ASR Error Correction in Low-Resource Burmese with Alignment-Enhanced Transformers using Phonetic Features

Ye Bhone Lin[1+], Thura Aung[1,2+], Ye Kyaw Thu[1,3*], Thazin Myint Oo[1]

[1]*Language Understanding Laboratory, Myanmar*
[2]*Department of Computer Engineering, KMITL, Bangkok, Thailand*
[3]*Language and Semantic Technology Research Team, NECTEC, Bangkok, Thailand*

Emails: yebhonelin10@gmail.com, 66011606@kmitl.ac.th, yekyaw.thu@nectec.or.th*, queenofthazin@gmail.com

***Abstract*—This paper investigates sequence-to-sequence Transformer models for automatic speech recognition (ASR) error correction in low-resource Burmese, focusing on different feature integration strategies including IPA and alignment information. To our knowledge, this is the first study addressing ASR error correction specifically for Burmese. We evaluate *five* ASR backbones and show that our ASR Error Correction (AEC) approaches consistently improve word- and character-level accuracy over baseline outputs. The proposed AEC model, combining IPA and alignment features, reduced the average WER of ASR models from 51.56 to 39.82 before augmentation (and 51.56 to 43.59 after augmentation) and improving chrF++ scores from 0.5864 to 0.627, demonstrating consistent gains over the baseline ASR outputs without AEC. Our results highlight the robustness of AEC and the importance of feature design for improving ASR outputs in low-resource settings.**

***Index Terms*—Burmese language, Automatic Speech Recognition, ASR Error Correction, IPA, Alignment, Transformer**

I. Introduction

Automatic Speech Recognition (ASR) systems often generate transcription errors, which can negatively affect downstream applications such as machine translation and information retrieval [1]. To address this, ASR Error Correction (AEC) has been proposed as a post-processing step to improve transcription quality without modifying the acoustic model [2]. Traditional AEC approaches relied on external language models for re-scoring ASR hypotheses [3], while more recent approaches explore Large Language Models (LLMs) for generative error correction, showing advantages over conventional language models [4].

End-to-end AEC methods based on Sequence-to-Sequence (S2S) architectures have gained popularity as they directly map erroneous transcripts to ground-truth text [5], [6]. Beyond text-only methods, some studies incorporate both acoustic features and ASR hypotheses to enable cross-modal AEC [7]–[9].

Recent advances highlight the effectiveness of S2S models; for example, [10] use a pre-trained S2S BART model as a denoising system to correct phonetic and spelling errors, fine-tuned on both synthetic and real ASR errors, and rescored using word-level alignments, achieving significant WER improvements on accented speech.

Several ASR datasets and systems have been developed for Burmese, including UCSY-SC1 [11] and large-vocabulary systems covering data collection, lexicon construction, and acoustic and language modeling [12]. MyanSpeech combines a CTC-Attention acoustic model with an RNN language model to improve alignment and recognition accuracy [13], while ChildASR targets primary-level students using a GMM-HMM model trained on 5 hours of child speech [14]. Additionally, myMediCon introduces a Burmese medical speech corpus and evaluates Transformer and RNN-based ASR models [15]. Despite having ASR datasets developed for Burmese, only two datasets are open-source corpora: OpenSLR80 [16] and FLEURS [17]. In the previous studies, phonetic information has also been shown to improve recognition performance in tonal languages such as Burmese [18].

Although existing systems improve acoustic modeling for low-resource Burmese, unlike post-OCR [19] and spelling error correction [20], ASR outputs still contain errors that can significantly affect downstream applications, and post-ASR error correction remains largely unexplored. To address this gap, we propose an alignment-guided Transformer Sequence-to-Sequence model with phonetic features based on the International Phonetic Alphabet (IPA) for low-resource Burmese AEC. Since training data is limited, we apply speech-level data augmentation and generate synthetic ASR errors from the augmented audio, rather than augmenting post-ASR text directly.

In this work, we present the AEC parallel corpus with data augmentation pipeline and a set of fine-tuned Whisper ASR models for Burmese general-domain speech

*Corresponding author.
+Equal contribution.

Preprint; Accepted to iSAI-NLP 2025

recognition. These resources aim to facilitate research and development in low-resource Burmese ASR and automatic error correction.

II. Dataset Preparation

We used two open-source Burmese speech corpora: OpenSLR80[1] [16] and FLEURS[2] [17]. Table I summarizes the amount of training and test data in hours. We also report the MOSNet scores for each dataset, which provide an automatic estimate of perceived speech quality. Both datasets achieve good MOSNet scores (4.06–4.14), indicating that the recordings are suitable for ASR.

