myMediCon: End-to-End Burmese Automatic Speech Recognition for Medical Conversations

Hay Man Htun*, Ye Kyaw Thu†, Hutchatai Chanlekha‡, Kotaro Funakoshi§, Thepchai Supnithi†

*Department of Electrical Engineering, Kasetsart University,
50 Ngamwongwan Road, Lat Yao, Chatuchak, Bangkok 10900, Thailand

†Language and Semantic Technology Research Team (LST),
Artificial Intelligence Research Group (AINRG),
National Electronics and Computer Technology Center (NECTEC),
112 Phahonyothin Road, Klong Nueng, Klong Luang, Pathumthani 12120, Thailand

‡Department of Computer Engineering, Kasetsart University,
50 Ngamwongwan Road, Lat Yao, Chatuchak, Bangkok 10900, Thailand

§Department of Information and Communication Technology, Tokyo Institute of Technology,
4259 Nagatsuta-cho, Midori-ku, Yokohama, Kanagawa 226-8503, Japan

hayman.h@ku.th, fenghtc@ku.ac.th, funakoshi@lr.pi.titech.ac.jp
{yekyaw.thu, thepchai.supnithi}@nectec.or.th

Abstract

End-to-End Automatic Speech Recognition (ASR) models have significantly advanced the field of speech processing by streamlining traditionally complex ASR system pipelines, promising enhanced accuracy and efficiency. Despite these advancements, there is a notable absence of freely available medical conversation speech corpora for Burmese, which is one of the low-resource languages. Addressing this gap, we present a manually curated Burmese Medical Speech Conversations (myMediCon) corpus, encapsulating conversations among medical doctors, nurses, and patients. Utilizing the ESPnet speech processing toolkit, we explore End-to-End ASR models for the Burmese language, focus on Transformer and Recurrent Neural Network (RNN) architectures. Our corpus comprises 12 speakers, including three males and nine females, with a total speech duration of nearly 11 hours within the medical domain. To assess the ASR performance, we applied word and syllable segmentation to the text corpus. ASR models were evaluated using Character Error Rate (CER), Word Error Rate (WER), and Translation Error Rate (TER). The experimental results indicate that the RNN-based Burmese speech recognition with syllable-level segmentation achieved the best performance, yielding a CER of 9.7%. Moreover, the RNN approach significantly outperformed the Transformer model.

Keywords: Speech corpus building, Automatic Speech Recognition, RNN, Transformer, low-resource, Burmese

1. Introduction

Automatic Speech Recognition (ASR) technology has emerged as a pivotal tool in the medical domain, offering a myriad of benefits to healthcare professionals and patients alike (Hoyt and Yoshihashi, 2010). Its ability to convert spoken language into written text has proved invaluable in scenarios such as real-time transcription of physician-patient interactions, which not only facilitates accurate documentation but also allows healthcare providers to focus more on the patient rather than on note-taking. Additionally, ASR has the potential to streamline workflow in medical settings by enabling voice-activated systems that assist in scheduling, data retrieval, and even preliminary diagnoses based on vocalized symptoms. However, based on studies conducted by Goss et al. (2019) and Sima (2016), the accuracy of speech recognition in the medical field varies, between 78% and 92%.

In recent years, the modeling and decoding processes in ASR systems have been noticeably simplified by end-to-end (E2E) architectures. A single neural network that implicitly incorporates the acoustic model (AM), pronunciation lexicon, and language model (LM) that make up a traditional hybrid system converts speech into text by propagating the model in the forward direction. E2E models were believed to achieve comparable results with the hybrid systems only when trained on large datasets, but since the majority of E2E model trainings are purely data-driven, the performance is highly dependent on the scale of training data (Li et al., 2021). Generally, ASR technologies require a large amount of speech data for a system to work

LREC-COLING 2024, pages 12032–12039

well (Scharenborg et al., 2017).

