



A Cascaded Burmese-to-Myeik Speech-to-Speech Translation System Using ASR, mBART-50 and MMS-VITS

Ye Thu Aung⁽¹⁾, Thet Thet Zin⁽²⁾ *

⁽¹⁾ University of Information Technology Yangon, Myanmar

⁽²⁾ Faculty of Information Science, University of Information Technology, Yangon, Myanmar

yethuaung@uit.edu.mm, thetthetzin@uit.edu.mm

ABSTRACT

Speech-to-speech translation system translates speech from one language to another in speech form. However, developing this type of system for a low-resource language is challenging because only a limited amount of text and speech data is available. This paper proposes a Burmese to Myeik speech-to-speech translation system, which is based on Automatic Speech Recognition (ASR), Machine Translation (MT), and Text-to-Speech (TTS). First, CNN-BiLSTM-CTC model was used to recognize Burmese speech and transform it into Burmese text. Then, the obtained text was translated into Myeik language with the help of the fine-tuned mBART-50 model. As the recognized text may contain errors caused by the ASR model, the MT model was trained on both clean Burmese sentences and ASR-generated sentences. To obtain Myeik speech from the translated Myeik text, a pretrained Burmese MMS-VITS model was fine-tuned on the Myeik speech dataset. The results of the experiments were obtained as follows: The ASR model obtained a CER of 21.26% and SER of 27.29%. The mBART-50 model obtained a BLEU score of 13.53 on the clean Burmese-to-Myeik test set. The robust MT model achieved a BLEU score of 13.09 when translating ASR-generated Burmese sentences. Finally, the experiments confirmed that the fine-tuned TTS model could generate intelligible Myeik speech. To conclude, the results of the experiments demonstrated that it is possible to develop a speech-to-speech translation system for the low-resource language based on a limited amount of training data.

Keywords: Burmese, Myeik, Speech-to-Speech Translation, Automatic Speech Recognition, Machine Translation, Text-to-Speech

1. Introduction

Speech is one of the most common means of communication among people; however, communication can become difficult when people communicate in different or regional dialects. Speech-to-speech translation (S2ST) can solve this problem of communication barriers between people by translating one's speech into another language and producing speech as an output. The S2ST systems have been developed using both cascaded and direct approaches [1], [2].

In a cascaded S2ST system, the translation process involves three main steps: Automatic Speech Recognition (ASR), Machine Translation (MT), and Text-to-Speech (TTS). First, the ASR model recognizes the source speech and outputs its textual representation. Then, the textual information is fed into the MT model, which translates it into the target language. Finally, the resulting text is used by the TTS model to produce the desired speech output as a result of the S2ST process.

Burmese is the official language of Myanmar, whereas Myeik is a dialect cluster mainly spoken in southern Myanmar. Previous studies have investigated speech recognition for the Myanmar language using both statistical and deep learning approaches [3], [4]. Statistical machine translation has also been studied for Myanmar-Myeik translation [5]. However, the resources, including speech and text data, for these languages are not as extensive as those of the major world languages, which complicates the process of developing a comprehensive speech-to-speech translation system.

Moreover, in a cascade approach, error propagation between different modules affects the outcome of the final translation. Speech recognition errors can propagate through the pipeline and cause errors in the final translation result [1]. For instance, if the ASR model recognizes a word in Burmese incorrectly, these recognition errors will be incorporated into the translation step, which may cause mistranslation errors at the word level, resulting in an incorrect Myeik word. Therefore, it is critical to assess not only the impact of each individual module on the final result but also the interaction between different modules in the cascaded system.

This paper introduces a cascaded Burmese-to-Myeik speech-to-speech translation system. The errors made at one stage can propagate to the following stages, leading to inaccurate results. For instance, the incorrect transcription of a Burmese word due to an error in the ASR model can cause a wrong translation in the MT component. As a result, errors that occurred at the ASR stage affect further processing, and the MT and TTS components cannot recover from them. It is especially challenging to attain a satisfactory balance of the components' performance as the system has to process various linguistic phenomena. Therefore, a multilingual cascaded system should be evaluated both quantitatively and qualitatively.

