

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

Ye Bhone Lin, Thura Aung, Ye Kyaw Thu*, Thazin Myint Oo



OUTLINES

1. Introduction

5. Result and Discussion

2. Dataset Preparation

6. Error Analysis

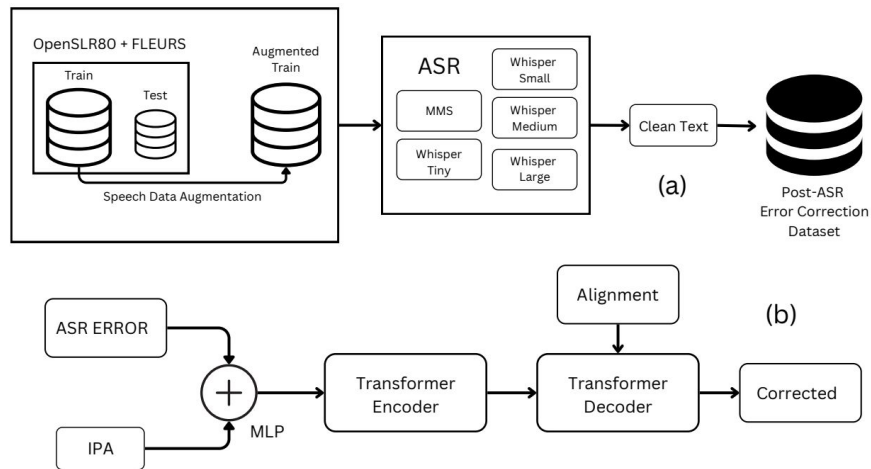
3. Materials

7. Future Work

4. Experimental Setup

8. Conclusion

1. Introduction



- Automatic Speech Recognition (ASR) systems transcribe spoken language into text.
- However, they often produce **errors**, especially in **low-resource languages** like Burmese, due to limited training data.
- These transcription errors degrade performance in downstream tasks such as translation or information retrieval.

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

1. Introduction (Cont'd)

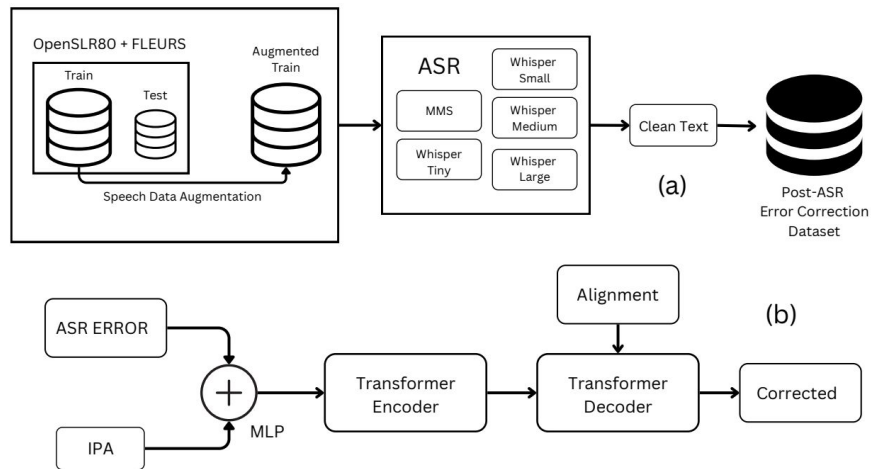


Fig. 1: (a) Post-ASR Dataset Preparation for ASR Error Correction (AEC) Training (b) Integration of Phonetic Features and Alignment in AEC.

- Recent Scientists, Fine-tuned BART model to correct ASR phonetic and spelling errors, improving WER on accented speech.
- The study proposes an **Alignment-Enhanced Transformer** that integrates **phonetic (IPA)** and **alignment** features for **ASR Error Correction (AEC)**.
- This is the **first work** to address Burmese ASR error correction with deep alignment and phonetic integration.

2. Dataset Preparation

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

TABLE I: Speech Data Information

Two main speech Open-Source corpora were used:

- **OpenSLR80:** 3.7 hours of training data, 0.42 hours of test data.
- **FLEURS:** 15.9 hours of training data, 1.64 hours of test data.

2. Dataset Preparation (Cont'd)

Data were **augmented (10%)** with:

- Pitch shift, speed perturbation, background noise,
- Vocal Tract Length Perturbation (VTLP), and
- Time and frequency masking.

