JUNLP@LT-EDI-2025: Efficient Low-Rank Adaptation of Whisper for Inclusive Tamil Speech Recognition Targeting Vulnerable Populations

Priyobroto Acharya[1], **Soham Chaudhuri**[2], **Sayan Das**[3],**Dipanjan Saha**[4], **Dipankar Das**[5]
[1]Dept. of Power Engineering, Jadavpur University, Kolkata, India
[2]Dept. of Electrical Engineering, Jadavpur University, Kolkata, India
[3,4,5]Dept. of CSE, Jadavpur University, Kolkata, India
{ priyobrotoacharya98, sohamchaudhuri.12.a.38,sayan.das200216,sahadipanjan6, dipankar.dipnil2005}@gmail.com

Abstract

Speech recognition has received extensive research attention in recent years. It becomes much more challenging when the speaker's age, gender and other factors introduce variations in the speech. In this work, we propose a fine-tuned automatic speech recognition model derived from OpenAI's whisper-large-v2. Though we experimented with both Whisper-large and Wav2vec2-XLSR-large, the reduced WER of whisper-large proved to be a superior model. We secured **4**th rank in the LT-EDI-2025 shared task. Our implementation details and code are available at our **`GitHub repository`**[1].

1 Introduction

Automatic Speech Recognition (ASR) has transformed the way humans interact with machines by enabling devices to understand spoken language. It plays a crucial role in enhancing accessibility for individuals with disabilities, such as the elderly and those with hearing or speech impairments(Yu and Deng, 2017; Malik et al., 2021). By allowing voice-based interaction, ASR improves ease of communication and overall quality of life for these groups.

While ASR systems have achieved impressive accuracy in languages like English, low-resource languages such as Tamil still face challenges (Ramesh and Gupta, 2021). Tamil, spoken by millions across Tamil Nadu, Sri Lanka, and Singapore, is linguistically rich and features numerous regional dialects, making speech recognition particularly complex. These challenges are amplified when recognizing speech from vulnerable populations, such as those with dysarthria or slurring (Christensen, 2013).

In this work, we focus on building an inclusive Tamil ASR system by fine-tuning the Whisper model (vasista22/whisper-tamil-large-v2), known for its strong multilingual performance (Radford et al., 2022). To make the fine-tuning process efficient, we use Low-Rank Adaptation (LoRA), which reduces the computational burden while maintaining high accuracy (Hu et al., 2021). Our training dataset includes Tamil speech samples from diverse dialects and speakers with impairments. The fine-tuned model achieves a Word Error Rate (WER) of **38.42%**, demonstrating significant improvement and the potential of Whisper models in developing accessible ASR systems for underrepresented languages.

2 Related Work

Automatic speech recognition (ASR) has evolved from hybrid Hidden Markov Model-Gaussian Mixture Model (HMM-GMM)(Xuan et al., 2001) frameworks to end-to-end deep learning systems. Early systems leveraged HMMs for temporal modeling and DNNs for acoustic feature extraction, achieving significant accuracy improvements over traditional methods. Transitioning to architectures like LSTMs and transformers enabled better sequential context capture, with models like Conformer and ContextNet integrating convolutional and self-attention mechanisms for spectral and global dependencies(Prabhavalkar et al., 2021). Self-supervised learning paradigms, such as wav2vec 2.0, further advanced low-resource ASR by leveraging unlabeled data for robust feature learning(Mainzinger and Levow, 2024)(Kheddara et al., 2024).

Recent efforts focus on domain-specific challenges, including elderly and vulnerable populations as well as low-resource speech recognition. (B et al., 2022)(Bartelds et al., 2023)presented findings from a shared task on Tamil ASR for vulnerable individuals, emphasizing the difficulty of recognizing atypical speech patterns in elderly

[1]`https://github.com/Priyobroto98/ASR-Tamil-LTEDI-2025`

and impaired speakers. Their work demonstrated the utility of HMM-DNN hybrid systems(Wang et al., 2019) and end-to-end models alongside data augmentation and transfer learning to improve robustness. In a follow-up shared task, (B et al., 2025) expanded the dataset and evaluated multilingual models (e.g., XLS-R, Whisper), showing that fine-tuning, domain adaptation, and acoustic normalization techniques effectively addressed speech variations and noise in low-resource settings. Similar advances include acoustic model adaptation using age-specific corpora like EARS and VOTE400, which reduce word error rates (WER) by 25% for elderly speech by mitigating spectral and prosodic variations. For low-resource languages, techniques like self-training and text-to-speech augmentation improve WER by 20–25%, as demonstrated for Gronings and Mvskoke. Transformer-based streaming architectures, employing time-restricted attention, balance latency and accuracy, while hybrid HMM-DNN systems remain relevant for stable frame-level processing. Despite progress, challenges persist in dataset diversity, real-time adaptation, and computational efficiency for edge deployment.

