Language Understanding λ(μ) Laboratory

Language Understanding Laboratory, Myanmar

Summer Internship (2025) Report

Study of Burmese Speech Processing using Transfer Learning Approaches

Submitted by

ASR Team and TTS Team

Under the supervision of

Prof. Ye Kyaw Thu
Affiliation of Supervisor

With mentorship from

Thura Aung
Affiliation of Mentor 1

Khaing Hsu Wai
Affiliation of Mentor 2

Team Members

Sai Wai Yan Phyo	Thiri Thaw	Moe Chan Myae Maung
Thet Htet San	Myat Oo Swe	Su Sandi Linn
Htet Arkar	Min Thiha Tun	Saw Zi Dunn
Kyi Thant Sin	Eaint Lay Hmone	Htwe Myat Cho
Yadana Myint Hein	Ye Bhone Lin	Cham Myae Phyo
Thant Htut Aung	Nann Oak(Caesar)	Kyawt Eaindray Win
Thet Su Sann	Phoo Pwint Cho Thar	Htut Ko Ko

Abstract

This report presents a comprehensive study on Burmese speech processing, leveraging a proprietary dataset of 56.23 hours and 21.71 hours. We explore various approaches in Automatic Speech Recognition (ASR) systems, emphasizing transfer learning techniques and post-asr correction approach. Our experiments utilize state-of-the-art models, including Whisper, and mT5 for post-ASR correction. We also investigate data augmentation strategies to increase diversity, which helps our models to generalize better and reduces overfitting. The performance and adaptability of the Whisper models (110.1% -> 15.66%) under these conditions are thoroughly evaluated, providing information on their effectiveness for Burmese speech processing and encouraging further exploration, as detailed in this report.

1 Introduction

Automatic Speech Recognition (ASR) for Burmese aims to convert spoken language into written text, thus improving accessibility and communication. Recent advances in ASR have been driven by transfer learning, fine-tuning, and end-to-end modeling techniques, which significantly enhance recognition accuracy.

ASR systems are now applied in various domains, including virtual assistants, smart devices, automotive voice control, healthcare documentation, and customer service call transcription. Although there are mature systems for major languages such as English, Chinese, and Japanese, Burmese ASR remains underdeveloped. This research addresses this gap by exploring transfer learning approaches for Burmese speech processing.

For Burmese ASR, we explore both transfer learning and zero-shot approaches using the OpenAI Whisper model and the Massively Multilingual Speech (MMS) system. Despite their success in high-resource settings, Whisper and similar models often encounter difficulties with the unique linguistic and grammatical properties of low-resource languages [3].

In recent years, several datasets for Burmese ASR have become available. The UCSY-SC1 corpus [10] represents one of the first resources, while the myMediCon project extends this effort with an 11-hour Burmese medical speech corpus and comparative evaluations of Transformer and RNN-based ASR models and obtain WER 23.1% in transformers.[5].

In this work, we fine-tuned Whisper models (Tiny, Small, and Medium) using three datasets: the open-source OpenSLR80 corpus (4.12 hours) [11], Google FLEURS (17.59 hours) [2], and our myMediTalk medical dataset (56.23 hours). The original Whisper Small model reported a word error rate (WER) of 110.1% on FLEURS [1]. By contrast, our experiments Whisper Small achieved 51.21% WER on the combined OpenSLR80 and FLEURS datasets, and 15.66% WER on the entire myMediTalk dataset, representing a substantial improvement from 110.1% to 15.66%. Addiontionally, Whisper Tiny, Small, and Medium models outperformed the original Whisper Tiny, Small, and Medium on each datasets.

We further examined ASR error correlations across Whisper Tiny, Whisper Small, and MMS models. The analysis was conducted by combining error patterns from these systems on the OpenSLR80 and FLEURS datasets. Without phonemic features, Transformer-based models showed limited improvement over prior ASR baselines. To address this limitation, we incorporated phonemic features derived from the International Phonetic Alphabet (IPA) and applied grapheme-to-phoneme (G2P) conversion [14], specifically tailored for low-resource Burmese.

2 Related Work

Several approaches have been explored for Burmese ASR. The myMediCon focused on end-to-end models, including Transformer and Recurrent Neural Network (RNN) architectures, for Burmese speech recognition [5]. Other end-to-end systems have applied Connectionist Temporal Classification (CTC), which addressed the low-resource nature of Burmese. Using a 26-hour Burmese speech corpus, one study with CTC achieved a Character Error Rate (CER) of 4.72% [6].

Earlier research also applied Hidden Markov Models (HMMs) for classification tasks and speech recognition [12]. The ChildASR system employed a Gaussian Mixture Model–based Hidden Markov Model (GMM-HMM). In this work, a dataset of 2,682 sentences,

totaling approximately 5 hours of speech from a Primary Myanmar Textbook, was used, achieving a Word Error Rate (WER) of 14.45% [9].

Building on previous research, it is evident that while Burmese ASR has benefited from methods such as HMM-based models, CTC-based approaches, and end-to-end architectures, the exploration of transfer learning and ASR error correlation remains limited. Our work introduces these techniques as a new contribution to low-resource Burmese ASR, demonstrating how transfer learning can enhance performance and how error correlation analysis provides deeper insight into model behavior and weaknesses. Together, these approaches represent a novel direction for advancing ASR in under-resourced languages.