Table I summarizes the amount of training and test data in hours. We also report the MOSNet [21] scores for each dataset, which provide an automatic estimate of perceived speech quality. To increase both the diversity and the amount of training data, we applied data augmentation to 10% of the training set for each method. Waveform-based augmentations included pitch shifting, speed perturbation, loudness adjustment, background noise addition, temporal shifting, and random cropping. Spectrogram-based augmentations included Vocal Tract Length Perturbation (VTLP) and time/frequency masking. These techniques enrich acoustic variability and improve model robustness. We applied augmentation only to the training split, keeping the test split intact. The augmentation process was done using `NLPAug`[3] Python library.

TABLE I: Speech Data Information

Dataset	Train Hr.	Test Hr.	MOSNet
OpenSLR80	3.70	0.42	4.06
FLEURS	15.95	1.64	4.14

For ASR, we experimented with Massively Multilingual Speech (MMS) [22] and Whisper [23] models in different sizes (tiny, small, medium, and large). Pretrained Whisper models performed poorly on Burmese, so we fine-tuned them using the available speech data to improve transcription quality. To generate ASR errors for AEC, we used both pretrained MMS and fine-tuned Whisper, applied on original as well as augmented data. Both ground-truth transcripts and ASR outputs (with errors) were segmented into syllables using `myWord`[4] [24]. Since ASR outputs often contained punctuation, unseen tokens, and special characters, we cleaned the text before feeding it into the error correction model.

Figure 1 (a) shows the process of dataset preparation for ASR Error Correction, and Table III shows the Post-ASR Dataset statistics: the number of sentences in parallel data and the total number of syllables (Syl) for original, augmented, and test sets, for both Err (ASR errors) and GT (ground truth).

[1]chuuhtetnaing/myanmar-speech-dataset-openslr-80
[2]chuuhtetnaing/myanmar-speech-dataset-google-fleurs
[3]https://github.com/makcedward/nlpaug
[4]https://github.com/ye-kyaw-thu/myWord

TABLE II: Pretrained ASR model information: model, number of parameters, and language coverage. * Whisper models were finetuned for Burmese in our work.

Model	Parameters	Languages
MMS-1B ASR	∼1B	1,100+
Whisper Tiny*	39M	99
Whisper Small*	244M	99
Whisper Medium*	769M	99
Whisper Large*	1.55B	99

TABLE III: Post-ASR Dataset statistics: number of sentences in parallel data and total number of syllables (Syl) for original, augmented, and test sets for both Err (error) and GT (groundtruth).

Dataset	Sentences	Err Syl	GT Syl
Original	31.7k	1.25M	1.22M
Augmented	55.9k	2.17M	2.19M
Test	3.19k	0.13M	0.12M

III. Methodology

A. Feature Extraction

a) IPA Features: We incorporated Phonetic (IPA-based) features to improve the training of the correction model. To extract IPA, Grapheme-to-IPA (G2IPA) conversion, we trained sequence tagging models using the myG2P word-level dictionary version 2.0 [25]. Among the methods tested, Conditional Random Fields (CRF) outperformed both Ripple Down Rule (RDR) and BiLSTM models, as shown in Table IV.

TABLE IV: Grapheme-to-IPA Conversion Results

	Precision	Recall	F1
RDR	0.84	0.84	0.85
CRF	**0.97**	**0.99**	**0.98**
BiLSTM	0.93	0.92	0.93

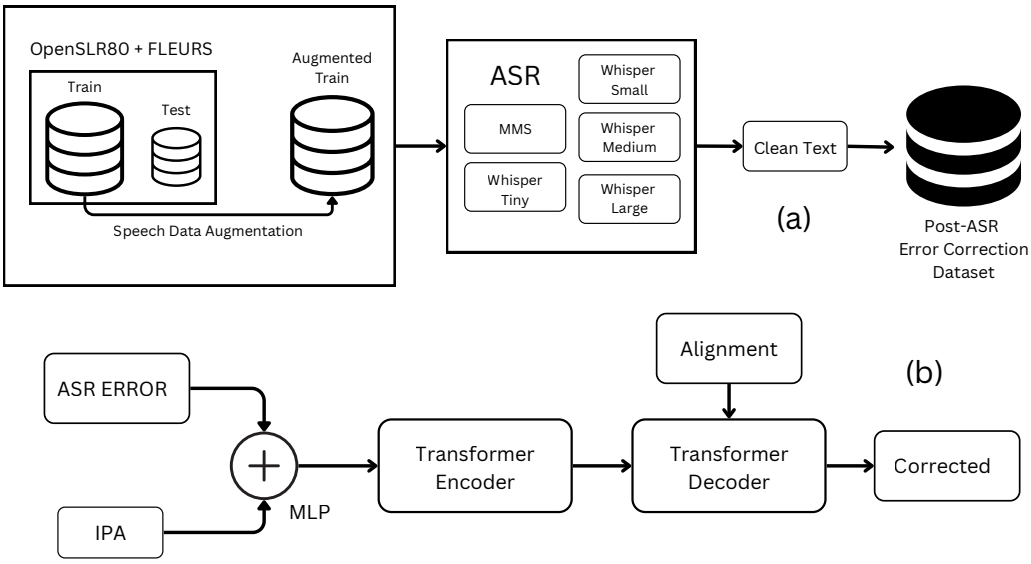
b) Alignment Feature: We used alignment to reduce hallucinations in the AEC models. For this purpose, we employed fast-align[5] [26], a log-linear re-parameterization of IBM Model 2 that addresses the strong assumptions of Model 1 and the over-parameterization of Model 2. Fast-align provides efficient inference, likelihood evaluation, and parameter estimation, and is consistently faster than IBM Model 4. Its high-quality alignments have been shown to improve downstream tasks, making it well-suited for guiding sequence-to-sequence AEC models.