3 Dataset Description and Analysis

The dataset focuses on addressing the challenges faced by vulnerable groups, specifically elderly individuals and transgender people in Tamil-speaking communities, where elderly individuals often encounter difficulties using digital tools in essential locations like banks, hospitals, and administrative offices, where speech-based systems could significantly ease their interactions (Gales et al., 2019; Liu and Lutters, 2021). Similarly, transgender individuals, frequently deprived of primary education due to societal prejudice, rely heavily on speech as their primary mode of communication (Pandey and Mishra, 2019; Bose et al., 2019). By capturing the spontaneous speech patterns of these groups, the dataset aims to facilitate the development of inclusive and accessible ASR systems that cater to their unique linguistic needs and daily life challenges (Albanie et al., 2020; Srinivasan et al., 2023).

The dataset contains **908 samples** totaling nearly 5 hours of speech. We have split the entire corpus into training (894 samples, 4.87 hours), validation (9 samples, 0.05 hours), and test sets (5 samples, 0.03 hours) for tracking the performance metrics at different stages of model development. In addition to this, we were provided with **2 hours** of high-quality audio speech data, which will be used for testing purposes after successfully training our best model and following best practices.

Set	Samples	Duration (hours)	Avg Duration (seconds)	Avg Text Length (chars)
Training	894	4.87	19.61	212
Validation	9	0.05	20.00	256
Test	5	0.03	20.00	229

Table 1: Dataset Statistics and Composition

We conduct **spectrogram analysis(Khodzhaev, 2024)** on the speech dataset to characterize the time-varying frequency properties of the audio signals. In figure-1 the analysis confirms that all samples exhibit dominant speech energy below **4 kHz**, with clearly observable formant structures.

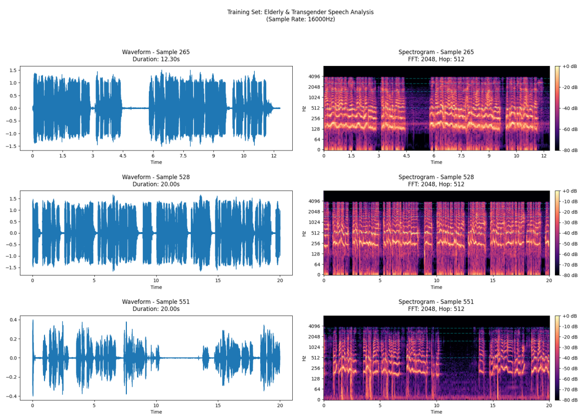

Figure 1: Representative spectrograms illustrating dominant speech energy and formant structures.

The overall spectral clarity and low background noise across all samples suggest high-quality recordings. These observations not only confirm the suitability of the data for further speech processing tasks—such as automatic speech recognition or speaker profiling (Nagrani et al., 2017; Yu et al., 2021), but also highlight the diversity in speaking styles and potential demographic differences among the speakers (Narayanan and Georgiou; **?**). Such variability is crucial for developing robust and inclusive speech systems that generalize well across different populations.

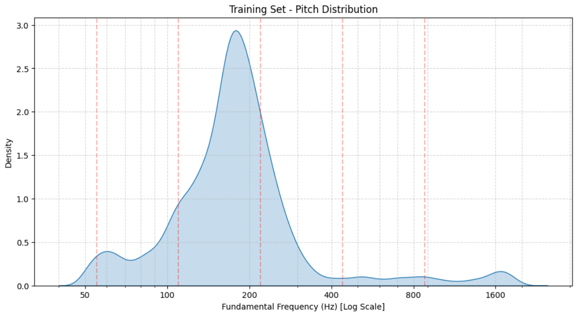

Figure 2: Pitch Distribution

In figure-2 the pitch distribution(Deruty et al., 2025) graph reveals a clear multimodal pattern, with a **dominant peak** near **200 Hz** and **secondary peaks** around **100 Hz** and at higher frequencies, indicating demographic diversity. The use of a logarithmic x-axis reflects the perceptual nature of pitch. Variations in peak heights highlight gender imbalance, which may introduce bias in ASR performance toward dominant voice types.