3 Medical Dataset

Several Burmese speech corpora exist for ASR research. The UCSY-SC1 Myanmar speech corpus contains over 42 hours of speech from news and daily conversation domains, collected from a large number of speakers [10]. Another example is a high-quality crowdsourced Burmese speech dataset with transcribed audio files recorded by volunteers, available on platforms such as `openslr.org` [11]. For the medical domain, the Burmese Medical Speech Conversations (myMediCon) corpus provides nearly 11 hours of speech among doctors, nurses, and patients, designed to improve medical ASR for Burmese [5].

Building upon myMediCon, we developed a substantially larger medical speech corpus, **myMediTalk**. Details are presented below.

3.1 Resource and Statistics

The myMediTalk corpus was recorded using conversational medical speech from 9 male and 12 female speakers. Audio was captured at a sampling rate of 16 kHz, with 50.63 hours (26,264 transcriptions) of training data and 5.60 hours (2,920 transcriptions) of testing data.

To assess audio quality, we employed both MOSNet and Signal-to-Noise Ratio (SNR). Traditionally, the Mean Opinion Score (MOS) depends on human listeners, which is subjective, inconsistent across individuals, and requires many participants for reliability [7]. MOSNet was developed to overcome these limitations by automatically predicting similarity scores, and preliminary studies show strong correlation with human ratings [7].

We also computed SNR to further evaluate audio quality. Table 15 shows the results.

Dataset	**Train (hrs)**	**Test (hrs)**	**MOSNet**	**SNR (dB)**
myMediTalk	50.63	5.60	4.04	23.58
OpenSLR80	3.70	0.42	4.06	36.51
FLEURS	15.95	1.64	4.14	25.07

Table 1: Speech dataset statistics and quality metrics.

From Table 15, it can be observed that both MOSNet and SNR values for myMediTalk are lower compared to other datasets. This indicates limited audio quality, which contributes to higher ASR error rates. Consequently, our study emphasizes ASR error correlation analysis and robust training methods to mitigate these challenges.

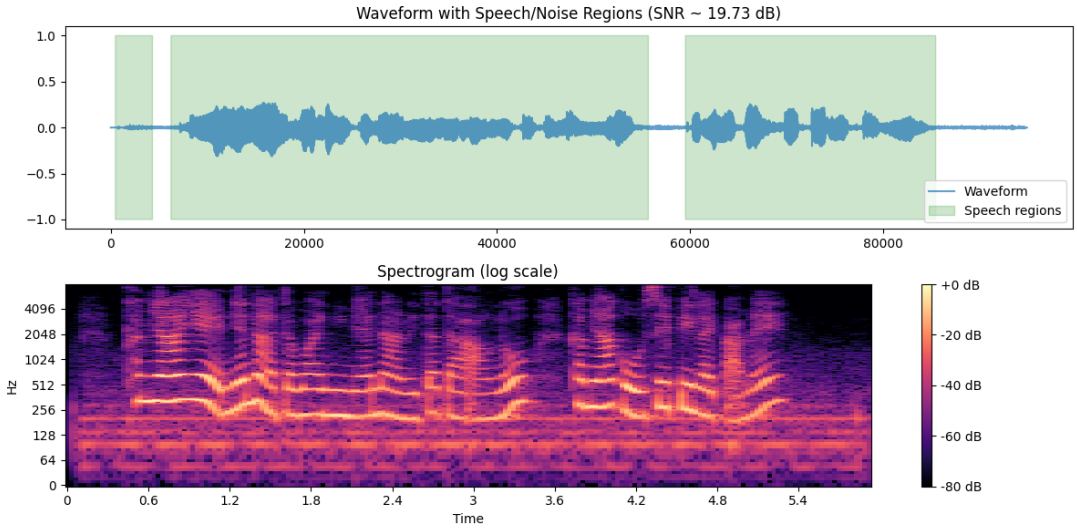

Figure 1: Speech/non-speech segmentation and spectrogram of a myMediTalk training utterance (SNR = 19.73 dB).

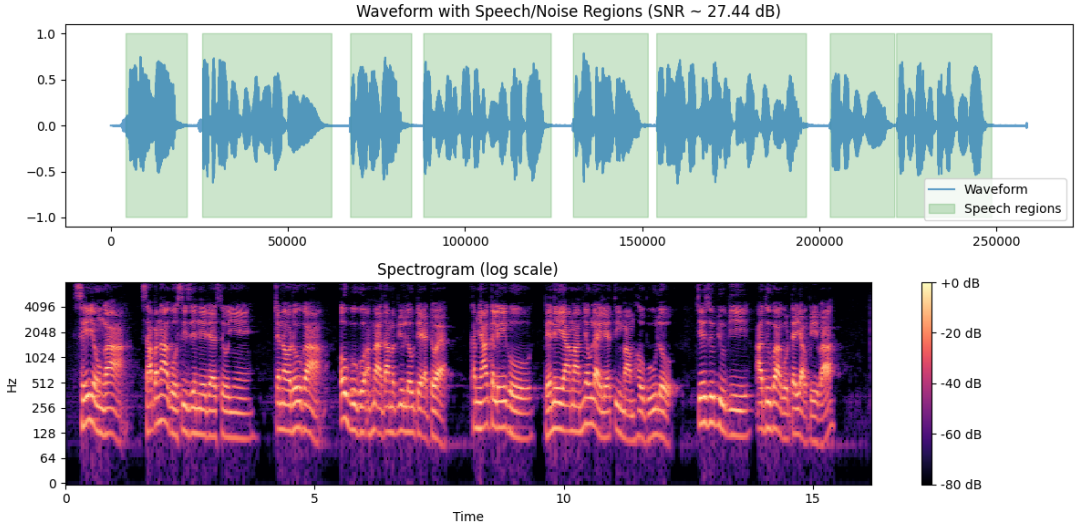

Figure 2: Speech/non-speech segmentation and spectrogram of a myMediTalk test utterance (SNR = 27.44 dB).