B. Feature Integration

For phonetic feature integration, IPA features were embedded and combined with word embeddings through

[5]http://github.com/clab/fast_align

Fig. 1: (a) Post-ASR Dataset Preparation for ASR Error Correction (AEC) Training (b) Integration of Phonetic Features and Alignment in AEC.



a multi-layer perceptron (MLP), allowing the Transformer to jointly leverage orthographic and phonetic representations. For alignment integration, we employed the alignment-assisted Neural Machine Translation (NMT) mechanism [27], where fast-align outputs were used as external constraints on the Transformer decoder's attention distributions. Figure 1 (b) shows the process of correcting ASR error using Phonetic (IPA) features and alignment information.

During training, these alignments guided and regularized attention weights, reducing spurious insertions or omissions and mitigating hallucinations. This supervision ensured that the model remained faithful to source–target correspondences while benefiting from the contextual modeling capacity of the Transformer.

C. Sequence-to-sequence Transformer

We adopted the Transformer-based S2S architecture, in which, the encoder captures contextual representations of the input sequence through multi-head self-attention, while the decoder generates the corrected output sequence by attending to both the encoder states and previously generated tokens. This architecture is well-suited for AEC as it balances fluency and contextual modeling with parallelizable training. We trained a S2S Transformer model under different settings: with and without phonetic (IPA-based) features and, with and without alignment information. Training was conducted on both the original and the augmented datasets, and we compared model performance across these configurations.

IV. Experimental Setup

For the ASR part, we used `transformer` library for Whisper finetuning and MMS pretraining inferencing. We used the `OpenNMT` toolkit [28] for training the sequence-to-sequence Transformer models. CRF-suite was employed for grapheme-to-IPA (G2IPA) conversion, while fast-align was used for alignment extraction.

For the Transformer architecture, both the encoder and decoder were configured with 4 layers, a hidden size of 512, and 512-dimensional word embeddings. The feed-forward network size was set to 2048 with 8 attention heads. Position encoding was enabled, and regularization included dropout (0.3) and attention dropout (0.3). Label smoothing was set to 0.1.

Training was performed with the Adam optimizer and the Noam learning rate schedule, with a learning rate of 0.1, 5,000 warm-up steps, and gradient accumulation of 4 steps. The maximum source and target sequence lengths were both limited to 200 tokens. The batch size was 64 (token-based normalization), and early stopping was applied with a patience of 4 validation checks to avoid overfitting. Models were trained for up to 200,000 steps, and the best 10 retained.

To assess the performance of our Transformer S2S based AEC models, we employed two widely used metrics: Word Error Rate (WER) and Character F-score (chrF++) [29]. WER captures the proportion of word-level substitutions, deletions, and insertions between the system output and the reference, making it a standard measure for correction accuracy at the lexical level. chrF++ evaluates similarity at the character n-gram level, which is particularly useful for morphologically rich languages and for capturing finer-

grained orthographic variations. Together, these metrics provide a balanced evaluation of both word-level accuracy and subword/character-level fidelity.

V. Result and Discussion

A. Across different AEC approaches

Across all configurations, every AEC-based approach consistently outperforms the baseline ASR output without AEC, confirming the effectiveness of error correction in improving recognition accuracy. It reduces average WER of ASR models (ASR: 51.56 → AEC: 37.07) and improves average chrF++ scores (ASR: 0.5864 → AEC: 0.638). The results show that WER generally increased after augmentation across all feature setups (e.g., AEC: 37.07 → 40.83; AEC + Alignment: 37.12 → 41.09), reflecting the added difficulty of word-level prediction under augmented variability. Crucially, however, all AEC-based approaches (with or without augmentation) still outperform the baseline ASR output without AEC, confirming the robustness of error correction despite the distribution mismatch introduced by augmentation. Interestingly, chrF++ demonstrates a mixed effect. For some setups, such as AEC + Alignment + IPA, chrF++ increased after augmentation (0.6044 → 0.6266), suggesting that the augmented data helped the models produce outputs with better character-level overlap with references. However, other setups such as AEC + IPA experienced a drop (0.6377 → 0.5936). Overall, the averaged chrF++ indicates that augmentation can enhance surface-level similarity in certain configurations, particularly when alignment features are included. These findings suggest that while data augmentation may slightly degrade word-level accuracy, it can improve subword or character-level similarity in specific feature configurations, especially those leveraging alignment-based representations.

