevakumar et al., 2025) contributed a valuable speech resource for emotion recognition research. Although important for affective computing, EmoTa is not structured as an ASR benchmark and does not support systematic evaluation of regional dialect robustness.

More broadly, dialect-focused benchmarks in multilingual settings such as IndicVoices-R (Javed et al., 2024)-have demonstrated the necessity of geographically diverse evaluation sets to measure accent and dialect sensitivity in ASR systems. These works highlight that models trained predominantly on standardised language varieties may exhibit systematic degradation when evaluated on regional speech.

4. Data Collection and Corpus Design

The dataset was collected from native Sri Lankan Tamil speakers across all 22 districts (out of 25)

[2] https://github.com/LTG-UoJ/TaLK-Corpus-Public

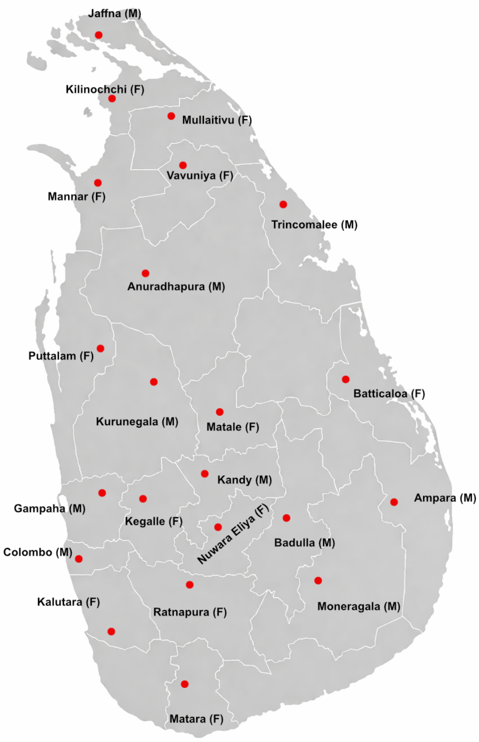


Figure 1: District-level distribution of speakers

of Sri Lanka to ensure comprehensive geographical coverage and dialectal diversity. The objective was to capture diverse variations in pronunciation, accent, and speaking style across different regions, making the corpus suitable for evaluation and diagnostic purposes in Automatic Speech Recognition (ASR) systems.

4.1. Speaker Selection

The corpus consists of 22 speakers (12 female and 10 male) representing diverse demographic backgrounds. The age distribution includes one speaker below 20 years, the majority between 20 and 60 years, and one speaker above 60 years. Participants were drawn from varied educational and occupational backgrounds to enhance representativeness for ASR development. All participants had previously sat for the GCE Ordinary Level (O/L) national examination in which mother tongue (Tamil) is an important subject.

4.2. Recording Protocol

Speech recordings were captured using the Sony PCM-A10 digital recorder and the DJI Mic Receiver connected to a mobile device. The DJI Mic features an Intelligent Noise Cancelling function, which was enabled to reduce background noise when recording with a smartphone. Recordings were conducted under controlled indoor environments with minimal background noise.

Participants were asked to speak spontaneously across multiple everyday domains, including school life, education, university experience, trips and travel, friends, family, places, festivals, funeral events, self-introduction, life experiences, and other common daily topics. These domains were selected to ensure lexical diversity and to reflect natural conversational contexts in Sri Lankan Tamil.

All recordings were standardised for ASR compatibility by converting the audio into mono-channel WAV format with a 16 kHz sampling rate. This configuration provides an optimal balance between speech quality and computational efficiency for modern speech recognition systems. The total duration of the collected speech data is approximately 1 hour and 33 minutes.