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

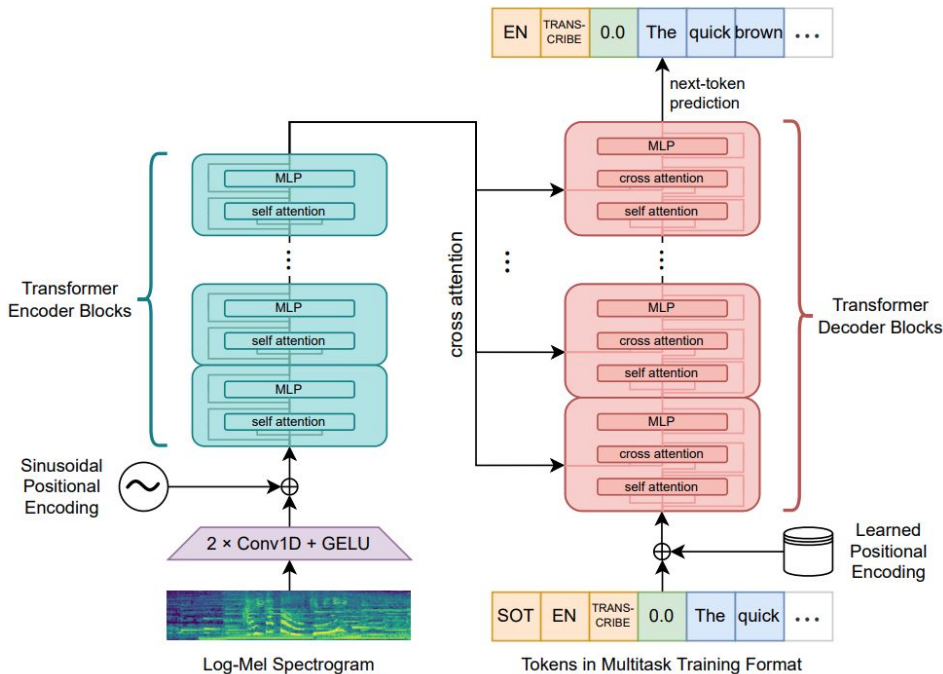
TABLE II: Post-ASR Dataset statistics: number of sentences in parallel data and total number of syllables (Syl) for original, augmented, and test sets for both Err (error) and GT (groundtruth).

Data were segmented into syllables using **myWord**, ensuring linguistic consistency between ASR output and target text.

2. Dataset Preparation (Cont'd)

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: Pretrained ASR model information: model, number of parameters, and language coverage. * Whisper models were fine tuned for Burmese



2. Dataset Preparation (Cont'd)

- Fine-tuned **Whisper models** — *Tiny, Small, Medium, and Large* on Burmese speech data to generate ASR outputs.
- Applied **Zero-shot learning** using **Meta MMS** to extend coverage for unseen speech patterns.
- Combined ASR hypotheses and ground truth transcriptions to build a **parallel dataset** for **ASR Error Correction (AEC)**.
- The resulting dataset contains aligned **error–correction pairs**, serving as input for AEC model training.

3. Materials

Framework: OpenNMT (Transformer sequence-to-sequence model).

ASR Sources: Whisper (Tiny, Medium, Large) and Meta MMS.

Feature Extraction Tools:

- **Phonetic (IPA):** myG2P v2.0 (CRF-based grapheme-to-phoneme conversion).
- **Alignment:** fast-align tool (for bilingual alignment scoring).

Hardware: NVIDIA GPU with 24GB VRAM, PyTorch backend.

No.	Words	Standard	Correct
1	သတင်းစာ (newspaper)	tha. tin: sa	dha- din: za
2	ပိုင်းခွဲ (denominator)	pain: chei	pain: gyei
3	ပုလဲ (pearl)	pu. le:	pa- le:
4	ပညာ (knowledge)	pa. nja	pjin nja
5	မင်းကြောင် (tatoo)	min kyaun	mhin gyaun

TABLE IV: Examples of contextually dependent pronunciations of some Myanmar words

4. Experimental Setup

Model Architecture: 4-layer Transformer encoder-decoder (512 hidden units, 8 heads).

Optimization: Adam optimizer with Noam learning rate scheduler.

Batch size: 4096 tokens.

Metrics:

- **Word Error Rate (WER)** — measures overall transcription accuracy.
- **Character F-score (chrF++)** — measures character-level fluency and correctness.

Baseline Models: Whisper Tiny, Medium, Large and MMS ASR outputs.