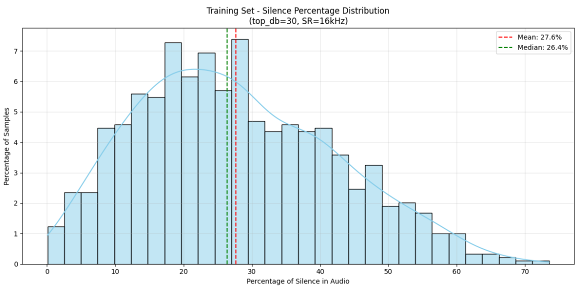

Figure 3: Silence Percentage Distribution

The dataset exhibits a bell-shaped silence distribution (jin Shim et al., 2024) (**mean 27.6%, median 26.4%**) with a right skew, where most samples contain **10–50%** silence (peaking at 25–30%) under a 30 dB/16 kHz detection threshold (refer Figure 3). This aligns with natural speech patterns, where pauses constitute approximately one-quarter of spoken content (Gold and Morgan, 2000), informing ASR design for effective endpoint detection and robustness (Ramírez et al., 2007). The balanced silence distribution facilitates training on realistic speech rhythms and timing structures (Jurafsky and Martin, 2000), improving temporal generalization in deployment scenarios.

From the analysis of temporal features (Figure 4), we found the audio dataset exhibits high-quality temporal features with segmented speech (amplitude ± 1.5 units) and precise silence intervals, evidenced by RMS energy drops to zero and spectral rolloff between 500–3500 Hz. Stable

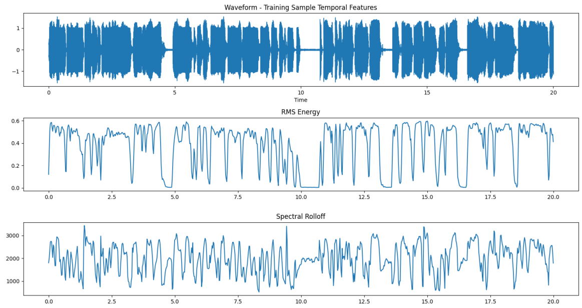

Figure 4: Audio Training Sample Temporal Features

RMS levels ($\sim$**0.4–0.5**) during speech segments indicate consistent articulation, while rolloff variations (**1000–3000 Hz**) reflect phonetic diversity, demonstrating complementary temporal-spectral features (waveform, energy, rolloff) that reveal controlled recording conditions ideal for training robust speech models requiring precise acoustic characterization (Rabiner and Schafer, 1978; Tolonen and Karjalainen, 2000; Purwins et al., 2019; Zhang et al., 2021).

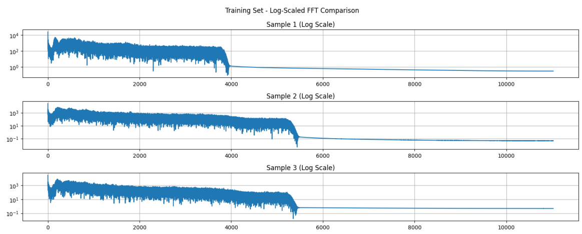

Figure 5: Log-Scaled FFT Comparison in Training Dataset

The **log-scaled FFT analysis** of the training dataset reveals concentrated spectral energy (**10^1–10^4** magnitude) in lower frequencies (**0–4000 bins**) with a sharp roll-off at **4000–5000 bins** across samples, indicating bandwidth-limited audio rich in harmonic content (refer Figure 5). Consistent noise floors (10^{-1}–10^0 magnitude) and spectral homogeneity suggest uniform recording/post-processing conditions, while the preserved harmonic structures and logarithmic energy distribution (aligning with auditory perception) highlight key perceptual features of speech signals (Choi et al., 2018; Deller et al., 1993; Verhelst and Roelands, 2000; Purwins et al., 2019).