3.2 Dataset Preparation

Before uploading the corpus to Hugging Face, we applied **voice source separation (VSS)** to remove background noises (e.g., car horns, mouse clicks) present in recordings collected across multiple environments. We experimented with several open-source VSS toolkits/models, including Asteroid, VoiceFixer-UNet, Demucs (UNet), Kim-UNet, SpeechBrain+Librosa, and SpeechBrain+SepFormer.

Based on Table 2, we selected Demucs (UNet) [4] as the preferred model due to its stability and natural voice preservation, which are critical in medical scenarios.

For the myMediTalk corpus, we uploaded the data to Hugging Face in **Parquet** format to enable efficient streaming without full downloads, avoiding out-of-memory (OOM) errors during training. This is especially beneficial for free platforms such as Google Colab (T4 GPUs) and Kaggle (P100 GPUs), where resources are limited.

Toolkit/Model	Qualitative Observations
Asteroid	Slightly lower output volume; minor residual noise remained.
VoiceFixer-UNet	Removed strong noise effectively, but sometimes altered transcripts and introduced unnatural voice quality.
Demucs (UNet)	Stable output; handled background noise moderately well, especially in low-noise conditions. Less effective in high-noise scenarios.
Kim-UNet	Strong noise reduction with minor high-frequency distortions.
SpeechBrain + Librosa	Enhanced loudness but weak in suppressing heavy noise.
SpeechBrain + SepFormer	Best separation overall, but occasionally altered speaker characteristics and masked parts of audio.

Table 2: Qualitative comparison of VSS toolkits on myMediTalk audio.

Dataset	Texts
OpenSLR80	253
FLEURS	384
Small Burmese Audio Corpus	2682

Table 3: ASR Error Correlation dataset composition.

We used the three pretrained ASR models that shown in 4 in both fine-tuning and zero-shot approaches, creating an **ASR Error Correlation dataset** from their outputs. For this dataset, we included only the OpenSLR80 test set, the FLEURS test set, and a small Burmese audio corpus (about 2,600 samples).

4 Methods

4.1 Whisper Model

Since the Whisper model uses a byte-level Byte Pair Encoding (BPE) tokenizer [1], which is designed primarily for English, we adapted it for Burmese by converting transcriptions into the Myanmar syllable format [13]. Using syllable-based targets helps guide the Whisper model more effectively, as Burmese is a syllable-timed language, and this representation improves alignment between speech and transcriptions. Additionally, the Whisper model is an encoder-decoder Transformer [1][15] that processes audio files segmented into 30-second clips. Audio features are extracted from the mel spectrogram using Whisper's FeatureExtraction module. Both preprocessing steps—tokenization and feature extraction—are integrated via the Whisper Processor for all Whisper model variants (Tiny, Small, and Medium) applied to the OpenSLR80 + FLEURS and myMediTalk datasets as shown in Table 15. This setup enabled effective fine-tuning and evaluation of the Whisper models across both the OpenSLR80 + FLEURS and myMediTalk datasets,

accommodating Burmese speech characteristics through syllable-based transcription representation.

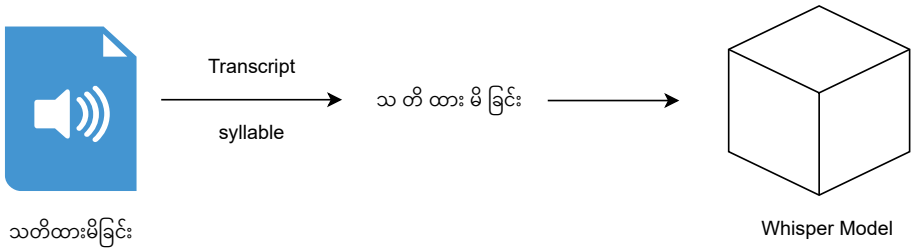

Figure 3: Pipeline showing syllable conversion of transcriptions before fine-tuning Whisper.

Model	Parameters	Languages
MMS-1B ASR	∼1B	1,100+
Whisper Tiny*	39M	99
Whisper Small*	244M	99
Whisper Medium*	769M	99

Table 4: Pretrained ASR models used in fine-tuning and zero-shot experiments. *Whisper models were finetuned for Burmese.

4.2 mT5 model

For the ASR error correction task, we used the multilingual T5 (mT5) model [16], a Transformer-based sequence-to-sequence model pretrained on large-scale multilingual corpora. The mT5 architecture is well-suited for text-to-text generation tasks, making it effective for correcting transcription errors generated by Automatic Speech Recognition (ASR) systems.

We constructed the **ASR Error Correction dataset** using the outputs from Whisper-Tiny, Whisper-Small, and MMS-1B ASR models. To build this dataset, we combined the OpenSLR80 test set, the FLEURS test set, and a small Burmese audio corpus (about 2,600 samples). This provided a diverse set of ASR outputs containing real-world error patterns across different datasets and models.