B. Without Augmentation

We first evaluated the ASR error correction models trained on the original dataset solely, without any augmented data. Across all ASR models, applying AEC yielded consistent improvements over the baseline (no AEC), confirming the robustness of error correction. Table V reports the WER and chrF++ scores for each feature configuration across the different ASR models. The results show that AEC consistently improves performance over the ASR models without AEC. For example, MMS sees a reduction in WER from 42.27 to 30.70 and an increase in chrF++ from 0.6126 to 0.6461 when AEC is applied. Incorporating alignment-based features further improves performance, yielding the best overall results for MMS with a WER of 30.21 and chrF++ of 0.6940. IPA features have a mixed effect: they improve chrF++ for some models (e.g., MMS) but can slightly increase WER for others. Among the Whisper models, the Small and Large variants achieve the lowest WER and highest chrF++ with AEC and alignment, demonstrating that error correction and alignment features are particularly effective for stronger ASR backbones. These findings indicate that training on the original clean data is sufficient for large models to achieve high accuracy, and careful feature selection (AEC and alignment) can maximize both word-level and character-level performance.

C. Effect of Data Augmentation

We investigated the impact of augmenting the training data on ASR error correction across multiple ASR models. Table V shows the WER and chrF++ scores for models trained on original data versus those trained with additional augmented data, evaluated on a fixed clean test set. Overall, the inclusion of augmented data consistently increased WER for all ASR models, indicating that the augmentation introduced variability that shifted the training distribution away from the clean test distribution. This distribution mismatch caused the models to perform worse at exact word-level recognition. In terms of chrF++, the effect of augmentation was more nuanced. For the smallest model, Whisper Tiny, chrF++ improved (0.5871 → 0.6245), suggesting that the augmented data helped the model generate outputs with higher character-level similarity to the references, even if full words were not always correct. For larger models (Whisper Small–Large, MMS), augmentation generally reduced chrF++, implying that these models, already strong on the original data, were negatively affected by the added variability. These results highlight that while augmentation can help under-parameterized models by providing additional variability, it may hinder performance for stronger models when evaluated on fixed clean data. Therefore, careful consideration is required when designing augmentation strategies, especially for high-performing ASR systems.

VI. Error Analysis

ASR Error Analysis Table VI presents the recognition outputs of different ASR models compared against the ground truth. The analysis shows that all models capture the general sentence structure, but frequent misrecognitions appear at the word and syllable level. For example, Whisper Medium and Large often substitute semantically related but phonetically mismatched words. This indicates that larger models, despite their improved fluency, tend to hallucinate words when uncertain. Smaller models such as Whisper Tiny exhibit more severe lexical errors with unrelated segments. MMS shows inconsistent syllable alignment, producing non-standard transcriptions such as "ဘခ|a- စတ်|ho" or repeated tokens ("a|x a|x"), reflecting unstable decoding on Myanmar text, which could hurt G2IPA conversion performance as well.

AEC Error Analysis Table VII compares different AEC approaches applied to Whisper Tiny output. The baseline Whisper Tiny transcription contains significant errors, e.g., "ယ|ja-" for "ဧရ|jei" and "ချောင်း|chaun:" for

Fig. 2: Average WER and chrF++ Score Across Different AEC Approaches Before and After Data Augmentation.

