Automatic Speech Recognition System for Malasar Language using Multilingual Transfer Learning

Basil K Raju, Leena G Pillai, Kavya Manohar, Elizabeth Sherly
Digital University Kerala
Thiruvananthapuram
Kerala, India

Abstract

This study pioneers the development of an automatic speech recognition (ASR) system for the Malasar language, an extremely low-resource ethnic language spoken by a tribal community in the Western Ghats of South India. Malasar is primarily an oral language which does not have a native script. Therefore, Malasar is often transcribed in Tamil script, a closely related major language. This work presents the first ever effort of leveraging the capabilities of multilingual transfer learning for recognising malasar speech. We fine-tune a pre-trained multilingual transformer model with Malasar speech data. In our endeavour to fine-tune this model using a Malasar speech corpus, we could successfully bring down the WER to 48.00% from 99.08% (zero shot baseline). This work demonstrates the efficacy of multilingual transfer learning in addressing the challenges of ASR for extremely low-resource languages, contributing to the preservation of their linguistic and cultural heritage.

1 Introduction

This work introduces an automatic speech recognition (ASR) system for the Malasar language, spoken by the indigenous Malasar community in the Western Ghats region of southern India[1]. The Malasar language is at risk due to its limited number of speakers, with just 7,760 reported in the 2001 census of India (Hazarika and Babu, 2023). This ASR system, developed with the goal of preserving the Malasar culture and language, is based on fine tuning the Whisper transformer model, which is pretrained on 680,000 hours of multilingual audio data, using transfer learning approach (Radford et al., 2023).

The Malasar language shares lexical similarities with multiple other languages, including Tamil, Malayalam, Muduga, Eravallan, and certain Irula dialects (Varghese, 2015). The different language influences in Malasar will result in frequent code-switching. The Malasar language is not well-documented in written form, and there is severe scarcity of quality audio data for training ASR models. The diversity in phonetic and linguistic characteristics along with the data scarcity contribute towards difficulty in building an ASR system for Malasar.

The objectives of this work include addressing data scarcity by collecting and curating Malasar speech data, developing an ASR model capable of handling code-switching and multilingualism, and contributing to the preservation of the Malasar linguistic heritage. This work represents a significant step towards safeguarding linguistic diversity in the face of endangered languages.

Contributions:

This work involved the creation of a Malasar speech corpus and the development of an ASR model for this endangered language. The key contributions are outlined below:

- Creation of a Malasar speech corpus, consisting of 5 hours of native Malasar speech accompanied by corresponding transcripts in the Tamil script.
- Development of a fine-tuned ASR model tailored for Malasar speech recognition.

2 Literature Review

Low-resource languages (LRLs) are languages with limited resources, such as data, tools, and expertise. Working with LRLs presents a number of challenges, including data scarcity, tool scarcity, and expertise scarcity. These challenges were addressed by using techniques such as data augmentation, transfer learning, and domain adaptation (Magueresse et al., 2020; Ranathunga et al., 2023).

[1] https://www.ethnologue.com/language/ymr/

The introduction of transformer models in ASR has indeed brought about a significant revolution in the field, especially for low-resource languages. ASR performance has improved drastically in recent years, mainly enabled by self-supervised learning (SSL) based acoustic models such as Wav2vec2 (Schneider et al., 2019; Conneau et al., 2020) and large-scale multi-lingual transformer models like Whisper (Radford et al., 2023).

Wav2vec2 is an encoder only transformer model. Using self supervised learning, this neural speech representation model is trained with a lot of unlabelled data (Schneider et al., 2019; Baevski et al., 2020), in a process referred to as pre-training. For example the XSL-R model, that preceded the Wav2vec2 model, was trained on 436K hours of publicly available unlabelled speech data from 128 languages (Conneau et al., 2020). The pre-trained model could be adapted with comparably less amount of labelled data for a specific task like ASR, making it suitable for low-resource languages. The process of fine-tuning the pre-trained encoder model involves adding a linear classification layer on top of the transformer and training the entire model by minimising the connectionist temporal classification (CTC) loss function (Graves et al., 2006).

Whisper transformer model, on the other hand is a sequence-to-sequence model which consists of an encoder and a decoder linked via a cross-attention mechanism (Radford et al., 2023). Unlike Wav2vec model, Whisper is trained on labelled speech data that amounts to 680K hours, of which 117K hours is non-English speech data in 99 different languages.

There has been efforts in the past to fine-tune both XLS-R and Whisper to make them transcribe seen and unseen languages, both high and low resource ones (Guillaume et al., 2022; Rouditchenko et al., 2023). Attempts to fine-tune ASR systems for Dravidian languages from X-LSR models have shown effective improvements in accuracy (Akhilesh et al., 2022; Manohar et al., 2023; Anoop and Ramakrishnan, 2023). In this work, we aim to undertake a groundbreaking initiative by developing an ASR system for the Malasar language using Whisper Transformer models.