There are very few R&D works on Burmese ASR (Naing et al., 2015; Mon et al., 2019a,b) and it still requires extensive experimentation compared to English. To the best of the authors' knowledge, there is scarce publicly available speech corpora for Burmese (Oo et al., 2020). The lack of resource in speech corpora for Burmese is one of the main barriers why ASR for Burmese is not as developed as English or other developed languages. Moreover, until now, there has been no Burmese speech corpus for the medical domain. Therefore, the main motivation for this study is to develop a publicly available medical domain speech corpus for Burmese and apply E2E ASR technologies for the evaluation of our proposed speech corpus.

This paper is organized as follows. A brief review of the related works is provided in Section 2. The nature of the Burmese language is described in Section 3. Details of the developed myMediCon corpus are given in Section 4. The methods applied and the experimental settings are explained in Section 5. This section introduces the characteristics of RNN (Subsection 5.1) and Transformer (Subsection 5.2) approaches used in our Burmese ASR models, presents data statistics for the system experiments (Subsection 5.3), and describes the evaluation metrics (Subsection 5.4). Section 6 reports the experimental results and discusses these results. Section 7 conducts an error analysis to examine the inaccuracies of the ASR models and suggests potential areas of improvement. Finally, Section 8 concludes the paper with future works.

2. Related Work

There are a few studies on Burmese ASR from different domains. A Burmese large vocabulary continuous speech recognition system for travel-tour domain was contributed by Naing et al. (2015). In this system, 3 different acoustic models were investigated using a created phonemically balanced corpus, which included 4K sentences and 40 hours of speech. These models included a Gaussian Mixture Model (GMM) and two Deep Neural Networks (DNNs). 100 utterances from 25 speakers in an open evaluation set were used in the experiment. The word error rate (WER) or syllable error rate (SER) results for the sequence discriminative training DNNs reached up to 15.63% and 10.87%, respectively.

Mon et al. (2019a) presented a Burmese speech corpus for ASR. In this study, a speech corpus called UCSY-SC1 (University of Computer Studies Yangon - Speech Corpus1) is developed for Burmese ASR research. The corpus consists of two different domain types: daily conversations and news. The experiments utilized various data sizes, with evaluation conducted on two distinct test sets: TestSet1 for web news and TestSet2 for recorded conversational data. Word error rates of 15.61% on TestSet1 and 24.43% on TestSet2 are the results of the Burmese ASR using this corpus.

Based on the existing literature and previous work concerning Burmese ASR systems, there appears to be no established end-to-end Burmese ASR system specifically tailored for the medical domain. Additionally, there are no publicly available resources for Burmese ASR within this domain. Given this gap, we aimed to develop our own myMediCon corpus encompassing medical domain. This initiative was undertaken to pioneer the investigation of the first end-to-end Burmese ASR system using our newly developed speech corpus from the medical domain. Subsequently, we evaluated the performance of this system employing deep learning models, namely, Recurrent Neural Networks (RNN) and Transformer architectures, within the ESPnet framework.

3. Burmese

There are approximately a hundred languages spoken in Myanmar. Burmese or Myanmar language is the official language of Myanmar (Htun et al., 2021). It is classified as a member of the Tibeto-Burman language family. It is also the most widely spoken language in Myanmar. About 32 million people speak Burmese as their first language, while another 10 million speak it as a second language. In Burmese text, words are represented as continuous strings of characters without any explicit word boundary markings. Notably, there are no spaces between words in Burmese. Burmese comprises 33 basic consonants, 12 vowels, 4 medials, and extension vowels, vowel symbols, devowelizing consonants, diacritic marks, specified symbols, and punctuation marks.

Burmese is a tonal language. This means that all syllables have prosodic features that are an integral part of their pronunciation and that affect word meaning. Prosodic contrasts involve not only pitch, but also phonation, intensity (loudness), duration, and vowel quality (Mon et al., 2019b). The phonology of Burmese is intricately structured, arising from the combination of vowels and consonants. The phonological structure of Burmese is defined by the utilization of singular vowels, or combinations of one vowel and consonant, represented by consonant combination symbols or sign Virama (“◌္”, U1039) (Consortium, 2023). In Burmese, each vowel has its distinct and specific sound. Due to the relative scarcity of the available language resources, Burmese is often considered to be an under-resourced language (Oo et al., 2020).