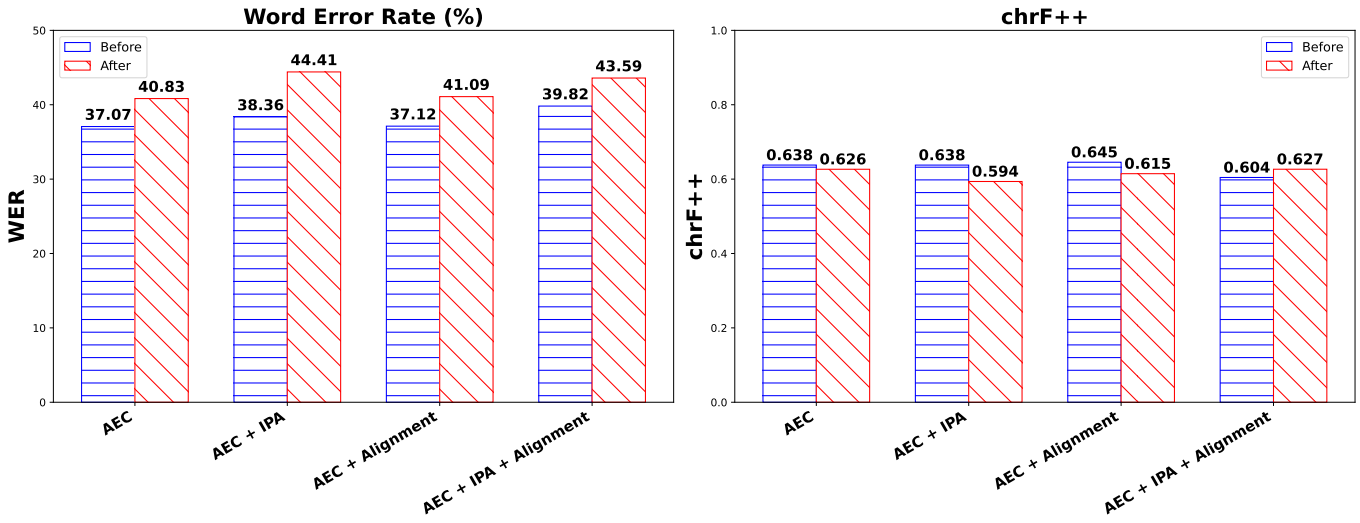


TABLE V: Comparison of ASR Error Correction models trained with Different Features on both Original and Augmented Data Across Different ASR systems (Best results per ASR model in **Bold**).

Original Data				+ Augmentation Data			
ASR Model	**Feature**	**WER**	**chrF++**	**ASR Model**	**Feature**	**WER**	**chrF++**
MMS	No AEC (Baseline)	42.27	0.6126	MMS	No AEC (Baseline)	42.27	0.6126
	+ AEC	30.70	0.6461		+ AEC	**33.76**	**0.6670**
	+ AEC + IPA	31.40	0.6923		+ AEC + IPA	38.85	0.6331
	+ AEC + Align	**30.21**	**0.6940**		+ AEC + Align	34.07	0.6646
	+ AEC + IPA + Align	35.03	0.6596		+ AEC + IPA + Align	36.57	0.6605
Whisper Tiny	No AEC (Baseline)	55.07	0.5483	Whisper Tiny	No AEC (Baseline)	55.07	0.5483
	+ AEC	**43.79**	**0.5871**		+ AEC	**48.28**	**0.6245**
	+ AEC + IPA	45.21	0.5797		+ AEC + IPA	52.00	0.5361
	+ AEC + Align	44.17	0.5868		+ AEC + Align	48.43	0.5546
	+ AEC + IPA + Align	45.48	0.5463		+ AEC + IPA + Align	51.37	0.5816
Whisper Small	No AEC (Baseline)	37.25	0.6534	Whisper Small	No AEC (Baseline)	37.25	0.6534
	+ AEC	**32.57**	0.6745		+ AEC	**36.22**	0.6302
	+ AEC + IPA	33.58	0.6709		+ AEC + IPA	38.57	0.6329
	+ AEC + Align	32.71	**0.6763**		+ AEC + Align	36.27	0.6484
	+ AEC + IPA + Align	33.71	0.6338		+ AEC + IPA + Align	38.23	**0.6727**
Whisper Medium	No AEC (Baseline)	72.18	0.5425	Whisper Medium	No AEC (Baseline)	72.18	0.5425
	+ AEC	**41.88**	0.6148		+ AEC	**46.01**	0.5884
	+ AEC + IPA	43.50	0.6030		+ AEC + IPA	49.41	0.5629
	+ AEC + Align	42.05	**0.6169**		+ AEC + Align	46.51	0.5844
	+ AEC + IPA + Align	43.21	0.5616		+ AEC + IPA + Align	49.86	**0.6041**
Whisper Large	No AEC (Baseline)	51.04	0.5752	Whisper Large	No AEC (Baseline)	51.04	0.5752
	+ AEC	**36.38**	**0.6666**		+ AEC	39.87	**0.6225**
	+ AEC + IPA	38.08	0.6425		+ AEC + IPA	43.21	0.6028
	+ AEC + Align	36.47	0.6527		+ AEC + Align	**40.17**	0.6211
	+ AEC + IPA + Align	41.66	0.6205		+ AEC + IPA + Align	41.94	0.6141

"ဧကြာင်း|gyaun:." These lexical substitutions break semantic consistency. Applying AEC substantially reduces such errors, recovering correct forms of keywords and improving alignment with the ground truth. The plain AEC approach fixes major errors but sometimes introduces new ones, such as replacing "မှ|mha." with "မှာ|mha" or ending with "တော့|do." instead of "သော|tho:". The addition of IPA information (+AEC +IPA) enhances phonetic matching, producing almost identical output to the ground truth. Alignment constraints (+AEC +Align) further stabilize word order but may still insert wrong tokens like "ဖို|pji.".