3 Malasar Speech Corpus

The Malasar language is primarily oral and does not possess a standardized script for written documentation. Given the linguistic similarities between Malasar and Tamil, transcription efforts often rely on using the Tamil script as a means of representing the spoken Malasar language in a written form.

Table 1: Details of Malasar speech corpus

Duration	5 Hours
Speakers	2
Language	Malasar
Script	Tamil
Sample rate	16 kHz
Domain	Stories

A corpus of Malasar speech to fine tune the transformer model was created from a spoken Bible database[2]. The original datasource contained audio recordings of Malasar speech, along with their transcriptions in Tamil script. The audio recordings are a total of 5 hours in length. The speakers consist of one male and one female in the age group of 30-35. The details about the dataset is described in Table 1.

4 Proposed Malasar ASR

Our proposed Malasar language ASR model is based on Whisper transformer architecture (Radford et al., 2023). Transformers offer a powerful architecture for ASR due to their ability to model long-term dependencies, efficient parallelization, self-attention mechanism, multilingual capabilities, and contextual understanding (Vaswani et al., 2017). Pretrained transformers offer added advantages for ultra-low resource languages as they can learn from a small corpora. These advantages have made transformers a popular choice for various natural language processing tasks.

4.1 Architecture

In this proposed method, we apply transfer learning on the Whisper transformer model using the target Malasar speech dataset. The fine-tuning process is all about updating the parameters of the pre-trained model through backpropagation with the Malasar language dataset. This enables the model to learn the language specific features and intricacies of the Malasar language, thereby improving its accuracy and performance.

Whisper's established proficiency in transcribing Tamil speech to Tamil script grants it a significant

[2]The datset was provided by Wycliffe India https://wycliffeindia.in/

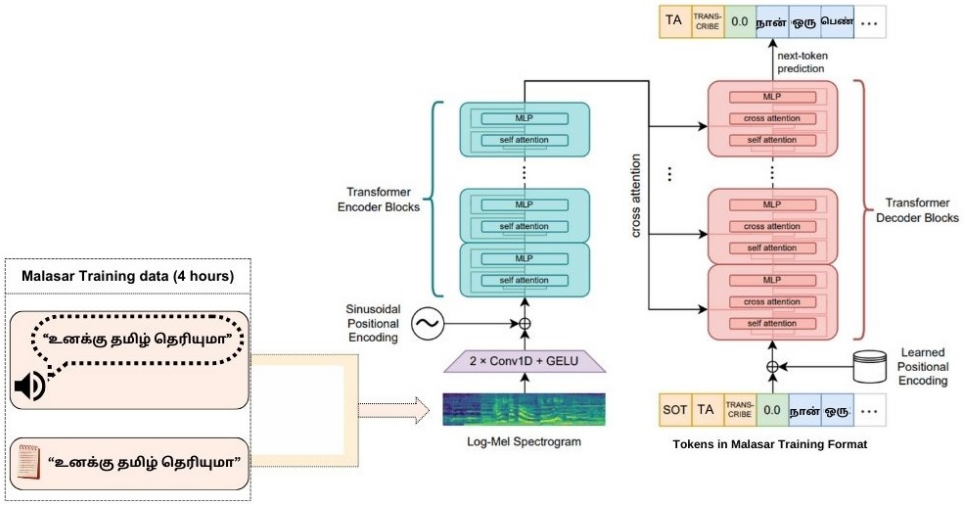

Figure 1: Block schematic representation of finetuned Whisper Transformer with Malasar speech data

advantage in transcribing Malasar, as Malasar is typically represented in Tamil script, a closely related language. This eliminates the need to learn a new script or spelling pattern, enhancing transcription accuracy and reducing the required training data for fine-tuning Whisper for Malasar. Consequently, Whisper becomes an efficient and effective choice for developing an ASR system tailored to Malasar.

The architecture of encoder-decoder structure in the Whisper transformer model, we propose to fine tune is shown in Fig. 1. Hidden state representations that capture key aspects of spoken language are created from the audio input by the encoder. The decoder then processes these hidden state representations to produce the matching text transcriptions. The language model is integrated into the system architecture itself in deep fusion. This contrasts with shallow fusion, in which the encoder and language model are integrated externally. The ability to train the whole system end-to-end using the same training data and loss function is one of the benefits of deep fusion which provide greater flexibility and overall better performance (Radford et al., 2023). The ASR model pipeline in Whisper can be divided into two components, (1) preprocessing and feature extraction, and (2) The encoder-decoder model.