To further enhance correction accuracy, we enriched the training data with **IPA (International Phonetic Alphabet)** representations and **G2P (grapheme-to-phoneme)**[14] features. These features provide linguistic guidance by encoding pronunciation cues, which help the mT5 model better capture the phonological structure of Burmese and correct transcription errors more effectively.

Figure 4 illustrates the complete ASR error correction pipeline using the mT5 model, where ASR outputs are combined, enriched with IPA and G2P features, and then used to fine-tune mT5.

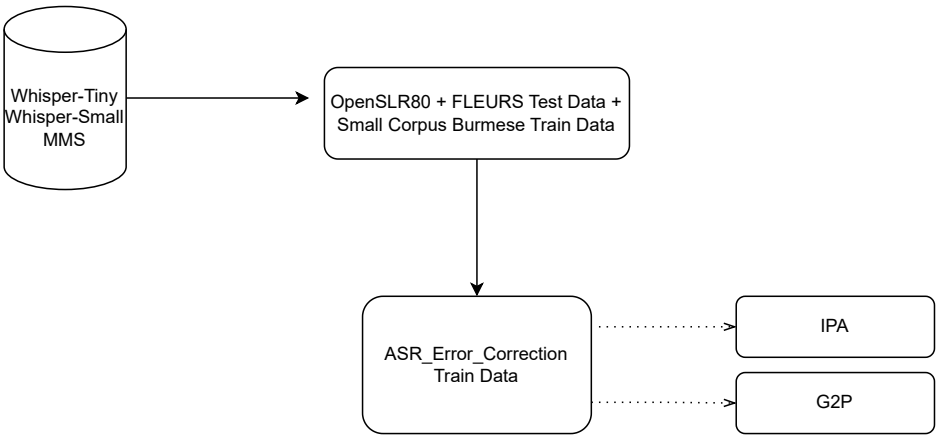

Figure 4: Pipeline of ASR error correction using the mT5 model.

4.3 Data Augmentation Strategy

To enhance the robustness of our Automatic Speech Recognition (ASR) models and reduce overfitting, we incorporated a variety of audio augmentation techniques. Data augmentation increases the diversity of training samples by simulating realistic acoustic conditions, different speakers, and recording artifacts that may not be present in the original data. This helps the models generalize better to new environments and unseen speakers. Data augmentation is particularly important for improving ASR performance in low-resource settings like Burmese. In this project, we designed an augmentation pipeline that operates on both waveform and spectrogram levels, followed by iterative model training and fine-tuning.

4.3.1 Waveform Augmentation

Waveform augmentations were applied directly on the raw audio signal. These transformations aim to reproduce natural variations that occur in real-world speech data. For example, loudness adjustment and noise injection simulate differences in recording equipment and background environments, while pitch shifting and speed perturbation mimic inter-speaker variations such as vocal characteristics and speaking rate [8]. Temporal manipulations such as cropping, masking, and time shifting introduce additional variability, while Vocal Tract Length Perturbation (VTLP) warps the spectral envelope to emulate differences in speaker physiology [8]. Together, these methods produce a rich set of acoustically diverse training samples.

4.3.2 Spectrogram Augmentation

In addition to raw waveform processing, we also applied augmentations on spectrogram representations. These transformations operate in the time-frequency domain and are particularly effective in preventing overfitting to narrow acoustic patterns. Time and frequency masking (as in SpecAugment) force the model to rely on contextual information rather than memorizing specific spectral [8]. Other techniques, such as time warping and frequency shifting, introduce variability in temporal and spectral dimensions, while dynamic range compression and band dropping simulate channel distortion and signal degradation. More advanced strategies, including resize-and-crop and patch swapping,

further diversify spectrogram structures by rescaling or rearranging time-frequency regions.

4.3.3 Integration with Model Training

The augmented data were systematically stored with updated metadata, allowing seamless integration into ASR training workflows. We first trained Whisper models with the augmented dataset, enabling them to capture greater variability in Burmese speech. The models were then fine-tuned on our proprietary recorded dataset to better align with the target domain. Following this, model predictions were used to generate additional transcriptions, which were further incorporated into training. This iterative cycle of augmentation, training, and fine-tuning allowed us to expand the effective dataset size and improve recognition performance under diverse acoustic conditions.

5 Experiments

5.1 Finetuning Whisper For Automatic Speech Recognition

We conducted two experiments using OpenSLR80 + FLEURS and myMediTalk datasets. For the OpenSLR80 + FLEURS dataset, we fine-tuned two Whisper models: Whisper Tiny and Whisper Small. To prevent out-of-memory (OOM) errors, we employed dataset streaming.

We fine-tuned the Whisper Tiny model with a training batch size of 16 and an evaluation batch size of 8 on a Google Colab environment using a T4 GPU. A learning rate of 1e-5 and warm up 500 steps was applied, which gradually reduced the word error rate (WER) during training. The model was trained for 20 epochs, with evaluation every 500 steps. At 4000 training steps, the model achieved a training loss of 0.065 and a validation loss of 0.202, resulting in a WER of 63.24%.