VII. Conclusion and Future Work

This work investigated ASR error correction (AEC) for low-resource Burmese across different feature configurations and training setups. Our findings show that all AEC-

TABLE VI: Error Comparison of Different ASR Output (Errors are colored in Red).

ASR Models	Sentence
Groundtruth	စိတ်|sei' ဝင်|win| စား|za: ဖွယ်|bwe ကောင်း|gaun: သော|tho: ရွာ|jwa သို့|dhou. အေး|ei: အေး|ei: လူ|lu လူ|lu ဖြင့်|hpjin. နာ|na ရီ|ji ဝက်|we' ခန့်|khan. လမ်း|lan: လျှောက်|shau' သွား|dhwa: ရ|ra— သည်|dhe
MMS	အ|a- စတ်|ho ဝင်|win စာ|za ဘဲ|be: ကောင်|gaun တော်|do ရွာ|jwa သို့|dhou. a|x a|x လူ|lu ဖြ|bu. င့်|bu. နိုင်|bu. ဝက်|we' ခန့်|khan. လမ်း|lan: လျှောက်|shau' သွား|dhwa: ရ|ra- သည်|dhe
Whisper Tiny	ဆိုက်|hsai' ဝင်|win ကြ|kya- ပဲ|pe: ကောင်း|kaun: လော်|lo ရာ|ja သို့|dhou. Aိတ်|x အေး|ei: ဘီ|bi လူ|lu ဖြင့်|hpjin. မ|ma- ဟုတ်|hou' ပွဲ|pwe: ခံ|gan နဲ့|ne: ရှောက်|shau' ဘွား|hpwa: ရာ|ja သည်|dhi
Whisper Small	စိတ်|sei' ဝင်|win စာ|sa- ဘက်|be' ကောင်း|kaun: လောင်|laun ရွာ|jwa သို့|dhou. အေး|ei: အေး|ei: လူ|lu ဖြင့်|hpjin. နာ|na ရီ|ji ဝက်|we' ခန့်|khan. လမ်း|lan: လျှောက်|shau' သွား|dhwa: ရ|ra- သည်|dhe
Whisper Medium	စိတ်|sei' ဝင်|win စား|za: ဝဲ|we: ကောင်း|gaun: လော်|lo ရာ|ja သူး|htei အင်|in အီ|i ယူ|ju ဖြင့်|hpjin. နိုင်|nain ဝက်|we' ခံ|gan နှင်း|nhin: ရှောက်|shau' သွား|thwa: ရ|ja. လေ|lei
Whisper Large	စိတ်|sei' ဝင်|win စာ|sa ပဲ|be: ကောင်း|gaun: လော်|lo ရ|ra' သူ|dhu ကျင်|gyin အေ|ei ဗျူ|hta ဖြင့်|hpjin. နိုင်|nain ဝက်|we' ခံ|khan နှံ|nhan. လျှောက်|shau' သွား|thwa: ရ|ja. လေ|lei

TABLE VII: Error Comparison of Different AEC approaches for Whisper Tiny Model (Errors are colored in Red).

AEC Approaches	Sentence
Correct	ရေ|jei စီး|zi: ကြောင်း|gyaun: မှ|mha. ထွက်|htwe' လာ|la သည်|thi နှင့်|ne. ရေ|jei ကူး|gu: ခြင်း|gyin: သည်|dhe ပုံ|boun မှန်|mhan အား|a: ဖြင့်|hpjin. မ|ma- ခက်|khe' ခဲ|ge: တော့|do. ပါ|ba
Whisper Tiny	ယ|ja- စီး|si: ချောင်း|chaun: မှ|mha. ထွက်|htwe' လာ|la သည်|dhe နှင့်|ne. ရေ|jei ကို|kou ချင်|gyo သည်|thi ပုံ|boun မှန်|mhan အား|a: ပြင်|bjin မ|ma- ချက်|che' ခဲ့|shoun: တော့|to. ပါ|ba
+ AEC	ရေ|jei စီး|zi: ကြောင်း|gyaun: မှာ|mha ထွက်|dwe' လာ|la သည်|dhe နှင့်|ne. ရေ|jei ကူး|gu: ခြင်း|gyin: သည်|dhe ပုံ|boun မှန်|mhan အား|a: ဖြင့်|hpjin. မ|ma- ခက်|khe' ခဲ|ge: သော|tho: ပါ|ba
+ AEC + IPA	ရေ|jei စီး|zi: ကြောင်း|gyaun: မှ|mha. ထွက်|htwe' လာ|la သည်|thi နှင့်|ne. ရေ|jei ကူး|gu: ခြင်း|gyin: သည်|dhe ပုံ|boun မှန်|mhan အား|a: ဖြင့်|hpjin. မ|ma- ခက်|khe' ခဲ|ge: တော့|do. ပါ|ba
+ AEC + Align	ရေ|jei စီး|zi: ကြောင်း|gyaun: မှာ|mha ထွက်|dwe' လာ|la ပြီ|pji နှင့်|ne. ရေ|jei ကူး|gu: ခြင်း|gyin: သည်|dhe ပုံ|boun မှန်|mhan အား|a: ဖြင့်|hpjin. မ|ma- ခက်|khe' ခဲ|ge: သော|tho: ပါ|ba
+ AEC + IPA + Align	ရေ|jei စီး|zi: ကြောင်း|gyaun: မှ|mha. ထွက်|htwe' လာ|la သည်|thi နှင့်|ne. ရေ|jei ကူး|gu: ခြင်း|gyin: သည်|dhe ပုံ|boun မှန်|mhan အား|a: ဖြင့်|hpjin. မ|ma- ခက်|khe' ခဲ|ge: တော့|do. ပါ|ba