4. Building myMediCon Corpus

Developing speech corpora is a fundamental step in the creation of speech processing systems such as ASR and Text-to-Speech (TTS), especially for low-resourced languages, it holds critical significance. Moreover, the performance of a speech recognizer is heavily contingent upon the quality and relevance of the speech corpora. The primary contribution of this work is the manual construction of a corpus called "myMediCon," consisting of Burmese Medical Speech Conversations.

The construction of a speech corpus can primarily be approached through two methods. The first method entails the collection of pre-existing speech data, which is then manually transcribed into text. The second method involves the creation of a text corpus initially, followed by the recording of speech as the collected text is read aloud. In this work, we employed the latter method.

Burmese Sentence:	ကင်ဆာ ဖြစ် နိုင် လား ၊ ဒေါက်တာ ။
English Translation:	Could it be cancer, Doctor?
Burmese Sentence:	ခင်ဗျား မှာ ဆီးချို ရောဂါ ၊ သွေးတိုး ရောဂါ ကဲ့သို့သော ရောဂါ အခြေအနေ မျိုး တွေ ရှိ ပါ သလား ။
English Translation:	Do you have any medical problems like diabetes mellitus, high blood pressure?
Burmese Sentence:	ခင်ဗျား မှာ ရင်ဘတ် အောင့် တာ ကဲ့သို့သော နှလုံး ရောဂါ ပြဿနာ မျိုး ရှိ ပါ သလား ။
English Translation:	Do you have any heart problems like angina?
Burmese Sentence:	သူ့ ကို အနာ သက်သာ ဖို့ အိုင်ဘူပရိုဖန် အစား ပါရာစီတမော သုံး ဖို့ အကြံပေး ပါ ။
English Translation:	Advise him to use paracetamol instead of ibuprofen for pain relief.

Table 1: Example word-level segmented Burmese sentences of the myMediCon corpus.

The medical sentences for the text corpus were sourced from the "Samson Handbook of Plab 2 and Clinical Assessment" (Samson, 2015). These sentences encompass conversations between patients and doctors, names of diseases and medicines, and treatment methodologies (San et al., 2022). Examples of word-level segmented Burmese sentences are illustrated in Table 1. Three university students manually translated the medical sentences from English to Burmese. The text corpus in the medical domain comprises a total of 14,592 medical sentences, encompassing 232,999 words and 15,431 unique words.

SpeakerID	Utterance	Duration
spk00	1001	2 hr, 1 min, 41 sec
spk01	14592	26 hr, 16 min, 35 sec
spk02	14592	37 hr, 33 min, 5 sec
spk03	1000	2 hr, 56 min 28 sec
spk04	1000	1 hr, 57 min, 40 sec
spk05	2000	3 hr, 52 min, 24 sec
spk06	1500	3 hr, 5 min, 21 sec
spk07	1000	1 hr, 48 min, 23 sec
spk08	500	59 min, 3 sec
spk09	500	46 min, 55 sec
spk10	4092	7 hr, 53 min, 9 sec
spk11	3000	4 hr, 48 min, 59 sec

Table 2: Total duration and utterances of each speaker in myMediCon corpus.

In constructing the speech corpus, a total of 14,592 Burmese medical sentences were collected. The recordings were conducted using a TASCAM (DR44-WL) recorder, an audio-recording software on a laptop (Dell i7 8th Gen), Voice Memos app from MacBook, built-in audio-recording apps on various Android devices (Oppo recorder on Oppo F 17, Redmi recorder on Redmi Note 6 Pro and Redmi Note 10 Pro, and Xiao Mi recorder on Mi Note 12 Pro) and Voice Memos apps on iOS devices (iPhone 7 +, iPhone 11 Pro Max, and iPhone 14 +). The corpus comprises utterances from 12 speakers, including 3 males and 9 females. Among these, 7 speakers are native Burmese speakers, while the remaining 5 are individuals from different ethnic nationalities of Myanmar namely, Pa'O, Kachin, Dawei, and Mon who speak Burmese as their second language. The age range of the speakers is between 20 and 30 years. The cumulative duration of the audio utterances is 93 hours, 59 minutes and 43 seconds, with individual audio utterance durations ranging from 1.8 to 35 seconds. The total duration alongside the number of utterances recorded by each speaker in the medical domain speech corpus is presented in Table 2.