4.3. Metadata Schema

All metadata for the **TaLK** corpus is stored in a single CSV file, where each row corresponds to an audio file and its associated transcription. Speaker-level information is repeated for every utterance to ensure each row is self-contained. The metadata includes the district of birth, gender, age band, job, religion, and domain(theme). Utterance-level information includes the duration of the recording, the filename of the audio file, and its transcript in Tamil script. In addition, IPA annotations are captured for each utterance to enable further phonological analysis. For the Jaffna district, IPA transcriptions were generated using the tool called ThamizhiIPA-Trans (Mahaganapathy et al., 2026) and validated by a linguist, while for other districts, annotations were manually created by the linguist. Table 1 provides a sample Metadata Entry from the TaLK Corpus. All audio files are stored in a dedicated folder, with filenames matching the entries in the CSV, allowing direct linking between metadata, audio, and transcripts. This format ensures compatibility with standard ASR processing pipelines and facilitates filtering or analysis by speaker attributes or recording characteristics.

4.4. File Naming Strategy and Anonymisation

An MD5-based hashing strategy was used to generate file names. Demographic attributes (Current Residence, Birth Place District, Age, and Gender) were combined into a unique input string and processed through MD5 so that in the future errors in file-naming can be easily tracked. A sequential

Field	Value
File name	0d2d12...9f850a_001.wav
Transcription	அங்க இருக்கும்போது...
IPA	*aŋga ɨrukkɨmbōðu...*
Birth Place (District)	Kilinochchi
Age Band	20-60
Gender	Feminine
Job	Student
Religion	Hindu
Theme	School
Duration	0:00:06

Table 1: Sample Metadata Entry from the TaLK Corpus

three-digit index was appended to maintain uniqueness across multiple files.

4.5. Ethics and Consent

All participants provided informed consent before recording, and their identities were anonymised to protect privacy. The data is licensed for research use, with restrictions preventing commercial exploitation, ensuring ethical compliance throughout data collection and usage.

5. Audio segmentation and Annotation

Speech segmentation was performed using Silero VAD v4.0[3] with a threshold of 0.5, a minimum speech duration of 250 ms, a minimum silence duration of 100 ms, a 512-sample window size, and 30 ms padding to produce clean speech segments. This process resulted in 1,214 utterances. Transcription was carried out manually by two trained linguists in dialect-preserving Tamil script and aligned to the segments. Text preprocessing and normalisation included Unicode normalisation, removal of punctuation, special characters, and extra or leading/trailing spaces. Quality control was conducted through manual checks, with double annotation considered for future validation.

6. Model benchmarking

The primary task supported by the TaLK corpus is the benchmarking of ASR systems for Sri Lankan Tamil. In addition, the detailed geographic and demographic metadata enables optional auxiliary tasks such as dialect or regional classification. In this section we report the performance of two widely used models Whisper V3 and MMS.

[3] https://github.com/aosfatos/silero-vad-v4

Model	WER	CER
Whisper v3	0.807	0.425
facebook/mms-1b-all	0.845	0.342

Table 2: ASR evaluation results on the TaLK dataset.

Model	TaLK	IndicVoices-Ta
Whisper v3	0.807	0.784
MMS	0.845	0.754

Table 3: WER comparison between TaLK (ours) and IndicVoices Tamil.

6.1. Metrics

Word Error Rate (WER) and Character Error Rate (CER) were used as the primary evaluation metrics. Scoring follows Tamil-aware tokenisation that respects the language's agglutinative morphology and script conventions. Punctuation marks and special characters are ignored during evaluation, and numerals are normalised consistently.

7. Baseline results

We present evaluation results for multilingual ASR models on Sri Lankan Tamil in zero-shot settings, assessing their performance without any Sri Lankan Tamil specific adaptation.

All experiments use the same audio preprocessing (mono, 16 kHz), normalisation, and scoring pipeline to ensure fair comparison across districts. We compute WER on the Tamil script transcripts and additionally report WER on Roman transliterations to separate orthographic effects from acoustic errors. Decoding parameters and any language-model usage are held constant across districts and documented for reproducibility.

