Front-End Fusion and Large-Scale Weakly Supervised Decoding Module based Myanmar Speech Recognition

Jian Cui
School of Information Science and Engineering
Yunnan University
Kunming,China
1120610855@qq.com

Jian Yang*
School of Information Science and Engineering
Yunnan University
Kunming,China
jianyang@ynu.edu.cn

***Abstract*—End-to-end (E2E) speech recognition based on deep neural networks has become the mainstream approach for building high-performance speech recognition systems. Training E2E models relies on large-scale datasets of "audio- text" pairs, which presents challenges for researching speech recognition in non-general-purpose languages under extremely low resource conditions. Unsupervised pre-trained models have achieved good performance in many low-resource automatic speech recognition and have been widely applied. In this paper, under extremely low-resource conditions, we build a baseline system for Myanmar language speech recognition based on speech self-supervised learning front-end models. Moreover, we explore methods to further improve the speech recognition performance for Myanmar language based on this baseline system, and propose and implement two optimization methods: (1) linear fusion of acoustic spectral features with self-supervised speech representation features to extract richer and more accurate speech features, and (2) introducing a multilingual multitask weakly supervised pre-trained Whisper model as the decoder module for Myanmar language speech recognition system. Experimental results show that the Myanmar language speech recognition baseline system constructed in this study has a character error rate of 21.5%. After introducing the proposed improvement methods in this study, the character error rate decreases to 16.0%.**

***Index Terms*—Low-resource, Myanmar, Transfer learning, Auto-regressive language model**

I. Introduction

Low-resource speech recognition is an extensively researched topic in the field of speech recognition. In the automatic speech recognition (ASR) task, a low-resource language typically refers to the lack of sufficient "audio-text" paired datasets that possess both an adequate quantity and quality for training the model effectively. Currently, the primary focus of research in very-low-resource speech recognition includes cross-lingual transfer learning [1], speech enhancement [2], adaptive learning [3], knowledge distillation [4], and unsupervised learning [5]. Self-Supervised Learning (SSL) [6] is a specific form of unsupervised learning that involves training models on large-scale unlabeled speech data using an auxiliary task. To enhance the accuracy of the speech recognition system for low-resource languages, the self-supervised pre-trained model, trained on extensive unlabeled speech data sequences, is combined with the migration learning technique. This process enables the transfer of knowledge from the pretraining phase to the low-resource speech data.

In this paper, we propose a front-end fused approach to combine traditional FBank features with speech representations from SSL model to extract high-dimensional audio features. Moreover, we introduce a pre-trained Whisper [7] decoder module that is capable of handling multiple languages and tasks. This module aids in building a robust acoustic model and improving the performance of the end-to- end speech recognition system for the Myanmar language. Experiments shows that combining high-dimensional speech features with a pre-trained Whisper decoder model significantly improves the performance of the Myanmar speech recognition model, which has limited resources. These enhancements improve the model's recognition accuracy while reducing the number of necessary training hyperparameters, shortening the training time, and mitigating the risk of overfitting during the training process.

The remaining parts of this paper will be organized as follows: Section II introduces the relevant models and methods, Section III discusses the experimental results, and Section IV presents the conclusions and future work.

II. Related Theories And Models

A. Myanmar Language

The Myanmar language is the official and primary language of the Republic of the Union of Myanmar, with a population of nearly 54 million speakers. Linguistically, the Myanmar language belongs to the Tibeto-Burman branch of the Sino-Tibetan language family. The Republic of the Union of Myanmar is a multi-ethnic country, where the majority of

*Corresponding author

the population, accounting for over 70%, are the Myanmar people. The Myanmar language serves as both the native and common language of the Republic of the Union of Myanmar.

In comparison to English and several other European languages, similarly to Chinese, Myanmar's writing system employs a left-to-right writing order and is composed of continuous substrings within a large character set. The characters are curved and lack distinct word boundaries, being presented in a curved form. Myanmar is an analytic language, and its basic word order is subject-object-verb. Linguistic experts in Myanmar classify the tonal system of Myanmar into four categories: high-falling tone, level tone, slight-falling tone, and mid tone.

In the Myanmar language, syllables are formed by combining consonant and vowel letters. Words in Myanmar can be categorized as monosyllabic, disyllabic, or polysyllabic, with monosyllabic words being the most common type. A syllable in Myanmar may have one or multiple characters. The Myanmar language comprises 33 basic consonant letters, 7 independent vowel characters, 7 non-independent vowel characters, 4 basic medials, 10 numeral symbols, and 2 punctuation marks [8].