4.2 Preprocessing and Feature Extraction

The feature extractor converts the raw audio signal into a sequence of features that can be used by the Whisper model. This involves resampling all audio to 16 kHz. Using windows of 25 milliseconds and a stride of 10 milliseconds, a log-magnitude Mel spectrogram representation with 80 channels is constructed. The input is scaled globally between -1 and 1, with the goal of achieving an essentially zero mean throughout the pre-training dataset to normalise the features (Radford et al., 2023).

4.3 Encoder-Decoder Structure

The encoder starts by employing a tiny stem to process this input representation. The stem uses the Gaussian error linear unit (GELU) activation function and has two convolution layers with a 3-filter width. The stride of the second convolution layer is two. The output of the stem is then further enhanced by sinusoidal position embeddings. The model takes the sequence of features as input and outputs a sequence of probabilities that correspond to the possible words that were spoken. The transformer normalises the encoder output using the final layer and pre-activation residual blocks. The decoder uses coupled input-output token representations and learnt position embeddings. It matches the encoder's width and the number of transformer blocks that are there. This model is pretrained on a large dataset of audio recordings and transcripts.

The encoder and decoder blocks are the vital designs of the transformer architecture. The encoder block entitled to process the input audio sequences and generates contextualised representations. It uses self-attention mechanisms to capture the dependencies and relationships between different input sequence and transfer this information to the following layers. The decoder block generates the output sequence in a sequential manner, by follow-

ing the encoder's contextualised representations and self-attention ensures that the relevant information are considered at each step in the output sequence.

5 Experimental Setup

Five hours of Malasar speech corpus and its corresponding Tamil transcript is utilised for the experiment described in this work. 80% of the corpus is used for fine-tuning the pretrained model, while the remaining 20% was allocated for validation and testing purposes. The data partition used for fine tuning, validation and testing is listed in Table. 2.

Table 2: Description of the speech datasets

	Dataset	Split
Train	Stories	80%
Validation	Stories	5%
Test	Stories	10%

The fine tuned model designed by optimizing several hyperparameters to achieve the highest performance possible such as training batch size, learning rate, warmup steps and evaluation batch size. We adopted a trial-and-error approach, where experimentation conducted with various values are assessed for their impact on the performance of the model. The best combination of hyperparameters based on WER is considered as the optimum parameter set for this problem.

Table 3: Training parameters

Parameters	value
Train Batch size	32
Learning rate	$1e^{-5}$
Warmup steps	100
Evaluation Batch size	16

In Table 3, the fine-tuning parameters are listed. The batch size, set to 32, controls the number of training samples processed in each step. A learning rate of $1e^{-5}$ ensures small parameter adjustments. Warmup steps are set at 100 during training, gradually increasing the learning rate for stability. Performance is evaluated regularly, with model checkpoints saved every 200 steps, and evaluation conducted at these intervals using a batch of 16 audio samples. A single Nvidia DGX A100 GPU with 80 GB RAM was used for fine-tuning experiments.

6 Training and validation

ASR model performance is measured in terms of word error rate (WER). WER is calculated on the basis of the number of words (N) in the ground truth speech transcript and the number of word insertions (I_w), deletions (D_w) and substitutions (S_w) in the predicted transcript (Eq. 1).

$$WER = \frac{(I_w + D_w + S_w)100}{(N)} \quad (1)$$

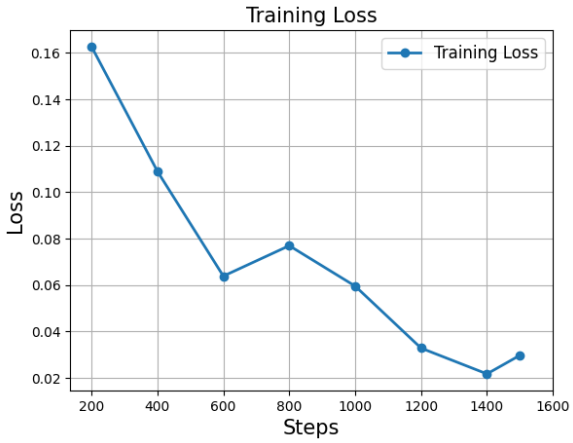

Figure 2: Reduction in training loss function

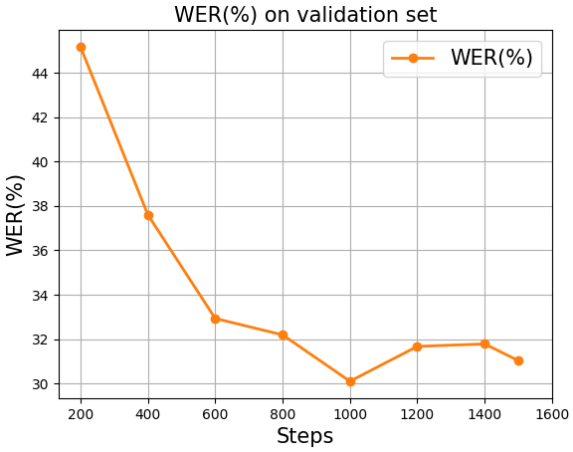

Figure 3: WER reduction on validation set during training

The reduction in WER on the validation dataset during each evaluation step is described in Figure 3. The model has gone through 1500 training steps with a WER of 31.03% on the validation set.