8. Results and Analysis

We evaluate on 1,214 utterances from 1 hour 33 minutes TaLK-Corpus. Table 2 reports WER and CER for the two zero-shot baselines. Whisper V3 outperforms Facebook's MMS under this evaluation. For comparison, we also report the performance of Whisper V3 and Facebook's MMS from the IndicVoices (Javed et al., 2024) study in Table 3, which shows that the models perform poorly for Sri Lankan Tamil compare to the Indian Tamil (or the major variety included in language models).

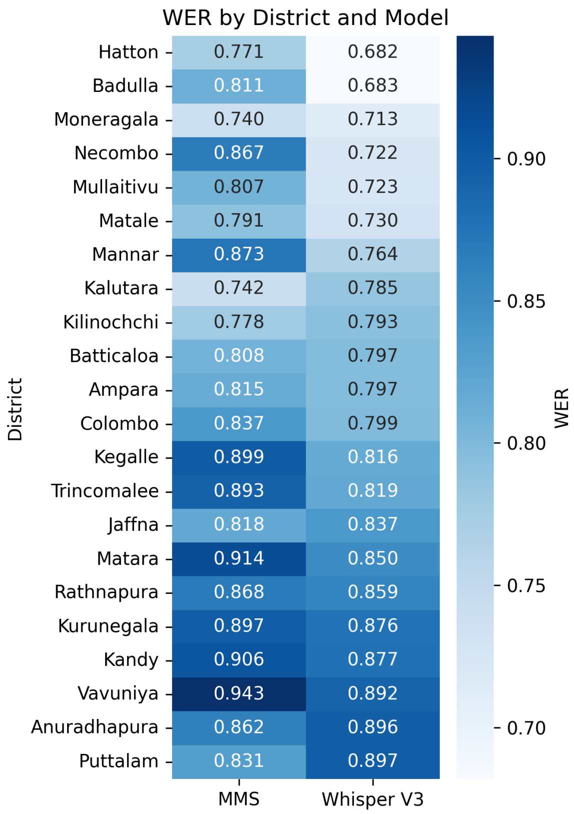


Figure 2: WER per district for MMS and Whisper V3.

Figure 2 provides a per-district breakdown. Whisper V3 shows notable variability across districts, with mean WER ranging from 0.672 to 0.903, indicating strong regional effects even under a single model. Hatton, Moneragala, and Badulla districts show comparatively lower WER for both MMS and Whisper V3. This is likely because the speech in these districts is closer to Indian Tamil, which the models were primarily trained on, so the ASR systems can recognise it more accurately than the more distinct Sri Lankan dialects in other districts.

Error analysis. Whisper V3 frequently code-switches to English for numerals and educational terms, which increases WER against Tamil-script references, for instance, as shown below, the ASR output contains code-switching and Roman script. Interestingly, the case markers present in the Tamil text are also reflected in the romanised text, but in Tamil script. For instance, கு *ku* (DAT marker) follows "lecture" in the ASR output below.

> **Ground truth:** இப்ப இங்க வந்து பார்த்தோம்னா இங்கிலீஷ் மீடியம் ரொம்பவே டி°ஃபிகல்ட் தான் என்ன சம்ரைம்ஸ் லெக்சருக்கு போனா தூங்கிருன்ட்டு வர மாதிரி தான் இருக்கும் ஏன் சொன்னா

> **ASR output:** இங்க வந்து பார்த்தாம்னா, **English Medium**, ரொம்பவே **Difficult** தான் **Sometimes**, **lecture** கு போனா, தூங்கிருந்து வர மாது தான் இருக்கும் ஏன் சொன்னாம்?.

These patterns highlight the importance of handling code-switching and numerals consistently in evaluation. The IPA annotation layer in TaLK enables future analysis of dialectal phonological patterns beyond orthographic WER.