4 Methodology and Implementation Details

In this study, speech recognition was performed using two pre-trained state-of-the-art models, Whisper and XLSR. Both models were trained on the Tamil corpus, and the best results were submitted for the competition.

The Whisper model (Radford et al., 2023) is a pre-trained automatic speech recognition (ASR) model trained on **680,000 hours** of multilingual and multitask supervised data sourced from the web. In our work, we have utilized **`vasista22/whisper-tamil-large-v2`**[2], which is a fine-tuned version of **`openai/whisper-large-v2`**[3] on the Tamil data available from multiple publicly available ASR corpora. This transformer-based encoder-decoder model processes log-Mel spectrograms through convolutional layers in the encoder and generates text autoregressively in the decoder. The model was further fine-tuned on a Tamil corpus of the given training dataset, providing a robust baseline for Tamil speech recognition.

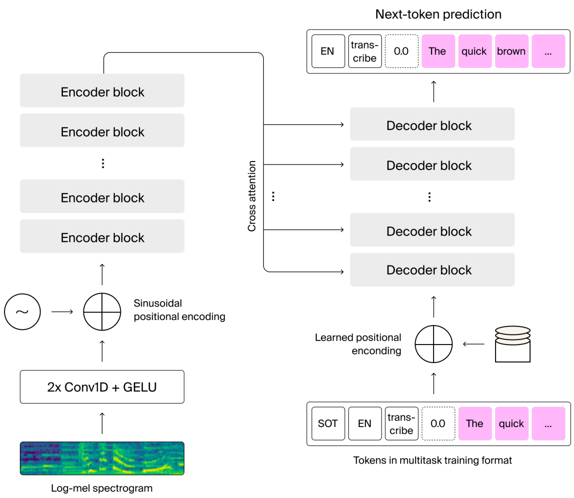

Figure 6: Whisper Model Architecture (https://openai.com/index/whisper/)

To adapt the 1.59-billion-parameter Whisper model efficiently, we utilize ***Low-Rank Adaptation (LoRA) (Hu et al., 2021)*** and ***Dynamic Rank Adaptation (DoRA) (Liu et al., 2024)***. These techniques freeze pre-trained weights and inject trainable low-rank matrices into specific transformer submodules, reducing computational overhead while preserving model performance (Xu et al., 2023).

LoRA decomposes weight updates (ΔW) into two low-rank matrices $\mathbf{A}$ and $\mathbf{B}$, where $\Delta W = \mathbf{BA}$. For a weight matrix $W \in R^{d\times k}$, the adapted weights become:

$$\begin{aligned} W' &= W + \Delta W \\ &= W + \mathbf{B}\cdot\mathbf{A}, \quad \mathbf{B} \in R^{d\times r}, \quad \mathbf{A} \in R^{r\times k} \end{aligned}$$

where $r \ll \min(d, k)$ is the rank of adaptation. This reduces trainable parameters from $\mathcal{O}(dk)$ to $\mathcal{O}(r(d + k))$.

We apply LoRA to the query, key, value, and output projection layers of each transformer block. To ensure stable training, weight scaling is used:

$$\Delta W = \alpha \cdot \frac{\mathbf{BA}}{r} \tag{1}$$

where α is a scaling factor (typically $\alpha \in [8, 32]$), introduced to stabilize updates for small r.

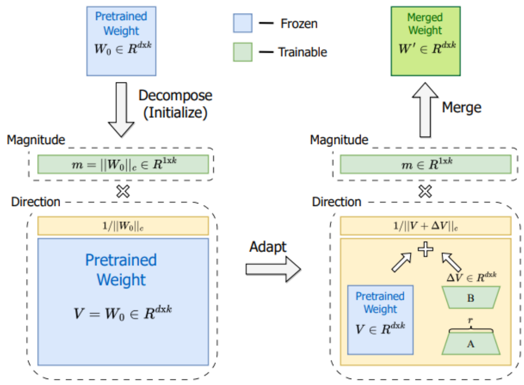

Figure 7: An overview of our proposed DoRA, which decomposes the pre-trained weight into magnitude and direction components for fine-tuning, especially with LoRA to efficiently update the direction component. Note that || · ||c denotes the vector-wise norm of a matrix across each column vector

DoRA extends LoRA by dynamically adjusting the rank r during training (Liu et al., 2024). It decomposes weights into magnitude (m) and direction ($\mathbf{V}$) components:

$$W = m \cdot \frac{\mathbf{V}}{\|\mathbf{V}\|_F} \tag{2}$$

where $\|\mathbf{V}\|_F$ is the Frobenius norm. During backpropagation, the gradient flows primarily through the direction $\mathbf{V}$, enabling more expressive parameterization even at low ranks.