B. Front-end Fusion

The front-end module in ASR tasks is encompassed by various aspects. Firstly, feature extraction plays a crucial role in converting the original audio signal into a feature representation that is compatible with the model. Commonly used features include MFCC, FBank, and Mel-Spectrogram. This process, being a fundamental problem in speech signal processing, can influence the performance of the model based on different front-end choices. Additionally, the ASR front-end can perform preprocessing operations to facilitate the training needs of the model. These operations include noise removal, speech quality enhancement, and audio length adjustment. Moreover, some tasks require more contextual information. To address this, the ASR front-end can expand the feature dimension by stacking multiple frames, incorporating time sliding windows, and utilizing other methods. By doing so, the model can capture longer contextual information effectively.

We applies a front-end fusion approach to enhance the performance of resource-limited ASR systems for the Myanmar language. Our approach combines FBank features with multiple self-supervised speech representation models. 1 illustrates the implementation details of the front-end fusion framework utilizing FBank features and self-supervised speech representation models. Our approach encompasses three distinct front-end models: one utilizing the FBank feature, another employing self-supervised speech representation model 1, and a third utilizing self-supervised speech representation model 2. Each front-end model autonomously processes the speech data and produces feature vectors. To ensure consistency, we projected these feature vectors to the same dimension using linear projection. Finally, the output features from each front-end model were merged by calculating the dot product. This projection step reduces redundancy within the feature vectors

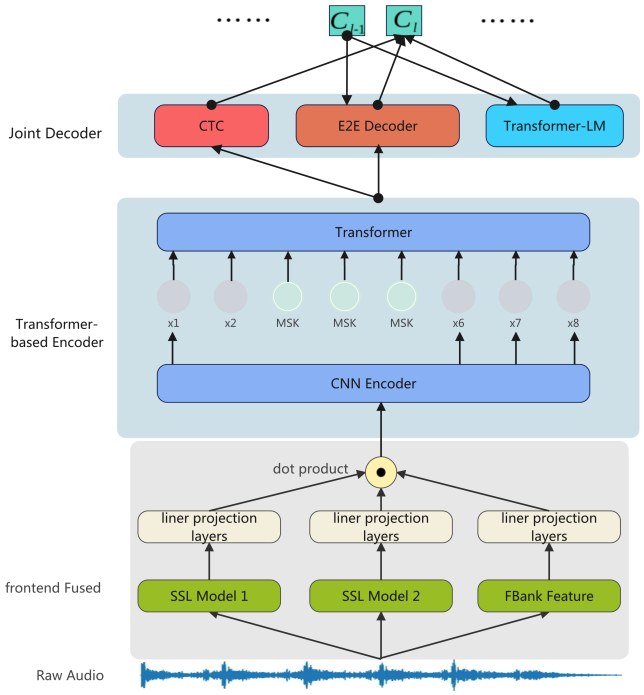

Fig. 1. Front-end fusion framework based on FBank and self-supervised speech representation.

of each front- end model, thus enhancing the robustness of the ASR model.

C. Hybrid CTC/Attention Joint Decoding Framework

The current speech recognition models utilize three frameworks, namely the Connectionist Temporal Classification (CTC) algorithm [9], the Recurrent Neural Network Transducer (RNN-T) algorithm [10], and the Encoder-Decoder algorithm based on the attention mechanism [11]. Among them, CTC decoding ensures a strict alignment between audio frames and labels. However, this algorithm assumes that the network outputs of different frames are conditionally independent, overlooking the correlated context labels in ASR tasks. ASR decoding based on the attention mechanism is prone to deletion or insertion errors, as it has flexible alignment characteristics. This can prematurely generate the end-of-sequence label (EOS) or excessively focus on repeated parts from previous inputs, resulting in overly long hypotheses.

We conducted our experiments using the open-source E2E speech processing toolkit ESPnet [12]. We adopts the Hybrid CTC/Attention joint decoding [13] architecture based on the shared encoder in 2, train the acoustic model using a multi-task learning approach, and uses Shallow Fusion [14] to combine the language model for re-scoring.