Similarly, the Whisper Small model was fine-tuned using the same batch sizes—16 for training and 8 for evaluation—on the same Google Colab T4 GPU setup. The learning rate remained at 1e-5, warm up 500 steps and the model was trained for 20 epochs up to 4000 steps, with evaluations every 500 steps. The fine-tuning process for Whisper Small mirrored that of Whisper Tiny, ensuring a consistent training environment. The Whisper Small model outperformed Whisper Tiny, achieving a training loss of 0.002 and a validation loss of 0.165 at 4000 steps, with a lower WER of 31.71%. This setup allowed a fair comparison of performance improvements between the two model sizes on our datasets. To put the chunk length 10 in parameters, the model can give the full audio transcriptions.

For the myMediTalk dataset, we fine-tuned the Whisper models in sequence—Tiny, Small, and Medium. During training, we encountered out-of-memory (OOM) errors while loading or streaming the dataset. To address this, we applied a chunking method, splitting the dataset into smaller portions consisting of 4 or 2 samples each. Each chunk contained approximately 6,500 or 10,000 samples. We fine-tuned Whisper Tiny, Small, and Medium on these chunks using the same hyperparameters: a learning rate of 1e-5, warmup steps of 100, and a maximum of 1000 training steps per chunk. Training was conducted on two T4 GPUs simultaneously, with each chunk trained independently for up to 1000 steps. Additionally, to handle potential interruptions due to GPU resource limitations, the training process includes a mechanism to save checkpoints and reload

the model. This allows training to resume seamlessly from the last saved state, ensuring robustness against GPU availability issues.

During inference, the Whisper model is limited to generating a maximum of 224 tokens, which corresponds to approximately 6 seconds of audio. To accommodate this limitation, the test audio was divided into chunks of 6 seconds each (chunk_length_s=6), with an overlap stride of 0.5 seconds (stride_length_s=0.5) between consecutive chunks. This approach ensures smooth transitions and reduces boundary artifacts when processing longer audio files.

ASR Model	**Train Hours**	**Parameter**	**Value**
Whisper Tiny*	4:05:25	chunk_length_s	10
Whisper Small*	4:44:52	chunk_length_s	10

Table 5: Whisper model training (OpenSLR80 + FLEURS) with exact runtime and inference parameter.

Model	**Subset Size**	**Samples per Batch**	**Training Time (hrs)**	**Parameter**	**Value**
Whisper Tiny	6,500	4	∼4	chunk_length_s	6
Whisper Small	6,500	4	∼8	chunk_length_s	6

Table 6: Whisper model training dataset subsets (myMediTalk) with runtime and inference parameter.

5.2 Post-ASR Correction with mT5

Parameters	**Values**
top_k	50
top_p	0.95
temperature	0.9
num_return_sequences	1
max_length	64

Table 7: mT5 ASR Error Correlation Inference

We selected the Multilingual T5 (mT5) model for post-ASR error correction due to its effectiveness as a compact language model that reduces word error rates (WER) from ASR outputs. For fine-tuning, we set the maximum input sequence length to 100 tokens, reflecting the relatively short average audio segments of our dataset. The model was trained on Google Colab using a T4 GPU (15 GB) and 12.7 GB RAM for 15 epochs. Evaluation and checkpoint saving occurred every 1000 training steps. After approximately 9,000 steps, the model showed significant convergence and strong performance on the ASR correction task. For inference, we used a single setup with mT5 incorporating phonetic features such as grapheme-to-phoneme (g2p) conversion and IPA representations to enhance accuracy. The fine-tuning settings remained the same for each feature type. While the small mT5 model with IPA features outperformed mT5 without any features and with only G2P features in terms of lower WER than the baseline, mT5 with G2P features achieved a better chrF++ score than the others.

6 Results & Discussion

6.1 Whisper Automatic Speech Recognition

From Table 8, it is clear that Whisper Small consistently performs better than Whisper Tiny across all training steps. For instance, Whisper Tiny reached a WER of 63.24%, while Whisper Small reduced it to 31.71% in training steps. Whisper Small also achieved a WER of 39.55%, which is a large improvement compared to the Original Whisper baseline at 110.1% WER [1].

However, performance is not uniform across all metrics. While Whisper Small had the lowest WER, Whisper Tiny achieved a higher chrF+ score of 0.6922, compared to 0.6202 for Whisper Small and 0.6711 for MMS. This shows that although Whisper Small is stronger in reducing word errors, Whisper Tiny and MMS still capture certain character-level similarities better.

The experimental results demonstrate clear differences in the quality of Myanmar transcription across the evaluated models in Table 10. In particular, Whisper Small shows strong performance not only in terms of word error rate (WER) but also in its ability to correctly represent Myanmar grammar and morphology. For example, in the phrase "ယင်း ဓနု ရာ", Whisper Small produces a grammatically valid rendering, while MMS outputs "ဓရ်ိး ဓနု ရာ", which is neither linguistically valid nor consistent with Myanmar orthography. Similarly, Whisper Tiny often substitutes unrelated tokens such as "ရင်း" or "ဓနု", indicating weaker representation of Myanmar grammar compared to Whisper Small.

These findings suggest that Whisper Small has implicitly learned the structural aspects of Myanmar grammar during training, allowing it to better generalize to unseen text. On the other hand, MMS, which was tested in a zero-shot setting, struggles with grammatical correctness despite producing some recognizable tokens. This highlights a key limitation of zero-shot approaches for low-resource languages: while they may capture broad phonetic correspondences, they often fail to encode fine-grained grammatical rules.