5. Results and Discussion

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

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)

5. Results and Discussion (Cont'd)

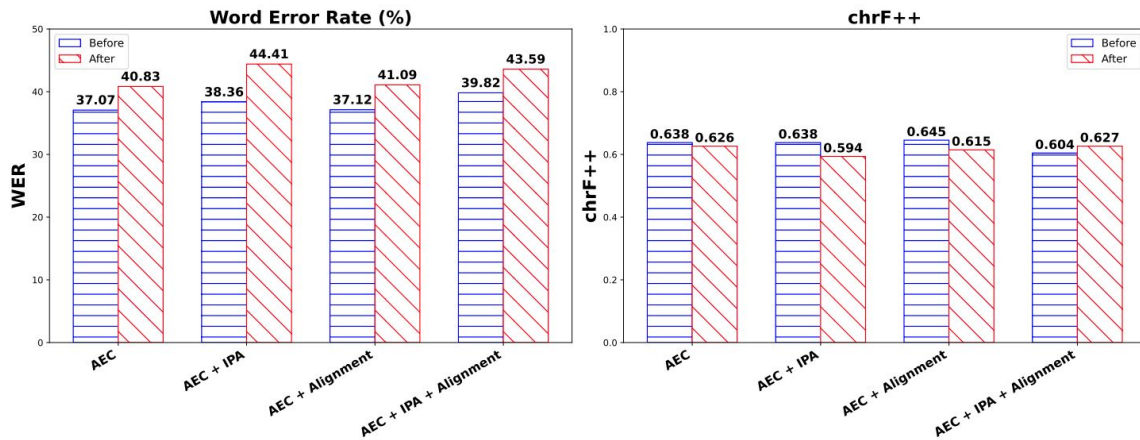


Fig. 2: Average WER and chrF++ Score Across Different AEC Approaches Before and After Data Augmentation.

- The **Word Error Rate (WER)** increased slightly after data augmentation across all AEC configurations.
- The **chrF++ scores** showed minor variations, with the *AEC + IPA + Alignment* approach showing a small improvement.
- Results suggest that data augmentation introduced some noise at the word level but helped maintain or improve **character-level consistency** in some models.
- The results validate that **linguistic-informed AEC models** can meaningfully enhance low-resource ASR systems without retraining acoustic models.

6. Error Analysis

1. ASR Error Analysis

All ASR models capture the general sentence structure but differ in accuracy at the **word and syllable level**.

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: ကောင်း gaun တော် do ရွာ jwa သို့ dhou. a x a x လူ lu ဖြင့် bu. င် 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. အိတ် x အေး ei: ဘီ bi လူ lu ဖြင့် hpjin. မ ma- ဟုတ် hou' ပွဲ pwe: ခဲ gan နဲ ne: ရှောက် shau' သွား 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' ခဲ gan နှင်း nhin: ရှောက် shau' သွား thwa: ရ ja. လေ lei
Whisper Large	စိတ် sei' ဝင် win စာ sa ဝဲ be: ကောင်း gaun: လော် lo ရာ ra' သူ dhu ကျင် gyin အေ ei ဂျူ hta ဖြင့် hpjin. နိုင် nain ဝက် we' ခဲ khan နှံ nhan. လျှောက် shau' သွား thwa: ရ ja. လေ lei

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

- **Whisper Medium and Large** produce fluent sentences but sometimes **hallucinate or substitute** words that sound similar.
- **Whisper Tiny and MMS** has **unstable decoding**, producing non-standard or repeated syllables such as “အ|a- စတ်|ho” or “a|x a|x”.

6. Error Analysis (Cont'd)

2. AEC Error Analysis

Applying AEC **reduces major lexical and syllable errors**, recovering correct forms closer to the ground truth.

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' ခဲ 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

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

- The **plain AEC** model corrects most key tokens but may introduce minor replacements (e.g., “မှ|mha” → “မှာ|mha”).
- **AEC + IPA** improves phonetic alignment, producing nearly identical output to the reference.
- **AEC + Align** enhances word order and structure but can still insert minor errors like “ပြီ|pji”.

7. Future Work

- Explore **smarter augmentation strategies** aligned with test data conditions.
- Extend AEC to **multimodal settings** (e.g., audio-text correction with visual cues).
- **Integrate large language models (LLMs)** to enhance contextual correction.
- Publicly **release fine-tuned Whisper and AEC models**, OpenNMT configurations, and the **parallel Burmese AEC dataset** to support further research in low-resource ASR.

8. Conclusion

- This work investigated **ASR Error Correction (AEC)** for **low-resource Burmese** using various linguistic and alignment features.
- All AEC-based methods consistently **outperformed baseline ASR outputs**, confirming the effectiveness of error correction.
- **Alignment features** delivered the strongest and most consistent improvements across models.
- **IPA (phonetic) features** provided mixed results helpful in smaller models, but less effective for large backbones.
- **Data augmentation** improved robustness in smaller models but sometimes caused a **distribution mismatch** in stronger models.
- Results highlight the **robustness of AEC** and the importance of **careful feature selection** for low-resource speech processing.

Thanks For Your Time
Q&A

References

- 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.
- 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.
- 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.
- 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.
- 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.
- 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.

References

- 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>
- 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.
- 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.
- 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.
- 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.
- 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.

References

- 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.
- 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.
- 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/>
- 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>
- 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>

References

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

References

- 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].
- 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.
- 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/>
- 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.
- 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/>

References

- 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/>
- 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>
- 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.