Hybrid CTC/Attention decoding combines the advantages of both the CTC and Attention mechanisms, enabling the processing of extended input and output sequences. This approach accurately captures the alignment between sequence data by using CTC and dynamically adjusts attention weights

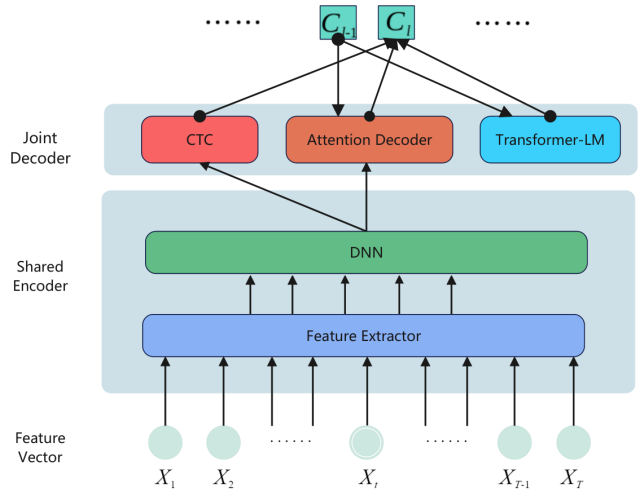

Fig. 2. Shared encoder hybrid CTC/Attention decoding architecture.

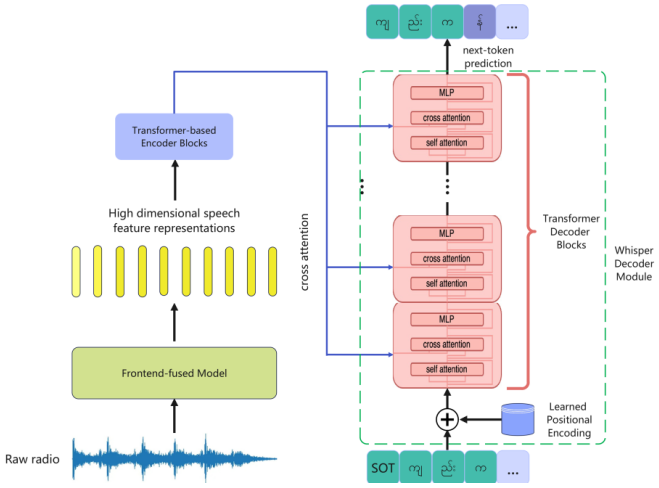

Fig. 3. Framework of the proposed approach.

to focus on relevant input sequence sections, thereby improving decoding performance.

The language model is a crucial component of the E2E ASR system, fundamental for achieving high recognition accuracy. It estimates the probability of hypothetical word sequences, learning the interrelationships between words referred to as language model scores. This training utilizes target language texts and involves optimizing cross-entropy using the entire text from the training set to train the neural network-based language model. Its objective is to capture the inherent relationship between natural language sequences and the words that follow. To assess the performance of the language model, perplexity (PPL) is used as a measure.

D. Whisper Model as Decoder

OpenAI has proposed and open-sourced the Whisper model, which presents a universal speech recognition model. The Whisper model employs the Transformer architecture to assess the resilience of speech processing systems trained with extensive weak supervision. It exhibits proficiency in handling long sequences and can be trained to perform various speech processing tasks, such as multilingual speech recognition, speech translation, speech recognition in informal settings, and speech activity detection.

The models are trained using 680,000 hours of audio data and corresponding transcriptions acquired from the Internet. They represent various inputs of task data as token sequences, which are predicted by the decoder model. The Whisper model consists of five different configurations, each varying in parameter size. Four of these pre-trained models are specifically trained for English in the ASR task, while the remaining five models are trained for multilingual data, serving both speech recognition and speech translation tasks. This allows for a trade-off between speed and accuracy. Fig.3 present the model structure of the front-end fusion module in combination with the multilingual weakly supervised Whisper decoder module.

III. EXPERIMENTAL RESULTS

Considering the limited availability of public Myanmar language corpus datasets, we decided to use a publicly available Myanmar speech corpus obtained from the OpenSLR acoustic and language resources library for our experiments. The OpenSLR identifier for the corpus is SLR80 [15], the International Standard Language Resource Number (ISLNR) (Mapelli et al., 2016) is 999-939-436-742-06. The dataset comprises 2528 audio segments as well as their corresponding original Myanmar transcription texts, with a total duration of 4.11 hours. In our methodology, we utilized the Moses [16] natural language processing tool to preprocess the Myanmar language text. We conducted tokenization using the Byte Pair Encoding (BPE) [17] segmentation algorithm to obtain subwords for recognition purposes.