Quantization to 8-bit precision was implemented using:

$$\mathbf{W}_{\text{int8}} = \text{quantize}\left(\frac{\mathbf{W} - \mu_{\mathbf{W}}}{\sigma_{\mathbf{W}}}\right)$$

where:

[2]https://huggingface.co/vasista22/whisper-tamil-large-v2
[3]https://huggingface.co/openai/whisper-large-v2

- $\mathbf{W}$ is the original full-precision weight matrix or tensor.
- $\mu_{\mathbf{W}}$ is the mean of the weight tensor $\mathbf{W}$, used for centering.
- $\sigma_{\mathbf{W}}$ is the standard deviation or scale factor of $\mathbf{W}$, used for normalization.
- $\mathbf{W}_{\text{int8}}$ is the quantized 8-bit integer representation of the normalized weights.
- $\hat{\mathbf{W}}$ is the dequantized approximation of the original weights in floating point.
- quantize$(\cdot)$ maps a real-valued input to discrete 8-bit integer levels (usually in the range $[-128, 127]$).

followed by dequantization:

$$\hat{\mathbf{W}} = \sigma_{\mathbf{W}} \cdot \mathbf{W}_{\text{int8}} + \mu_{\mathbf{W}}$$

Training employed mixed-precision arithmetic (FP16) with the **AdamW** optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-6}$), a learning rate of 10^{-5} with 50 warmup steps, and gradient accumulation over 2 steps. Only **2.99%** of parameters (47.5M out of 1.59B) were trainable through selective application of LoRA to the query, key, and value projection layers.

During implementation, a comprehensive data preprocessing pipeline was constructed using WhisperProcessor components, which extract audio features with a sampling rate of **16kHz** and prepare corresponding text transcriptions for supervised training. We have used a custom `DataCollatorSpeechSeq2SeqWithPadding` that effectively handles variable-length audio inputs and properly masks padding tokens in labels with -100 to be ignored during loss calculation. The combined use of 8-bit quantization, LoRA, and DoRA reduced memory requirements by 4 times compared to full-precision fine-tuning and achieved a **97%** reduction in trainable parameters without significant accuracy degradation, demonstrating the efficacy of parameter-efficient methods (Dettmers et al., 2023) for large-scale ASR (Radford et al., 2023) adaptation.

On the other hand, we fine-tuned the pretrained **anuragshas/wav2vec2-xlsr-53-tamil**[4] checkpoint with the Hugging Face `Trainer` API. The model is

[4]https://huggingface.co/anuragshas/wav2vec2-xlsr-53-tamil

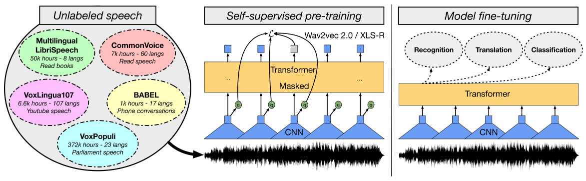

Figure 8: Fine-tuning XLSR for Tamil ASR with Transformers. (`https://huggingface.co/blog/fine-tune-xlsr-wav2vec2`)

a `Wav2Vec2ForCTC` type model (Conneau et al., 2021) and was fine-tuned with full-scale fine-tuning, without layer freezing or modifications. Connectionist Temporal Classification (CTC) loss was used during training and performance was tracked with Word Error Rate (WER) and Character Error Rate (CER). Mixed precision training was activated with `fp16=true`, and the best model was chosen based on the minimum WER on the evaluation set. Gradient accumulation with an accumulation step of 2 was used to stabilize training and mimic larger batch sizes.

5 Result and Discussion

Submissions to the Shared Task on Speech Recognition for Vulnerable Individuals in Tamil were evaluated using the **Word Error Rate (WER)** between the ASR hypotheses and the reference human transcriptions for the evaluation set (Morris et al., 2004).