5. Methods and Experimental Setup

We developed end-to-end ASR models utilizing two neural network architectures, RNN (Recurrent Neural Network) and Transformer, to evaluate the performance of these models on our self-compiled speech corpora from medical domain. For this endeavor, we employed ESPnet (Watan-

abe et al., 2018), a comprehensive end-to-end speech processing toolkit that encompasses a range of speech processing tasks such as end-to-end speech recognition, text-to-speech conversion, speech translation, speech enhancement, speaker diarization, and spoken language understanding, among others. ESPnet leverages PyTorch as its deep learning engine and adheres to Kaldi-style data processing, feature extraction, formatting, and recipe protocols, thereby furnishing a complete setup for an array of speech processing experiments[1] (Inaguma et al., 2020).

5.1. Recurrent Neural Network (RNN)

Recurrent Neural Networks (RNNs) are a class of deep learning (DL) methods employed in ASR, known for their flexibility in formatting both bidirectional and unidirectional models (Suyanto et al., 2020). They are used to detect patterns in sequential data due to their inherent capability to capture temporal dependencies (Schmidt, 2019).

In our research, we devised an efficacious ASR model utilizing the ESPnet framework and an architected RNN design. A significant emphasis was placed on hyperparameter tuning to optimize the model configuration for peak performance. A pivotal component of our ASR system is the encoder, based on the Visual Geometry Group-RNN (VGG-RNN) architecture. Comprising four bidirectional LSTM (Long Short-Term Memory) layers with an output size of 320, it is tailored to capture complex temporal patterns in the input speech data. Particularly, we employed a projection mechanism and adjusted dropout rates to enhance the model's generalizability while mitigating overfitting.

The accuracy of transcription is bolstered by the decoder, a unidirectional LSTM layer with 320 hidden units, further refining the learned representations. Our training strategy amalgamated a robust scheduling mechanism with the Adadelta optimizer, set at a learning rate of 1.0. Specifically, we utilized the ReduceLROnPlateau scheduler with patience set to 1 and mode set to minimize, ensuring dynamic adjustment of learning rates contingent on validation performance.

5.2. Transformer

Transformer is a sequence-to-sequence (seq2seq) architecture originally proposed for neural machine translation (NMT) that has rapidly replaced recurrent neural networks (RNNs) in natural language processing tasks. Transformer learns sequential information via a self-attention mechanism instead of the recurrent connection employed in RNNs (Karita et al., 2019).

[1]Available at https://github.com/espnet/espnet

In this study, we used a transformer architecture to design and implement an ASR model. The configuration of our model was adjusted to maximize performance, balancing important factors like attention heads, linear units, and dropout rates. Based on the transformer framework, the encoder has 12 blocks with four attention heads each, enabling complex feature extraction and representation. We utilized attention dropout and positional dropout techniques to improve the model's learning capabilities. Additionally, our model gains from convolutional input layers, which improves its capacity to recognize complex patterns in the speech data.

The decoder, which consists of three transformer blocks, uses a similar attention configuration to ensure efficient context integration while decoding. Additionally, our model includes label smoothing with a weight of 0.1 and joint Connectionist Temporal Classification (CTC) - attention training with a weight of 0.3 for the CTC that has been carefully calibrated. We used a warmup learning rate scheduler that linearly increases and exponentially decreases the learning rate during training to ensure stable convergence. The training process was optimized using the Adam optimizer with a learning rate of 0.001. Our ASR system is built to be reliable, utilizing 256-dimensional embeddings and 4 attention heads, which helps it capture intricate linguistic patterns in speech data.

5.3. Data Statistics

In the ASR system experiments, we utilized a Burmese speech corpus consisting of 5,400 Burmese sentences, with each speaker contributing 450 sentences. These sentences were randomly selected from our comprehensive myMediCon and text corpus developed for the medical domain. It includes a total of 82,218 words and 7,212 unique words. There is no overlap between our extracted sentences. In the extracted speech corpus, the total duration of audio utterances is 10 hours, 46 minutes, and 1.35 seconds. Each audio utterance ranges between 1.8 and 35 seconds in duration. The corpus comprises recordings from a total of 12 speakers, including 3 males and 9 females. The total duration and the amount of utterances recorded by each speaker in the extracted medical domain speech corpus for the experiments of ASR models are shown in Table 3.