The corpus consists of recordings from 20 female speakers in a calm and noise-free environment. The speakers were all volunteer participants aged between 25 and 35. The durations of the audio utterances are between 2.5 and 12 seconds, with the majority of the utterances having durations between 4 to 7.5 seconds. The sentence lengths are between 55 and 360 Unicode characters, These clips are manually segmented and aligned with their corresponding transcription texts, and saved in a consistent format as single- channel WAV files. The audio files have a sampling rate of 16 kHz and a depth of 16 bits. we randomly allocated 1848 sentences from the corpus for the training set, 200 sentences for the validation set, and 480 sentences for the test set. Table I provides specific statistical data for the Myanmar corpus.

A. Experimental Setup

For our experiments, we employed a joint decoding approach based on the Hybrid CTC/Attention decoding architecture with a shared encoder. The language model is trained using Transformer-based model on the training set text. The target language for our experiments was Myanmar, with a

TABLE I
MYANMAR CORPUS DATA STATISTICS

Dataset	Audio Duration(h)	Sentences	words	characters
Training set	3	1848	16391	129866
Validation set	0.33	200	1566	12511
Test set	0.78	480	4145	33189
Total	4.11	2528	22102	175566

subword size of 150 for BPE segmentation. The CTC fine-tuning target vocabulary consisted of 196 Myanmar character subwords, a space marker, a special CTC blank symbol, an unknown character, and start/end symbols.

Model training configuration, we using the HuBERT [18] Large and wav2vec2.0 [19] Large pretrained model as front-end, combined with attention-based encoder-decoder models as the baseline systems. During the training process, the network layers before the front-end model are frozen, while the final projection layer is randomly initialized. Model training involves utilizing labeled datasets. The encoder comprises 12 layers with 4 attention heads, and the model has an output dimension of 512. The feedforward layer's dimension is set to 2048. Dropout rates of 0.1 are applied to the neural network layers and positional layers. The model is optimized during training using the Adam optimizer with a learning rate of 0.002. Spectral augmentation is applied to the acoustic features. The weight for CTC-Loss is set to 0.3.

The improved system employs a front-end fusion framework, as illustrated in Fig.3, incorporating the FBank feature extraction model, the HuBERT Large self-supervised speech representation model, and the Wav2vec2.0 Large model. Each front-end model processes the speech data input independently, generating feature vectors that are subsequently transformed to a unified feature dimension using a linear projection layer. Finally, the output features from each front-end are combined using a dot product with a projection dimension of 100. All other experimental configurations remain the same as the baseline system.

The decoding parameters are standardized. Specifically, the weight of the CTC decoding is set at 0.5, the weight of the language model is set at 0.3, and the beam search is conducted using a beam size of 15.

B. Experimental Results

Table II presents the experimental results comparing the word error rate (WER) and character error rate (CER) on the validation/test set through different front-end fusion methods. We utilize a shallow fusion re-scoring approach with the Transformer language model. The front-end fusion method employing FBank and two self-supervised speech representation model achieved 62.8% WER and 18.2% CER on the test set. Compared to the baseline system, it reduced the WER by 4.2% and the CER by 3.2%.

In this paper, we propose integrating the Whisper model decoder module to replace the existing system, which relies on the attention decoder model. This enhancement is achieved by utilizing a front-end fusion model optimization technique. For the encoder, we configure 12 layers with the attention header set to 4. The model's output dimension is set to 512, and the feed-forward layer has an internal dimension of 2048. As for the decoder, we select the Whisper Base model, which consists of 6 residual attention layers. The neural network layer's deactivation rate and positional deactivation rate are both set to 0.1.