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

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 transcriptions.

During the fine-tuning phase, a close watch was kept on the WER and **Character Error Rate (CER)** of both models, which were trained for the same number of epochs (Hori et al., 2017).

Model	Val. Loss	WER(%)	CER(%)
whisper-tamil-large-v2	0.540	69.4	26.1
wav2vec2-large-xlsr-53-tamil	1.727	94.0	44.2

Table 2: ASR Model Performance Comparison

We compared both the models' WER and CER. Since the `whisper-tamil-large-v2` model demonstrated significantly lower WER and CER than the `wav2vec2-large-xlsr-53-tamil` model, we selected it for generating transcriptions for the test dataset and submitted those results for final evaluation.

Team Name	WER	Rank
CrewX	31.9	1
NSR	34.85	2
Victory	34.93	3
JUNLP	38.42	4
SSNCSE	42.3	5

Table 3: Team-wise WER and Rank

We achieved a WER of **38.42** on the test dataset, which helped us secure the **4$^{\text{th}}$** rank in the shared task. This performance demonstrates the robustness of parameter-efficient fine-tuning strategies for multilingual ASR tasks on low-resource and demographically sensitive datasets (Hsu et al., 2021).

6 Limitations

Despite its contributions, this work has several limitations. The dataset's limited size and dialectal diversity may hinder generalization, particularly for underrepresented Tamil accents (Addanki et al., 2022). Computational constraints restricted the exploration of more complex architectures and large-scale training (Gaido et al., 2021). Evaluation primarily relied on WER, which may not fully reflect real-world intelligibility or user-centric performance, especially for vulnerable populations (Falk and Chan, 2007; Meng et al., 2021). The model's performance varied across regional pronunciations, suggesting a need for more balanced data. Additionally, the absence of human-centered evaluations, such as user studies or error analysis on critical phrases, limits insights into practical usability (Amershi et al., 2019). Resource limitations also prevented extensive hyperparameter tuning and ablation studies. Broader metrics, including semantic accuracy and user satisfaction, could better assess assistive utility (Baker et al., 2020). Finally, ethical considerations, such as bias mitigation and inclusivity in data collection, were not thoroughly examined (Hovy and Prabhumoye, 2021). Addressing these gaps in future work could enhance robustness and fairness in Tamil speech recognition.

7 Future Scope

To overcome these limitations and extend the impact of this study, several avenues for future work are proposed. Expanding the dataset to include speakers from a wide range of demographics and regions, as well as recording audio in diverse environmental conditions, could enhance the model's robustness and adaptability (Ko et al., 2017; Besacier et al., 2014). Incorporating advanced architectures and exploring multilingual frameworks may further improve performance (Pratap et al., 2020; Conneau et al., 2021). Real-world deployment possibilities, such as live transcription services and language learning tools for vulnerable groups, offer practical applications of this research (Albanie et al., 2020; Srinivasan et al., 2023). Collaborations with local communities and organizations to co-develop datasets and validate findings can ensure inclusivity and greater acceptance of the model in real-world scenarios (Bender et al., 2021).

8 Conclusion

This work presents JUNLP's efficient approach to building an inclusive Tamil Automatic Speech Recognition (ASR) system for vulnerable populations, including elderly and transgender speakers. Using parameter-efficient fine-tuning (PEFT) methods Low-Rank Adaptation (LoRA) and Dynamic Rank Adaptation (DoRA), we adapted the multilingual Whisper-large-v2 model for low-resource Tamil speech with demographic variation. Our model achieved a Word Error Rate (WER) of 38.42% on the LT-EDI-2025 evaluation set, securing 4th place. By freezing Whisper's 1.59B pretrained weights and injecting low-rank matrices, we reduced trainable parameters by 97% (47.5M) and memory usage by 4 times, enabling fine-tuning on limited hardware. DoRA's decomposition improved expressiveness, and 8-bit quantization with mixed-precision training stabilized optimization. Trained on 908 speech samples (5 hours) reflecting dialectal diversity, the model showed promise in inclusive ASR. Limitations include dataset size, regional bias, and reliance on WER. Future directions include expanding diverse corpora and integrating user-centered evaluations. This study affirms PEFT-enhanced Whisper models as viable for equitable ASR in Tamil.