Step	**Whisper Tiny**			**Whisper Small**		
	Training Loss	Validation Loss	WER	Training Loss	Validation Loss	WER
1000	0.544	0.506	86.88	0.115	0.155	43.77
2000	0.140	0.215	68.23	0.027	0.148	35.76
3000	0.087	0.199	64.51	0.010	0.173	35.19
4000	0.065	0.202	63.24	**0.002**	**0.165**	**31.71**

Table 8: Comparison of Whisper Tiny vs Whisper Small on OpenSLR80 + FLEURS on Training Data

Model	**WER**	**chrF++**
Whisper Tiny	45.73	**0.6922**
Whisper Small	**39.55**	0.6202
MMS	45.36	0.6711

Table 9: Whisper Model on OPENSLR80 + FLEURS on Test Data

Model Approaches	
Ground Truth	အ လောင်း သည် ယင်း နေ ရာ တွင် ပေါ် လာ သည် မှာ တစ် ရက် ပင် ရှိ ပြီ ဟု ရဲ ဌာ န က ဆို ပါ သည်
Whisper Tiny	အ လောင်း သည် ရင်း နေ့ ရာ တွင် ဘော လာ သည် မှာ သစ် ရစ် ပင် ရှိ ဖြည် ဟို ရဲ ထား နာ က ဆို ပါ သည်
Whisper Small	အ လောင်း သည် ရင်း နေ ရာ တွင် ပေါ် လာ သည် မှာ တ ရစ် ပင် ရှိ ပြီး ဟု ရဲ ဌာ န က ဆို ပါ သည်
MMS	အ လောင်း သည် ရေ့် နေ ရာ တွင် ပေါ် လာ သည် မှာ သ ရပ် ပင် ရှိ ပြီး ဟု ရဲ ဌာ န က ဆို ပါ သည်

Table 10: Whisper Tiny and Small evaluating on OpenSLR80 + FLEURS.

Model	WER	chrF++
Whisper Tiny	28.31	0.7145
Whisper Small	**15.66**	**0.8695**
Whisper Medium	24.96	0.7493
Whisper Tiny (5000+ rows)	29.09	0.7104
Whisper Small (5000+ rows)	**19.18**	**0.8092**
Whisper Tiny (5000+ rows + Augmentation 2%)	30.49	0.7062

Table 11: Whisper Model on myMediTalk on Test Data

Table 11 presents the quantitative evaluation of different Whisper variants on the myMediTalk test dataset. Among the models trained with the same setup, Whisper Small achieved the lowest word error rate (WER) of 15.66 and the highest chrF++ score of 0.8695, showing a clear advantage over Whisper Tiny (WER 28.31, chrF++ 0.7145) and Whisper Medium (WER 24.96, chrF++ 0.7493). When trained with subset datasets from myMediTalk (5000+ rows), performance shifted: Whisper Small reached a WER of 19.18 with chrF++ 0.8092, while Whisper Tiny degraded slightly to WER 29.09 and chrF++ 0.7104. Adding augmentation (2%) to Whisper Tiny further increased performance (WER 30.49, chrF++ 0.7062). These results confirm that Whisper Small consistently provides more reliable performance than other Whisper models.

To better understand transcription quality, Table 12 provides sample outputs alongside the ground truth. Here, Whisper Small generally follows the grammatical and orthographic structures of Myanmar, but still produces critical errors such as replacing "စျာပန" with "စာပန", and "မြှုပ်" with "မျှုတ်". These substitutions indicate lexical confusion despite overall structural accuracy. In contrast, Whisper Tiny and Whisper Medium exhibit more severe deviations, such as producing invalid or nonsensical tokens ("ဇားပန", "စားပနား", "ဦး ကူး ကို ကို အောင် ကို အောင် ကူ ကို"). These outputs not only increase the WER but also reduce intelligibility. Augmented Whisper Tiny also showed instability, inserting unnatural token sequences ("ဦး နေ့ က ဦး စု", "သာ့း"), which highlights that naive augmentation may not generalize well in Myanmar transcription but more generalize than Whisper Tiny.

Taken together, the results suggest that Whisper Small provides the best balance between quantitative performance (WER, chrF++) and qualitative transcription fidelity, correctly handling Myanmar grammar in most cases. However, error analysis reveals that even Whisper Small struggles with certain phonetic confusions and semantic consistency. Whisper Tiny and Medium are more prone to grammar-breaking errors, while augmentation without careful design can worsen transcription quality. This indicates that scaling model size and training data is beneficial, but further adaptation, such as Myanmar-specific fine-tuning, is essential for robust performance in low-resource languages like Myanmar.