9. Availability, Licensing, and Reproducibility

The TaLK dataset has been made publicly available to support research in dialect-aware Sri Lankan Tamil speech processing. It includes clean, segmented 16 kHz mono WAV audio files and Tamil script and IPA transcriptions. The dataset is released under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license. This permits sharing, adaptation, and use for non-commercial research, specifically for evaluation, and academic purposes, provided appropriate credit is given to the creators (citation required).

10. Conclusion

We introduce the TaLK Corpus, a Sri Lankan Tamil speech corpus comprising district-level data for benchmarking ASR applications for Sri Lankan Tamil, along with rich metadata that can be used for in-depth dialectal studies and phoneme-level benchmarking. The benchmark enables reproducible, dialect-aware evaluation of ASR systems across all districts, highlighting meaningful regional variability in zero-shot performance. Our baseline results and error analysis show that current multilingual ASR systems struggle with dialectal variation and code-switching, reinforcing the need for geographically diverse evaluation and richer Sri Lankan Tamil resources. TaLK-Corpus provides a concrete starting point for more inclusive ASR research and for future expansions in speaker coverage and recording conditions. The TaLK corpus repository is available at `https://github.com/LTG-UoJ/TaLK-Corpus-Public`.

11. Limitations and Future Work

The current dataset is small to support model training; therefore, model evaluation was conducted in a zero-shot setting. In addition, the present version of TaLK includes only one speaker per district. Consequently, it is not yet possible to fully disentangle district-level dialect effects from speaker-specific characteristics such as speaking rate, articulation style, or recording variability. The reported district-wise WER differences should therefore be interpreted as preliminary and indicative rather than definitive evidence of systematic dialect-level model performance differences.

WER evaluation is further affected by orthographic or spacing differences in the manually created ground truth. For example, when the model predicts "இருக்கும் போது" as two tokens while the ground truth is "இருக்கும்போது" is one token, the WER metric counts this as an error despite the prediction being linguistically correct. Such discrepancies are inherent to human annotation and can slightly inflate reported error rates, especially in low-resource languages with flexible orthographic conventions. Therefore, WER should be interpreted as indicative of overall performance rather than exact linguistic correctness.

This work is part of a broader initiative to develop a Sri Lankan Tamil speech corpus. We are currently expanding the dataset by including multiple speakers per district with balanced demographic representation, increasing the total number of recording hours through longer and more natural conversations, and incorporating diverse recording conditions (e.g., varying noise levels and channels). In addition to direct speech recordings, we plan to compile speech data from publicly available sources, such as YouTube, with appropriate consent and ethical compliance, to further enhance district-level coverage.

12. Acknowledgements

This research is part of the Sri Lankan Tamil Corpus (TaLK Corpus) project[4] at the University of Jaffna and is supported by a Google Research Scholar Award to Kengatharaiyer Sarveswaran. The authors thank Ms Sumirtha Karunakaran for her assistance with data collection. We also thank all participants and collaborators who contributed to the development of the corpus.

References

Harsh Ahlawat, Naveen Aggarwal, and Deepti Gupta. 2025. Automatic Speech Recognition: A survey of deep learning techniques and approaches. *International Journal of Cognitive Computing in Engineering*, 6:201–237.

B. Bharathi, Bharathi Raja Chakravarthi, et al. 2022. Findings of the Shared Task on Speech Recognition for Vulnerable Individuals in Tamil. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*.

B. Bharathi, Bharathi Raja Chakravarthi, et al. 2024. Overview of the Third Shared Task on Speech Recognition for Vulnerable Individuals in Tamil. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*.

Akriti Dhasmana, Aarohi Srivastava, and David Chiang. 2026. Dialect Matters: Cross-Lingual ASR Transfer for Low-Resource Indic Language Varieties. *arXiv preprint arXiv:2601.04373*.

P. Jain and A. Bhowmick. 2025. Comparative performance analysis of end-to-end ASR models on Indo-Aryan and Dravidian languages within India's linguistic landscape. *Journal of Audio, Speech, and Music Processing*, 10(2025).