We performed three experiments by random sub-sampling validation to evaluate the performance of our ASR models. The detailed statistics on the train, development or validation, and test sets for the three experiments are displayed in Table 4, Table 5, and Table 6. In each experiment, the sentences from all sets are not overlap. In the first experiment, the training set comprised

11 speakers, with 2 males and 9 females, while the development set included 2 speakers, with 1 male and 1 female. The test set consisted of 12 speakers, with 3 males and 9 females. Similarly, in the second experiment, the training set comprised 11 speakers, with 3 males and 8 females, while the development set involved 2 speakers, both females. The test set included 12 speakers, with 3 males and 9 females. In the third experiment, there were 12 speakers in each set, with 3 males and 9 females distributed equally.

SpeakerID	Utterance	Duration
spk00	450	55 min, 50 sec
spk01	450	39 min, 57 sec
spk02	450	1 hr, 16 min, 18 sec
spk03	450	1 hr, 25 min 51 sec
spk04	450	55 min, 44 sec
spk05	450	50 min, 30 sec
spk06	450	53 min, 46 sec
spk07	450	38 min, 23 sec
spk08	450	52 min, 57 sec
spk09	450	42 min, 45 sec
spk10	450	46 min, 53 sec
spk11	450	47 min, 02 sec

Table 3: Total duration and utterances of each speaker in the extracted Burmese Speech corpus for the system.

Datasets	Utterance	Duration
Train	4200	8 hr, 23 min, 1 sec
Dev	600	1 hr, 9 min, 45 sec
Test	600	1 hr, 13 min, 14 sec

Table 4: Statistics on train, development and test sets used for the first experiment of the system.

Datasets	Utterance	Duration
Train	4200	7 hr, 50 min, 28 sec
Dev	600	1 hr, 43 min, 27 sec
Test	600	1 hr, 12 min, 12 sec

Table 5: Statistics on train, development and test sets used for the second expriment of the system.

Datasets	Utterance	Duration
Train	4200	8 hr, 25 min, 43 sec
Dev	600	1 hr, 9 min, 45 sec
Test	600	1 hr, 9 min, 7 sec

Table 6: Statistics on train, development and test sets used for the third experiment of the system.

5.4. Evaluation Metrics

Word Error Rate (WER), Character Error Rate (CER), and Translation Error Rate (TER) are common metrics used in evaluating the performance of ASR, Optical Character Recognition (OCR), and Machine Translation (MT) systems, respectively. WER is calculated as $\frac{S+D+I}{N}$, where S, D, and I represent the number of substitutions, deletions, and insertions, respectively, and N is the total number of words in the reference. Similarly, CER is computed as $\frac{S+D+I}{N}$, but at the character level. TER extends the idea to machine translation evaluation, and is calculated as $\frac{S+D+I+Sh}{N}$, where Sh represents the number of shifts, which are word movements in the text. These metrics provide a quantitative measure of the accuracy and quality of the respective systems, with lower values indicating better performance.

6. Results and Discussion

In this section, we present the average and the best evaluation results of the three experiments for End-to-End (E2E) Burmese ASR. Three separate experiments were used for the evaluation, each with a different random split dataset.

The RNN model's average performance throughout all experiments offers favorable results. In RNN model, for the word-level corpus type, the average WER was 40.2%, and the average CER and TER were 13.6% and 20.7%, respectively. Concurrently, for the syllable-level corpus type, the RNN model achieved an average WER of 25.8%, with average CER and TER of 14.1% and 20.9%, respectively.

The Transformer model performed at a competitive rate with slightly higher average error rates than the RNN model. The Transformer model yielded an average WER of 53.4%, CER of 18.5%, and TER of 27.3% for word-level corpus type. Similarly, the Transformer model produced an average WER of 26.4% for syllable-level corpus type, with averages for CER and TER of 12.7% and 19.5%, respectively.