based approaches, regardless of augmentation, consistently outperform the baseline ASR outputs without AEC, highlighting the effectiveness of error correction in improving recognition quality. Alignment features provided the strongest gains, especially when combined with AEC, while IPA features yielded mixed effects depending on the backbone ASR model. Data augmentation was found to introduce variability that increased WER on clean test data, although it sometimes improved character-level similarity (chrF++) for smaller models. This suggests that augmentation can be beneficial in under-parameterized scenarios but may lead to distribution mismatch for stronger ASR backbones. Overall, these results underscore the robustness of AEC and the importance of careful feature design. Future work will explore augmentation strategies better matched to test conditions, as well as extending AEC to multimodal scenarios [7] and integrating large language models [30] for enhanced correction capabilities.

For future work, we plan to release the fine-tuned Whisper models and AEC models, including `OpenNMT` configuration files, along with the parallel corpus of ASR outputs, ground-truth and their corresponding AEC-corrected transcripts, to support further study in low-resource Burmese ASR error correction.

References

[1] A. Kannan, Y. Wu, P. Nguyen, T. N. Sainath, Z. Chen, and R. Prabhavalkar, "An analysis of incorporating an external language model into a sequence-to-sequence model," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5828–5832.

[2] S. S. Sodhi, E. K.-I. Chio, A. Jash, S. Ontañón, A. Apte, A. Kumar, A. Jeje, D. Kuzmin, H. Fung, H.-T. Cheng *et al.*, "Mondegreen: A post-processing solution to speech recognition error correction for voice search queries," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. ACM, 2021, pp. 3569–3575.

[3] T. Tanaka, R. Masumura, H. Masataki, and Y. Aono, "Neural error corrective language models for automatic speech recognition," in *Proceedings of Interspeech*, 2018, pp. 401–405.

[4] C.-H. H. Yang, Y. Gu, Y.-C. Liu, S. Ghosh, I. Bulyko, and A. Stolcke, "Generative speech recognition error correction with large language models," *arXiv preprint arXiv:2309.15649*, 2023.

[5] A. Mani, S. Palaskar, N. V. Meripo, S. Konam, and F. Metze, "Asr error correction and domain adaptation using machine translation," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6344–6348.

[6] J. Liao, S. Eskimez, L. Lu, Y. Shi, M. Gong, L. Shou, H. Qu, and M. Zeng, "Improving readability for automatic speech recognition transcription," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 5, pp. 1–23, 2023.

[7] B. Mu, Y. Li, Q. Shao, K. Wei, X. Wan, N. Zheng, H. Zhou, and L. Xie, "Mmger: Multi-modal and multi-granularity generative error correction with llm for joint accent and speech recognition," 2024. [Online]. Available: https://arxiv.org/abs/2405.03152

[8] S. Radhakrishnan, C.-H. H. Yang, S. A. Khan, R. Kumar, N. A. Kiani, D. Gomez-Cabrero, and J. Tegnér, "Whispering llama: A cross-modal generative error correction framework for speech recognition," in *Proceedings of the 2023 Conference on*

Empirical Methods in Natural Language Processing (EMNLP), 2023.

[9] C. Chen, R. Li, Y. Hu, S. M. Siniscalchi, P.-Y. Chen, E. Chng, and C.-H. H. Yang, "It's never too late: Fusing acoustic information into large language models for automatic speech recognition," *arXiv preprint arXiv:2402.05457*, 2024.

[10] S. Dutta, S. Jain, A. Maheshwari, S. Pal, G. Ramakrishnan, and P. Jyothi, "Error correction in asr using sequence-to-sequence models," *arXiv preprint arXiv:2202.01157*, 2022.