Tahir Javed, Janki Nawale, Eldho George, Sakshi Joshi, Kaushal Bhogale, Deovrat Mehendale, Ishvinder Sethi, Aparna Ananthanarayanan, Hafsah Faquih, Pratiti Palit, Sneha Ravishankar, Saranya Sukumaran, Tripura Panchagnula, Sunjay Murali, Kunal Gandhi, Ambujavalli R, Manickam M, C Vaijayanthi, Krishnan Karunganni, Pratyush Kumar, and Mitesh Khapra. 2024. IndicVoices: Towards building an inclusive multilingual speech dataset for Indian languages. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10740–10782, Bangkok, Thailand. Association for Computational Linguistics.

Ahrane Mahaganapathy, Sumirtha Karunakaran, Kavitha Navakulan, and Kengatharaiyer Sarveswaran. 2026. Bridging dialectal variation: A phonetic transcription tool for Tamil. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 234–241, Rabat, Morocco. Association for Computational Linguistics.

Mozilla Foundation. 2019. Mozilla Common Voice. `https://commonvoice.mozilla.org`.

JILL CR Newbigin. 2019. Evolution of Tamil Language: A Historical Study. *Journal of Emerging Technologies and Innovative Research (JETIR)*, 6(3):115–123. © 2019 JETIR1903O19.

S. Nishanth, Shruthi Rengarajan, Burugu Rahul, and G. Jyothish Lal. 2025. NSR@LT-EDI-2025: Automatic Speech Recognition in Tamil. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*.

OpenSLR. 2019. OpenSLR: Free Speech and Language Resources. `http://www.openslr.org`.

[4] `https://sites.google.com/univ.jfn.ac.lk/talkcorpus/home`

Hemant Palivela, Meera Narvekar, David Asirvatham, Shashi Bhushan, Vinay Rishiwal, and Udit Agarwal. 2025. Code-switching asr for low-resource indic languages: A hindi-marathi case study. *IEEE Access*, 13:9171–9198.

Kabilan Prasanna and Aryaman Arora. 2024. Irumozhi: Automatically classifying diglossia in Tamil. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3096–3103.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Francis Rakotomalala, Hasindraibe Randriatsarafara, Hajalalaina Aimé Richard, and Ravonimanantsoa Ndaohialy Manda Vy. 2021. Voice User Interface: Literature Review, Challenges and Future Directions. *SYSTEM THEORY, CONTROL AND COMPUTING JOURNAL*, 1:65–89.

Kengatharaiyer Sarveswaran, Gihan Dias, and Miriam Butt. 2021. ThamizhiMorph: A morphological parser for the Tamil language. *Machine Translation*, 35(1):37–70.

Jubeerathan Thevakumar, Luxshan Thavarasa, Thanikan Sivatheepan, Sajeev Kugarajah, and Uthayasanker Thayasivam. 2025. EmoTa: A Tamil Emotional Speech Dataset. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 193–201, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Preethi Thinakaran, Malarvizhi Muthuramalingam, Anushiya Rachel Gladston, P Vijayalakshmi, Hema A Murthy, T Nagarajan, et al. 2025. SIToBI–A Speech Prosody Annotation Tool for Indian Languages. *arXiv preprint arXiv:2502.09661*.

Naveed Unjum, Stephanie Evert, Kengatharaiyer Sarveswaran, Ruvan Weerasinghe, and Nevidu Jayatilleke. 2026. Llms for low-resource languages: A survey.

P. Yasmini. 2017. The contrast between jaffna tamil and upcountry tamil: A dialectological study. In *The Third International Conference on Linguistics in Sri Lanka, ICLSL 2017*, page 143. Department of Linguistics, University of Kelaniya, Sri Lanka. Conference proceedings.

A. Zeidan. 2020. Languages by Total Number of Speakers. https://www.britannica.com/topic/languages-by-total-number-of-speakers-2228881. Accessed: 2024-12-29.

Author Index