The best evaluation outcomes among the three experiments for E2E Burmese ASR models employing RNN and Transformer architectures are depicted in Table 7 and Table 8, respectively. Furthermore, we conducted a comparison between word-level and syllable-level segmentations. Given that a syllable is a fundamental unit in Burmese (Thu et al., 2021), we utilized a syllable segmentation tool named "sylbreak" to convert Burmese words to syllable levels[2].

[2]Available at https://github.com/ye-kyaw-thu/sylbreak

Corpus Types	WER%	CER%	TER%
Word	36.4	11.5	17.3
Syllable	19.0	9.7	14.7

Table 7: The best evaluation results of RNN model for E2E Burmese ASR with word error rate (WER), character error rate (CER) and translation error rate (TER) among the three experiments.

Corpus Types	WER%	CER%	TER%
Word	50.7	17.3	25.5
Syllable	23.1	11.2	17.1

Table 8: The best evaluation results of Transformer model for Burmese ASR with word error rate (WER), character error rate (CER) and translation error rate (TER) among the three experiments.

Observing the results, it is discernible that syllable-level segmentation significantly enhances the accuracy of both RNN and Transformer models. For instance, transitioning from word to syllable segmentation reduces the WER from 36.4% to 19.0% in the RNN model, and from 50.7% to 23.1% in the Transformer model. This pattern of improvement is consistent across all three evaluation metrics, underscoring the efficacy of syllable-level segmentation in capturing the nuances of Burmese speech, a language where syllables play a fundamental role.

Upon comparing the RNN and Transformer models, the RNN model exhibits superior performance across all metrics and segmentation levels. For example, at syllable-level segmentation, the RNN model achieves a lower WER of 19.0%, compared to the Transformer's WER of 23.1%. The trend remains consistent in CER and TER metrics as well. These findings are pivotal as they not only demonstrate the relative robustness of the RNN model in handling Burmese speech recognition tasks but also emphasize the importance of choosing an appropriate segmentation level based on the linguistic characteristics of the target language. This insight is crucial for further advancements in Burmese ASR systems, potentially aiding in the development of more accurate and efficient models.

Furthermore, it is worth noting that the lower performance of the Transformer model could be attributed to our low-resource setting. Despite having a larger corpus of approximately 93 hours, 59 minutes, and 43 seconds, due to limited GPU and computing resources, we opted to use only about 11 hours from the developing corpus for the experiments, which was built with data from 12 speakers. Transformers typically require extensive data and meticulous hyperparameter tuning, including warm-up procedures, to achieve optimal performance, which might not have been feasible given our resource constraints and the experimental setup.

7. Error Analysis

Error analysis was conducted to examine the inaccuracies of the ASR models in generating hypotheses, utilizing SCLITE (NIST, 2021) for the analysis. Tables 9 and 10 present the top 10 error examples from the RNN models. We discovered that the majority of the issues stemmed from phonetic discrepancies, text encoding, or word segmentation challenges. A phonetic error occurs when the model estimates a word that phonetically resembles another word (for example; "နှင့်" is pronounced as "nhin." and "နဲ့" is pronounced as "ne."). Some errors are related to incorrect typing order of Burmese syllables within the corpus, leading to different syllable segmentation and thereby impacting the ASR model's performance (for example; "င့်" and "နဲ့"). Our analysis suggests that enlarging the speech corpus and normalizing the text transcriptions could mitigate such errors.

No.	Frequency	REF ⇒ HYP
1:	26	မှ ⇒ မှာ
2:	17	ရ ⇒ ရာ
3:	12	တ ⇒ တစ်
4:	9	င့် ⇒ နဲ့
5:	9	င့် ⇒ မြင်
6:	9	သော ⇒ တော့
7:	8	စ ⇒ စား
8:	8	စာ ⇒ စား
9:	8	တဲ့ ⇒ နဲ့
10:	8	တွင်း ⇒ တွင်