Model Approaches	Transcription Output
Ground Truth	ဆေး ရုံ က စ္စာ ပ နာ ကို စီ စဉ် နေ တယ် ဆို ရင် ကျွန် တော် တို့ က အုတ် ဂူ ကို အ မှတ် အ သား မ ပြူ လုပ် တဲ့ အ တွက် ခင် ဗျား က လေး ကို ဘယ် နေ ရာ မှာ မြှပ် လိုက် လဲ သိ မှာ မ ဟုတ် လို့ ကျေး ဇူး ပြု ပြီး အ သု ဘ ကို တက် ရောက် ပေး ပါ
Whisper Tiny	ဆေး ရုံ က ဇား ပ နာ ကို စီ စဉ် နေ တယ် ဆို ရင် ကျွန် တော် တို့ က အုပ် ဂူ ကူး ကိုင် ကူ ကို အ မှာ အ သာ မ ပြူ လုပ် တဲ့ အ တွက် က လေး ကို ဘယ် နေ့ ရာ မှ မျှပ် လိုက် လဲ သိ မှာ မ ဟုတ် ဘူး လို့ ကျေး ဇူး ပြု ပြီး အာ သူ ပါ ကို တက် ယောက် ပေး ပါ
Whisper Tiny + Augmentation	ဆေး ရုံ က စား ပ နာ ကို စီ စစ် နေ တဲ့ ဆို ရင် ကျွန် တော် တို့ က ဦး နေ့ က ဦး စု ကို အ မှတ် အ သာေျး ပြူ လုပ် တဲ့ အ တွက် ခင် ဗျား က လေး ကို ဘယ် နေ ရာ မှာ မျှံ့ လိုက် လဲ သိ မှာ မ ဟုတ် ဘူး လို
Whisper Small	ဆေး ရုံ က စာ ပ နာ ကို စီ စဉ် နေ တယ် ဆို ရင် ကျွန် တော် တို့ က အုပ် ကူ ကို အ မှတ် အ သား မ ပြူ လုပ် တဲ့ အ တွက် ခင် ဗျား က လေး ကို ဘယ် နေ ရာ မှာ မျှတ် လိုက် လဲ သိ မှာ မ ဟုတ် ဘူး လို့ ကျေး ဇူး ပြု ပြီး အား သူ ဘ ကို တက် ရောက် ပေး ပါ
Whisper Medium	ဆေး ရုံ က စား ပ နား ကို စီ စစ် နေ တယ် ဆို ရင် ကျွန် တော် တို့ က ဦး ကူး ကို ကို အောင် ကို အောင် ကူ ကို အ မှတ် အ သား မ ပြူ လုပ် တဲ့ အ တွက် ခင် ဗျား က လေး ကို ဘယ် နေ ရာ မှာ မျှပ် လိုက် လဲ သိ မှာ မ ဟုတ် ဘူး လို့ ကျေး ဇူး ပြု ပြီး အား သူး ပါ ကို တက် ယောက် ပေး ပါ

Table 12: Transcription outputs of Whisper Tiny, Whisper Tiny with Augmentation, Whisper Small, and Whisper Medium evaluated on myMediTalk.

6.2 Post-ASR Correction with mT5

System / Approach	WER	chrF++	Δ WER vs. Baseline	Δ chrF++ vs. Baseline
Baseline	71.87	0.4244	0.00	0.00
mT5 G2P	81.25	0.4645	+9.38	+0.0401
mT5 IPA	**64.51**	0.4633	-7.36	+0.0389
mT5	69.50	**0.4716**	-2.37	+0.0472

Table 13: Comparison of ASR performance across different systems and approaches.

We clearly observe differences in training and validation loss across mT5, mT5 + IPA, and mT5 + G2P as shown in Table 14. mT5 has the lowest train loss and mT5 + G2P has the lowest validation loss, mT5 + IPA outperformed with baseline WER 71.87% to the WER 64.51% even though mT5 wihout feature has 69.50% and mT5 + G2P features has WER 81.25% shown in Table 13.

The comparison in Table 15 shows that **Whisper Small** produces outputs that, at

Step	mT5		mT5 + IPA		mT5 + G2P	
	Train	Eval	Train	Eval	Train	Eval
1000	2.3435	1.7770	2.6164	1.9373	2.2943	1.9261
2000	1.3035	1.3365	1.3625	1.4363	1.4721	1.3720
3000	1.4303	1.0257	1.3797	1.1516	1.2927	1.0819
4000	1.0028	0.8861	1.1582	0.9059	1.0790	0.9013
5000	0.7064	0.7309	0.7019	0.7865	0.7341	0.7581
6000	0.6910	0.6389	0.9860	0.6710	0.4052	0.6628
7000	0.4623	0.5822	0.6672	0.5896	0.4125	0.5639
8000	0.3027	0.5213	0.3533	0.5421	0.3163	0.5109
9000	0.5235	0.4700	0.2428	0.4802	0.3782	0.4605
10000	0.1541	0.4519	0.2995	0.4491	0.2907	0.4406
11000	0.2460	0.4280	0.3351	0.4262	0.2025	0.4074
12000	0.2846	0.3927	0.1861	0.4139	0.1820	0.3902
13000	0.1243	0.3995	0.1382	0.3974	0.2088	0.3760
14000	**0.0891**	**0.4071**	**0.1363**	**0.4041**	**0.2028**	**0.3613**

Table 14: Comparison of training and validation loss at every 1000 steps across mT5, mT5 + IPA, and mT5 + G2P.

first glance, appear closer to the ground truth than the Post-ASR correction models. Its sentence structure looks more fluent and natural, which may give the impression of higher accuracy. However, a closer inspection reveals key issues.

Most Post-ASR correction outputs (**mT5** and **mT5+G2P**) fail to properly complete the Burmese sentence and truncate before reaching a natural sentence-ending particle. In contrast, **mT5+IPA** is the only Post-ASR correction system that generates a complete sentence, even though the lexical content differs slightly from the ground truth. This suggests that the IPA representation provides stronger guidance for sentence-level coherence.