TABLE II
COMPARISON OF WER (%) AND CER (%) WITH DIFFERENT FRONT-END FUSION METHODS ON TEST SET

Front-end	Test	
	WER	**CER**
FBANK	71.6	26.8
FBANK+HuBERT Large	67.0	21.4
FBANK+Wav2vec2.0 Large	66.2	20.6
FBANK+HuBERT Large+Wav2vec2.0 Large	**62.8**	**18.2**

TABLE III
COMPARISON OF WER (%) AND CER (%) WITH DIFFERENT WHISPER MODEL DECODER MODULES ON TEST SET

Front-end	Models		Test	
	Encoder	**Decoder**	**WER**	**CER**
FBANK + HuBERT Large + Wav2vec2.0 Large	Transformer	Transformer	62.8	18.2
		Whisper tiny	61.6	16.9
		Whisper base	**61.0**	**16.6**
		Whisper small	63.6	19.4
		Whisper medium	66.8	21.4

Table III presents the comparative experimental results of WER and CER on the test set using different Whisper decoder modules. The study utilizes a Transformer-based language model with shallow fusion re-ranking. The most successful performance is obtained by combining the front-end fusion method based on FBank and a self-supervised speech representation model with the Whisper base decoder module. The results on the test set show 61.0% WER and 16.6% CER. Implementing the front-end fusion approach leads to 1.8% decrease in WER and 1.6% decrease in CER.

IV. CONCLUSIONS AND FUTURE WORKS

The paper aims to enhancing the performance of a low-resource Myanmar end-to-end speech recognition system with using minimal supervised training data. We propose a front-end fusion training approach that combines the FBank front-end with self-supervised speech representation models. Additionally, we optimize and enhance the baseline system by incorporating the Whisper decoder module, which is designed for multiple languages. The experimental results yield the following findings: (1) the front-end fusion method enables the extraction of superior speech features from a limited-scale Myanmar "audio-text" datasets. (2) for non- general Myanmar languages, the most effective performance is achieved through the fusion of the traditional FBank feature model with self-supervised speech representation models such as HuBERT Large and Wav2vec2.0 Large, combined with the Whisper

decoder module. (3) the experimental results of different sizes of Whisper models reveal that Whisper Tiny and Base models outperform other models, suggesting that larger models with more parameters require a larger training datasets. The experimental results verify the effectiveness of the above models and methods, which can further improve the performance of E2E speech recognition system for Myanmar with low resource conditions.

In this study, we did not consider any data enhancement methods. As a result, in our upcoming research, we aim to apply data enhancement techniques to ASR tasks in the following study. These techniques include varying speech rate, pitch, and simulating vocal tract. By enhancing the diversity of training data to enhance the overall robustness and performance of the E2E speech recognition system.

ACKNOWLEDGMENT

This work is supported by Science and Technology Innovation 2030-Major Project (No.2020AAA0107901) and Natural Science Foundation of China (No.61961043).

REFERENCES

[1] Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. *arXiv preprint arXiv:2005.00052*, 2020.

[2] S Pascual, A Bonafonte, and J Serra. Speech enhancement generative adversarial network.

[3] Kaisheng Yao, Dong Yu, Frank Seide, Hang Su, Li Deng, and Yifan Gong. Adaptation of context-dependent deep neural networks for automatic speech recognition. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 366–369. IEEE, 2012.

[4] Yevgen Chebotar and Austin Waters. Distilling knowledge from ensembles of neural networks for speech recognition. In *Interspeech*, pages 3439–3443, 2016.

[5] Alexei Baevski, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. Unsupervised speech recognition. *Advances in Neural Information Processing Systems*, 34:27826–27839, 2021.

[6] Xuankai Chang, Takashi Maekaku, Pengcheng Guo, Jing Shi, Yen-Ju Lu, Aswin Shanmugam Subramanian, Tianzi Wang, Shu-wen Yang, Yu Tsao, Hung-yi Lee, et al. An exploration of self-supervised pretrained representations for end-to-end speech recognition. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 228–235. IEEE, 2021.

[7] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

[8] Jian Yang et al. Burmese word segmentation method and implementation based on crf. In *2018 International Conference on Asian Language Processing (IALP)*, pages 340–343. IEEE, 2018.

[9] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006.

[10] Alex Graves. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*, 2012.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[12] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, et al. Espnet: End-to-end speech processing toolkit. *arXiv preprint arXiv:1804.00015*, 2018.

[13] Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi. Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253, 2017.

[14] Ding Zhao, Tara N Sainath, David Rybach, Pat Rondon, Deepti Bhatia, Bo Li, and Ruoming Pang. Shallow-fusion end-to-end contextual biasing. In *Interspeech*, pages 1418–1422, 2019.

[15] 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).

[16] Hieu Hoang and Philipp Koehn. Design of the moses decoder for statistical machine translation. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 58–65, 2008.

[17] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

[18] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

[19] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.