Table 9: Top 10 errors of RNN ASR model with syllable segmentation

No.	Frequency	REF ⇒ HYP
1:	10	မှ ⇒ မှာ
2:	9	တာလဲ ⇒ လဲ
3:	9	တဲ့ ⇒ နဲ့
4:	9	နှင့် ⇒ နဲ့
5:	7	အလေးချိန် ⇒ ကိုယ်အလေးချိန်
6:	7	ကံမကောင်းစွာ ⇒ စွာ
7:	7	ဖြစ်စေ ⇒ စေ
8:	6	မှာ ⇒ ထဲမှာ
9:	6	တာကို ⇒ ကို
10:	8	ထွက် ⇒ သွေးထွက်

Table 10: Top 10 errors of RNN ASR model with word segmentation

8. Conclusion and Future Work

In this work, we have unveiled a pioneering endeavor towards low-resourced End-to-End (E2E) Burmese ASR in the medical domain utilizing RNN and Transformer architectures within the ESPnet framework. Through the meticulous curation of a specialized Burmese Speech corpus, encompassing nearly 11 hours of medical dialogues from 12 distinct speakers, we have laid a foundational stone for exploring ASR potentials in under-resourced languages like Burmese. Our comprehensive evaluation underscores the significance of appropriate segmentation methodologies demonstrating superior performance with syllable-level segmentation over word-level segmentation, especially in the context of the RNN model which notably outperformed the Transformer model across all evaluation metrics.

The empirical results accentuate the RNN model's robustness, achieving the lowest CER of 9.7% amidst all configurations, thereby shedding light on its suitability for Burmese ASR tasks. Moreover, the noteworthy reduction in WER and TER with syllable-level segmentation articulates the importance of aligning segmentation strategies with the linguistic nuances of the target language. Our error analysis elucidates that addressing phonetic discrepancies, text encoding, and word segmentation challenges, possibly through enlarging the speech corpus and normalizing text transcriptions, holds the potential to further refine the accuracy of the ASR models.

In conclusion, this study not only contributes a valuable resource to the sparse landscape of Burmese ASR but also provides pivotal insights into the interplay of segmentation strategies and model architectures. The promising results, despite the corpus' size limitations, invoke an optimistic outlook for further enhancements in Burmese ASR accuracy and efficiency through extended speech corpora, advanced end-to-end ASR techniques, data augmentation, and transfer learning explorations. These directions will be the main focus of our future work as we move forward to advance Burmese ASR, particularly in domain-specific applications like medical dialogues, towards real-world usability and higher accuracy benchmarks. In addition, we are planning to publish our developed myMediCon corpus, which focus on the medical domain for future Burmese speech researchers.

9. Acknowledgements

This research is financially supported by National Research Council of Thailand (NRCT), Thailand Advanced Institute of Science and Technology (TAIST), National Science and Technology Development Agency (NSTDA), Tokyo Institute of Technology, and Faculty of Engineering, Kasetsart University (KU) under the TAIST Tokyo Tech Program. The Language Understanding Lab in Myanmar made significant contributions to the development of the myMediCon corpus, including corpus design, translation from original English sentences, and support for recording devices. We would also like to extend our gratitude to all speakers for their kind assistance with recording.

10. References

The Unicode Consortium. 2023. Myanmar, Range: 1000–109F, The Unicode Standard, Version 15.1. https://www.unicode.org/charts/PDF/U1000.pdf. [Online; accessed 19-October-2023].

Foster Goss, Suzanne Blackley, Carlos Ortega, Leigh Kowalski, Adam Landman, Ct Lin, Marie Meteer, Samantha Bakes, Stephen Gradwohl, David Bates, and Li Zhou. 2019. A clinician survey of using speech recognition for clinical documentation in the electronic health record. *International Journal of Medical Informatics*, 130.

Robert Hoyt and Ann Yoshihashi. 2010. Lessons learned from implementation of voice recognition for documentation in the military electronic health record system. *Perspectives in health information management / AHIMA, American Health Information Management Association*, 7:1e.

Hay Man Htun, Ye Kyaw Thu, Hlaing Myat Nwe, May Thu Win, and Naw Naw. 2021. Statistical machine translation system combinations on phrase-based, hierarchical phrase-based and operation sequence model for burmese and pa'o language pair. *Journal of Intelligent Informatics and Smart Technology*, 2(2):1–9. Submitted July 16, 2021; accepted October 13, 2021; revised October 20, 2021; published online October 31, 2021.