[11] A. N. Mon, W. P. Pa, and Y. K. Thu, "Ucsy-sc1: A myanmar speech corpus for automatic speech recognition," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 4, pp. 3194–3202, 2019.

[12] H. M. S. Naing, A. M. Hlaing, W. P. Pa, X. Hu, Y. K. Thu, C. Hori, and H. Kawai, "A myanmar large vocabulary continuous speech recognition system," in *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2015, pp. 320–327.

[13] S. S. Su Yee, W. Lai Lai Phyu, W. P. Pa, M. Myint Yain, S. W. Maw, and N. Chan, "Myanspeech: Joint ctc-attention and rnn language model for end-to-end read speech recognition," in *2025 IEEE Conference on Computer Applications (ICCA)*, 2025, pp. 1–8.

[14] H. H. Moe and H. Mar Soe Naing, "Childasr: Child automatic speech recognition for myanmar language," in *2025 IEEE Conference on Computer Applications (ICCA)*, 2025, pp. 1–6.

[15] H. M. Htun, Y. Kyaw Thu, H. Chanlekha, K. Funakoshi, and T. Supnithi, "myMediCon: End-to-end Burmese automatic speech recognition for medical conversations," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Torino, Italia: ELRA and ICCL, May 2024, pp. 12 032–12 039. [Online]. Available: https://aclanthology.org/2024.lrec-main.1051/

[16] Y. M. Oo, T. Wattanavekin, C. Li, P. De Silva, S. Sarin, K. Pipatsrisawat, M. Jansche, O. Kjartansson, and A. 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)*. Marseille, France: European Language Resources Association (ELRA), May 2020, pp. 6328–6339. [Online]. Available: https://www.aclweb.org/anthology/2020.lrec-1.777

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

[18] A. N. Mon, W. P. Pa, and Y. K. Thu, "Exploring the effect of tones for myanmar language speech recognition using convolutional neural network (cnn)," in *Proceedings of the 15th International Conference of the Pacific Association for Computational Linguistics (PACLING)*, Yangon, Myanmar, 2017, pp. 334–345.

[19] T. Aung, Y. K. Thu, and M. N. Oo, "myocr: Optical character recognition for myanmar language with post-ocr error correction," in *2024 19th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, 2024, pp. 1–6.

[20] E. T. Phyu, Y. K. Thu, H. Chanlekha, K. Funakoshi, and T. Supnithi, "Exploring the impact of error type features integration on transformer-based myanmar spelling correction," in *2024 21st International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 2024, pp. 365–372.

[21] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, "Mosnet: Deep learning based objective assessment for voice conversion," in *Proc. Interspeech 2019*, 2019.

[22] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi, A. Baevski, Y. Adi, X. Zhang, W.-N. Hsu, A. Conneau, and M. Auli, "Scaling speech technology to 1,000+ languages," *arXiv*, 2023.

[23] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356

[24] Y. K. Thu, "myWord: Syllable, Word and Phrase Segmenter for Burmese," https://github.com/ye-kyaw-thu/myWord, Sep. 2021, [Online; accessed 13-Jun-2024].

[25] H. Htun, N. H. Aung, S. S. Moe, W. T. Zaw, N. N. Oo, T. Supnithi, and Y. K. Thu, "Grapheme-to-ipa phoneme conversion for burmese (myg2p version 2.0)," *Journal of Intelligent Informatics and Smart Technology*, vol. 1, no. 1, April 2021.

[26] C. Dyer, V. Chahuneau, and N. A. Smith, "A simple, fast, and effective reparameterization of IBM model 2," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, L. Vanderwende, H. Daumé III, and K. Kirchhoff, Eds. Atlanta, Georgia: Association for Computational Linguistics, Jun. 2013, pp. 644–648. [Online]. Available: https://aclanthology.org/N13-1073/

[27] H. Xi, F. Zhang, and Y. Wang, "Transformer machine translation model incorporating word alignment structure," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 1010–1019, 2024.

[28] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. Rush, "OpenNMT: Open-source toolkit for neural machine translation," in *Proceedings of ACL 2017, System Demonstrations*, M. Bansal and H. Ji, Eds. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 67–72. [Online]. Available: https://aclanthology.org/P17-4012/

[29] M. Popović, "chrF: character n-gram F-score for automatic MT evaluation," in *Proceedings of the Tenth Workshop on Statistical Machine Translation*, O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, C. Hokamp, M. Huck, V. Logacheva, and P. Pecina, Eds. Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 392–395. [Online]. Available: https://aclanthology.org/W15-3049/

[30] R. Sachdev, Z.-Q. Wang, and C.-H. H. Yang, "Evolutionary prompt design for llm-based post-asr error correction," 2024. [Online]. Available: https://arxiv.org/abs/2407.16370