7 Result and Discussion

We evaluate the performance of the first ever Malasar language ASR on the held out test dataset.

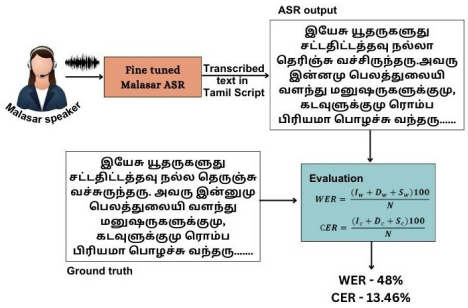

Figure 4: Performance of Malasar ASR. Sample output from the ASR is indicated against the ground truth transcript.

Being a Dravidian language, Malasar has a morphologically complex word structure. It results in unreasonably high WER, even if a single space was introduced in the decoded text when compared to the ground truth transcript (Anoop and Ramakrishnan, 2023). A better indicator of performance hence would be character error rate (CER), where the error rates are calculated at character level according to eq. 2.

$$CER = \frac{(I_c + D_c + S_c)100}{(N)} \quad (2)$$

The comparative analysis of WER and CER between the zero shot baseline (Whisper small model) and the fine-tuned models illustrates significant improvements in the accuracy of the ASR system. The baseline model exhibits high error rates in both WER (99.08%) and CER (34.74%), as indicated in Table. 4. In contrast, the fine-tuned model yields notably reduced error rates with a WER of 48.00% and a CER of 13.46%. These results were additionally verified by linguistic experts at Wycliff India. These improvements underscore the effectiveness of the fine-tuning process, enhancing the ASR system's ability to accurately transcribe spoken language into written text, a valuable development in speech recognition tasks.

Table 4: Performance Evaluation

Models	WER (%)	CER (%)
Baseline	99.08	34.74
Fine Tuned	**48.00**	**13.46**

8 Conclusion

This work presents the preliminary results of applying transfer learning on a transformer-based ASR system tailored for the endangered Malasar language. Our experimental results show a substantial reduction in WER as well as CER on the fine-tuned model, in comparison with the zero shot baseline. The WER and CER on the unseen test dataset are respectively 48.00% and 13.46%. Looking forward we plan to collaborate with the lingustic experts to expand the domain of training dataset and improve the transcription accuracy further. Moreover, the ASR system should be integrated into language learning and preservation efforts to maximise its impact on the Malasar community.

Acknowledgements

The Malasar speech corpus creation of this work was supported by the Wycliffe India. Wycliffe India is an interdenominational, non-sectarian and non-profit mission organization. The authors are grateful for the support and linguistic assistance provided by the organization.

References

A Akhilesh, Brinda P, Keerthana S, Deepa Gupta, and Susmitha Vekkot. 2022. Tamil speech recognition using xlsr wav2vec2.0 & ctc algorithm. In *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6.

C S Anoop and A G Ramakrishnan. 2023. Exploring a unified asr for multiple south indian languages leveraging multilingual acoustic and language models. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 830–837.

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.

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*.

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*, pages 369–376.

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 *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 170–178, Dublin, Ireland. Association for Computational Linguistics.

Chaya R Hazarika and Bontha V Babu. 2023. Prevalence of hypertension in indian tribal population: a systematic review and meta-analysis. *Journal of Racial and Ethnic Health Disparities*, pages 1–17.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

Kavya Manohar, Gokul G. Menon, Ashish Abraham, Rajeev Rajan, and A. R. Jayan. 2023. Automatic recognition of continuous malayalam speech using pretrained multilingual transformers. In *2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS)*, pages 671–675.

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.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

Andrew Rouditchenko, Sameer Khurana, Samuel Thomas, Rogerio Feris, Leonid Karlinsky, Hilde Kuehne, David Harwath, Brian Kingsbury, and James Glass. 2023. Comparison of Multilingual Self-Supervised and Weakly-Supervised Speech Pre-Training for Adaptation to Unseen Languages. In *Proc. INTERSPEECH 2023*, pages 2268–2272.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Bijumon Varghese. 2015. *The Tribes of Palakkad, Kerala: A Sociolinguistic Profile*. SIL Electronic Survey Reports.

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.