References

Kartik Addanki, John J. Godfrey, and Sanjeev Khudanpur. 2022. Acce: A benchmark for evaluating asr robustness to accent variations. In *Proceedings of Interspeech*.

Samuel Albanie, Gül Varol, Liliane Momeni, and 1 others. 2020. Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues. *European Conference on Computer Vision (ECCV)*.

Saleema Amershi, Daniel Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, and Penny Collisson. 2019. Guidelines for human-ai interaction. In *CHI Conference on Human Factors in Computing Systems*.

Bharathi B, Bharathi Raja Chakravarthi, Subalalitha Cn, Sripriya N, Arunaggiri Pandian, and Swetha Valli. 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*, pages 339–345, Dublin, Ireland. Association for Computational Linguistics.

Bharathi B, Bharathi Raja Chakravarthi, Sripirya N, Rajeswari Natarajan, Rajalakshmi R, Suhasini S, and Swetha Valli. 2025. Overview of the fourth shared task on speech recognition for vulnerable individuals in tamil. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion*, Naples. Association for Computational Linguistics.

Ryan Baker, Yi Zhang, Zach Traylor, Mohit Doss, and Kristen Shinohara. 2020. Evaluating the effectiveness of speech recognition for individuals with speech impairments. In *ASSETS*.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making more of little data: Improving low-resource automatic speech recognition using data augmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 747–766, Toronto, Canada. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of FAccT*, pages 610–623.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

Debasmita Bose, Naveena Karusala, and Denzil Ferreira Chattopadhyay. 2019. Voice as agency: Gender identity in voice user interfaces. In *ACM CHI Conference on Human Factors in Computing Systems (CHI)*.

Keunwoo Choi, György Fazekas, Mark Sandler, and Kyunghyun Cho. 2018. A comparison of audio signal preprocessing methods for deep neural networks on music tagging. In *Proceedings of the 26th European Signal Processing Conference (EUSIPCO)*, pages 1870–1874. EURASIP.

Heidi Christensen. 2013. Automatic speech recognition for disordered speech. In *Handbook of Speech Communication*, pages 549–566. Walter de Gruyter.

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.

John R Deller, John H L Hansen, and John G Proakis. 1993. *Discrete-time Processing of Speech Signals*. Macmillan Publishing Company.

Emmanuel Deruty, Luc Leroy, Yann Macé, and David Meredith. 2025. Methods for pitch analysis in contemporary popular music: Highlighting pitch uncertainty in primaal's commercial works. *arXiv preprint arXiv:2502.08131*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Tiago H. Falk and William Y. Chan. 2007. Performance measures for voice-controlled interfaces with limited training data. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1169–1179.

Lorenzo Gaido, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2021. A survey on computational limitations and challenges in end-to-end speech recognition. *Computer Speech & Language*, 67:101178.

Mark Gales, Kate Knill, and Phil Woodland. 2019. Speech technology for health care: Opportunities and challenges. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6825–6829.

Ben Gold and Nelson Morgan. 2000. *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. John Wiley & Sons.

Takaaki Hori, Shinji Watanabe, Yu Zhang, and William Chan. 2017. Advances in joint ctc-attention based end-to-end speech recognition with a deep cnn encoder and rnn-lm. In *Proc. Interspeech*.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, and 1 others. 2021. Robust speech recognition via large-scale weak supervision. In *Proceedings of ICASSP*, pages 7383–7387. IEEE.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*. ICLR.

Hye jin Shim, Md Sahidullah, Jee weon Jung, Shinji Watanabe, and Tomi Kinnunen. 2024. Beyond silence: Bias analysis through loss and asymmetric approach in audio anti-spoofing. *arXiv preprint arXiv:2406.17246*.

Daniel Jurafsky and James H Martin. 2000. *Speech and Language Processing*. Prentice Hall.

Hamza Kheddara, Mustapha Hemis, and Yassine Himeur. 2024. Automatic speech recognition using advanced deep learning approaches: A survey. *arXiv preprint arXiv:2403.01255*.

Zulfidin Khodzhaev. 2024. A practical guide to spectrogram analysis for audio signal processing. *arXiv preprint arXiv:2403.09321*.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2017. A study on data augmentation of reverberant speech for robust speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5220–5224. IEEE.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*.