Moreover, Whisper Small, while fluent, introduces substitution errors (e.g., using "တိုင် ရ" instead of "ဒဏ် ရာ"), which changes the meaning. In critical applications such as medical dialogue, these subtle substitution errors can lead to harmful misinterpretations, even if the sentence appears grammatically correct.

Model Approaches	
Ground Truth	ဘု ရင် မ ကြီး ရဲ့ ဒဏ် ရာ တွေ မှာ တွေ့ ရ တယ် ဗျ ။
Whisper Small	ဘု ရင် မာ ကြီး ရဲ့ တိုင် ရ တွေ မှာ တွေ့ ရ တွေ ပါ
mT5	ဘု ရင် မ ကြီး ရဲ့ ဒိုင် ယာ ရီ မှာ တွေ့ ရ တဲ့
mT5 + G2P	ဘု ရင် မ ကြီး ရဲ့ ဒိုင် ယာ တွေ မှာ တွေ့ ရ တဲ့
mT5 + IPA	ဘု ရင် မ ကြီး ရဲ့ အ တိုင်း ပဲ ရှိ တယ် ။

Table 15: Post-ASR Error Correction Output.

7 Conclusion & Future Work

As we move toward speech technologies to promote access and human–machine interaction, it is vital to understand how transfer learning approaches can assist low-resource languages such as Burmese. This study showed substantial improvements in transcription accuracy, in comparison to zero-shot baselines, by fine-tuning Whisper models on several datasets together with the newly compiled myMediTalk medical corpus. The magnitude of these improvements, however, was contingent upon model size and linguistic characteristics: *Whisper Small* consistently outperformed both *Whisper Tiny* and *MMS*, whereas post-ASR correction using *mT5* provided an additional boost when enriched with IPA representations. These findings suggest that large multilingual models hold promise for extending speech technologies to Burmese. Nevertheless, their effectiveness depends strongly on incorporating linguistic features and on the careful, deliberate design of training datasets.

In the future, our work points to several directions. Collecting more speech data that covers a wider range of speaker,ages and dialects will make the model more reliable and inclusive.It will also be important to test with newer multilingual models. On the post-ASR side, transfer-based approaches that fine-tune multilingual error-correction models specifically for Burmese hold a lot of promise. Taken together, these steps could help move Burmese ASR and TTS from research prototypes toward practical systems that perform well in the diverse, real-world contexts where they are most need.

References

[1] Tao Xu Greg Brockman Christine McLeavey Ilya Sutskever Alec Radford, Jong Wook Kim. Robust speech recognition via large-scale weak supervision. *arXiv preprint*, page 24, 2022.

[2] Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech, 05 2022.

[3] Xabier de Zuazo, Eva Navas, Bon Saratxaga, and Inma Hernaez Rioja. Whisper-lm: Improving asr models with language models for low-resource languages. *arXiv preprint*, pages 1–26, 2025.

[4] Alexandre Défossez. Hybrid spectrogram and waveform source separation. In *Proceedings of the ISMIR 2021 Workshop on Music Source Separation*, 2021.

[5] Hay Man Htun, Ye Kyaw Thu, Hutchatai Chanlekha, Kotaro Funakoshi, and Thepchai Supnithi. mymedicon: End-to-end burmese automatic speech recognition for medical conversations. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1–8. ELRA and ICCL, May 2024.

[6] Laet Laet Lin Khin Me Me Chit. Exploring ctc based end-to-end techniques for myanmar speech recognition. *International Conference on Intelligent Computing Optimization*, pages 1–9, 2020.

[7] Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang. Mosnet: Deep learning-based objective assessment for voice conversion. *INTERSPEECH*, pages 1–5, 2019.

[8] Edward Ma. Nlp augmentation. https://github.com/makcedward/nlpaug, 2019.

[9] Htet Htet Moe and Hay Mar Soe Naing. Childasr: Child automatic speech recognition for myanmar language. In *2025 IEEE Conference on Computer Applications (ICCA)*, pages 1–6, 2025.

[10] Aye Nyein Mon, Win Pa Pa, and Ye Kyaw Thu. Ucsy-sc1: A myanmar speech corpus for automatic speech recognition. *International Journal of Electrical and Computer Engineering (IJECE)*, 9(4):3194–3202, 2019.

[11] Yin May Oo, Theeraphol Wattanavekin, Chenfang Li, Pasindu De Silva, Supheakmungkol Sarin, Knot Pipatsrisawat, Martin Jansche, Oddur Kjartansson, and Alexander Gutkin. Burmese Speech Corpus, Finite-State Text Normalization and Pronunciation Grammars with an Application to Text-to-Speech. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*, pages 6328–6339, Marseille, France, May 2020. European Language Resources Association (ELRA).

[12] Arnaud Martin Siwar Jendoubi, Boutheina Ben Yaghlane. Belief hidden markov model for speech recognition. *arXiv preprint*, pages 1–5, 2015.

[13] Ye Kyaw Thu. sylbreak. `https://github.com/ye-kyaw-thu/sylbreak`, 2017.

[14] Ye Kyaw Thu, Win Pa Pa, Yoshinori Sagisaka, and Naoto Iwahashi. Comparison of grapheme-to-phoneme conversion methods on a myanmar pronunciation dictionary. In *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP)*, pages 11–22, Osaka, Japan, 2016. COLING.

[15] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. Scipy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, 17:261–272, 2020.

[16] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *arxiv*, pages 1–17, 2021.