Hirofumi Inaguma, Shun Kiyono, Kevin Duh, Shigeki Karita, Nelson Yalta, Tomoki Hayashi, and Shinji Watanabe. 2020. ESPnet-ST: All-in-one speech translation toolkit. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 302–311, Online. Association for Computational Linguistics.

Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin,

Ryuichi Yamamoto, Xiaofei Wang, Shinji Watanabe, Takenori Yoshimura, and Wangyou Zhang. 2019. A comparative study on transformer vs rnn in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 449–456.

Mohan Li, Catalin Zorila, and Rama Doddipatla. 2021. Transformer-based online speech recognition with decoder-end adaptive computation steps. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 1–7.

Aye Nyein Mon, Win Pa Pa, and Ye Kyaw Thu. 2019a. UCSY-SC1: A Myanmar speech corpus for automatic speech recognition. *International Journal of Electrical and Computer Engineering (IJECE)*, 9:3194–3202.

Khaing Zar Mon, Ye Kyaw Thu, Hay Man Htun, Zun Hlaing Moe, Thida San, Hnin Aye Thant, and Reenu. 2019b. Study on extremely low-resource automatic speech recognition (asr) with burmese, shan, and pa'o languages. In *The 4th ONA Conference*, Phnom Penh, Cambodia. Ministry of Posts and Telecommunications.

Hay Mar Soe Naing, Aye Mya Hlaing, Win Pa Pa, Xinhui Hu, Ye Kyaw Thu, Chiori Hori, and Hisashi Kawai. 2015. A myanmar large vocabulary continuous speech recognition system. In *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 320–327.

NIST. 2021. National Institute of Standards and Technology (NIST), Speech recognition scoring toolkit (sctk), Version 2.4.12. https://github.com/usnistgov/SCTK. [Online; accessed 19-October-2023].

Yin May Oo, Theeraphol Wattanavekin, Chenfang Li, Pasindu De Silva, Supheakmungkol Sarin, Knot Pipatsrisawat, Martin Jansche, Oddur Kjartansson, and Alexander Gutkin. 2020. Burmese speech corpus, finite-state text normalization and pronunciation grammars with an application to text-to-speech. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6328–6339, Marseille, France. European Language Resources Association.

C. Samson. 2015. *Samson Handbook of PLAB and Clinical Assessment*. [Online]. Available: https://books.google.co.th/books?id=UUT0swEACAAJ.

Mya Ei San, Ye Kyaw Thu, Thepchai Supnithi, and Sasiporn Usanavasin. 2022. Improving neural machine translation for low-resource english-myanmar-thai language pairs with switchout data augmentation algorithm. In *2022 17th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, pages 1–6.

Odette Scharenborg, Francesco Ciannella, Shruti Palaskar, Alan W. Black, Florian Metze, Lucas Ondel, and Mark Hasegawa-Johnson. 2017. Building an asr system for a low-resource language through the adaptation of a high-resource language asr system: Preliminary results. Unpublished manuscript.

Robin M. Schmidt. 2019. Recurrent neural networks (rnns): A gentle introduction and overview.

Ajami Sima. 2016. Use of speech-to-text technology for documentation by healthcare providers. *The National medical journal of India*, 29, 3:148–152.

Suyanto Suyanto, Anditya Arifianto, Anis Sirwan, and Angga P. Rizaendra. 2020. End-to-end speech recognition models for a low-resourced indonesian language. In *2020 8th International Conference on Information and Communication Technology (ICoICT)*, pages 1–6.

Ye Kyaw Thu, Hlaing Myat Nwe, Hnin Aye Thant, Hay Man Htun, Htay Mon, May Myat Myat Khine, Mi Hsu Pan Oo, Mi Pale Phyu, Nang Aeindray Kyaw, Thazin Myint Oo, Thazin Oo, Thet Thet Zin, and Thida Oo. 2021. sylbreak4all: Regular expression based syllable breaking tool for nine major ethnic languages of myanmar. In *Proceedings of the 16th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP 2021)*, pages 1–6, virtual conference.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. ESPnet: End-to-end speech processing toolkit. In *Proceedings of Interspeech*, pages 2207–2211.