Shu Liu and Wayne G. Lutters. 2021. Speech interfaces for older adults: Supporting aging in place. *ACM Transactions on Accessible Computing (TACCESS)*, 14(3):1–26.

Julia Mainzinger and Gina-Anne Levow. 2024. Fine-tuning asr models for very low-resource languages: A study on mvskoke. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 76–82, Bangkok, Thailand. Association for Computational Linguistics.

Anas Malik, Zubayer Hossain, and Firoj Alam. 2021. Automatic speech recognition: A review. In *2021 2nd International Conference on Artificial Intelligence and Data Analytics (CAIDA)*, pages 1–6. IEEE.

Ziqiang Meng, Shuai Li, Dong Yu, and 1 others. 2021. A survey on speech evaluation metrics. *APSIPA Transactions on Signal and Information Processing*, 10:1–14.

Alan Morris, Victor Maier, and Phil Green. 2004. Spoken language understanding: Systems for extracting semantic information from speech. *IEEE Signal Processing Magazine*, 21(5):67–76.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: A large-scale speaker identification dataset. *Interspeech*, pages 2616–2620.

Shrikanth Narayanan and Panayiotis Georgiou. Real-time emotion detection from speech using audio and voice quality features. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Shweta Pandey and Akhilesh Mishra. 2019. Transgender inclusion and voice technology: A social justice perspective. *Technology in Society*, 59:101–120.

Rohit Prabhavalkar, Takaaki Hori, Tara N. Sainath, Ralf Schlüter, and Shinji Watanabe. 2021. End-to-end speech recognition: A survey. *arXiv preprint arXiv:2108.10520*.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, and 1 others. 2020. Mls: A large-scale multilingual dataset for speech research. *Interspeech*.

Hendrik Purwins, Bo Li, Tuomas Virtanen, Jan Schlüter, Shuo-Yan Chang, and Tara N. Sainath. 2019. Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 13(2):206–219.

Lawrence R Rabiner and Ronald W Schafer. 1978. Digital processing of speech signals. *Prentice-Hall Signal Processing Series*.

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.

Alec Radford, Jong Wook Kim, Tian Xu, Greg Brockman, and Christine McGuffie. 2022. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 19123–19139. PMLR.

C Ramesh and Ramesh Gupta. 2021. Challenges in speech recognition for low-resource and regional indian languages. In *IEEE International Conference on Communication and Signal Processing*, pages 145–150. IEEE.

Javier Ramírez, Juan M Górriz, and Juan C Segura. 2007. Voice activity detection. fundamentals and speech recognition system robustness. *Robust speech recognition and understanding*, 1:1–22.

Ramya Srinivasan, Shalini Aravindhan, and Ravi Mahalingam. 2023. Inclusive speech recognition: Annotating transgender and elderly tamil voices. In *Proceedings of LT-EDI*.

Tuomas Tolonen and Matti Karjalainen. 2000. A computationally efficient multipitch analysis model. *IEEE Transactions on Speech and Audio Processing*, 8(6):708–716.

Werner Verhelst and Maarten Roelands. 2000. Perceptual audio coding based on a signal-adaptive psychoacoustic model. *IEEE Transactions on Speech and Audio Processing*, 8(3):330–338.

Dong Wang, Xiaodong Wang, and Shaohe Lv. 2019. An overview of end-to-end automatic speech recognition. *Symmetry*, 11(8):1018.

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *arXiv preprint arXiv:2312.12148*.

Guorong Xuan, Wei Zhang, and Peiqi Chai. 2001. Em algorithms of gaussian mixture model and hidden markov model. In *Proceedings of the IEEE International Conference on Image Processing (ICIP 2001)*, pages 733–736. IEEE.

Chenglin Yu, Ming Yin, Shuai Wang, Zhenhua Chen, and Xiaodong Wang. 2021. M2met: A multi-modal multi-genre dataset for speaker profiling and recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3730–3743.

Dong Yu and Li Deng. 2017. Automatic speech recognition. In *Automated Speech Recognition*, pages 1–18. Springer.

Yu Zhang, Yashesh Wu, William Chan, Navdeep Jaitly, and Quoc V. Le. 2021. Benchmarking robust speech recognition: Background noise, babble, and channel variation. In *Proc. Interspeech*, pages 3076–3080.