The paper proposes a cascaded Burmese to Myeik speech-to-speech translation system. A CNN-BiLSTM model with Connectionist Temporal Classification (CTC) is used for Burmese speech recognition [4]. For machine translation, the pretrained mBART-50 model [8] is fine-tuned using Burmese-Myeik parallel text. The MT model is further trained with both clean Burmese text and ASR-generated Burmese text to make it more suitable for the output produced by the ASR model. For Myeik speech generation, a pretrained Burmese MMS-VITS model [9] is fine-tuned using Myeik speech data. Finally, the ASR, MT, and TTS components are connected to form the complete Burmese-to-Myeik speech to speech translation system.

The main aim of the system is to develop and evaluate a complete Burmese-to-Myeik speech-to-speech translation system using limited language resources. The ASR, MT, and TTS components are first evaluated separately and are then tested as an integrated system. The effect of ASR errors on machine translation is also considered by using ASR-generated Burmese text during MT fine-tuning.

2. Related Work

Several studies have been carried out on speech recognition, machine translation, and speech synthesis for low-resource languages. For Myanmar languages, previous research has also explored different statistical and deep learning methods. This section reviews the studies that are closely related to the ASR, MT, and TTS components used in the proposed system.

2.1 Automatic Speech Recognition

Early research on Myanmar speech recognition mainly focused on statistical methods. Mon et al. [3] developed a continuous Myanmar speech recognition system using Hidden Markov Model (HMM)-based acoustic models with Gaussian Mixture Model (GMM) and Subspace Gaussian Mixture Model (SGMM). Their experiments used five hours of speech data. The HMM-SGMM model reported a WER of 7.6% for the closed test set and 31.9% for the open test set.

More recent studies have used deep learning for Myanmar speech recognition. Chit and Lin [4] developed end-to-end Myanmar ASR system using CTC with CNN and Bidirectional Long Short-Term Memory (BLSTM). Their study examined character, syllable, and sub word units using about 26 hours of Myanmar speech audios. Their best character-level CNN-BLSTM model reported a Character Error Rate (CER) of 4.72%, and the syllable-based model obtained an SER of 12.38%. Their work is closely related to the ASR component of this study because CNN, bidirectional recurrent layers, and CTC are also used for Burmese speech recognition.

2.2 Machine Translation

Machine translation between Myanmar and related languages has been studied using both statistical and neural approaches. Oo et al. [5] studied Statistical Machine Translation (SMT) between Myanmar and Myeik using a parallel corpus of approximately 10,000 sentences. Their experiments compared phrase-based, hierarchical phrase-based, and Operation Sequence Model (OSM) methods. Word-level and syllable-level segmentation were also examined. Among these methods, OSM produced the highest BLEU and RIBES scores, while syllable segmentation was found to be suitable for Myanmar-Myeik translation.

Later, Oo et al. [6] studied Neural Machine Translation (NMT) between Myanmar language and Dawei language which is another language variety of Myanmar. Their experiments used a parallel corpus of around 9,000 sentences and tested RNN and Transformer-based models with syllable segmentation. Their experiments showed that the RNN model using LSTM performed well for Myanmar-Dawei translation. This work showed that neural translation methods can also be applied to translation between Myanmar and a related low-resource language.

Pretrained multilingual models have also been developed to improve machine translation when training data are limited. Liu et al. [7] introduced mBART, which is pretrained on large-scale monolingual text from different languages. The pretrained model can then be adapted to a particular translation task through fine-tuning with parallel sentences. An extended version, mBART-50, supports multilingual translation across a larger number of languages [8]. In this study, mBART-50 is used as the base model for Burmese-to-Myeik translation module and is fine-tuned using prepared Burmese-Myeik parallel sentences.

2.3 Text-to-Speech

Text-to-Speech (TTS) generates speech from written text. Recent advances in neural networks have enabled more natural speech synthesis. The neural networks have been trained on text-speech paired data; therefore, they learned what patterns in text correspond to what sound. The article proposes Tacotron 2, which is designed by Shen et al. [10], uses a sequence-to-sequence architecture together with a neural vocoder. The model converts text into a Mel-spectrogram, which is then used to reconstruct the speech waveform.

Later, Kim et al. [11] later introduced VITS, an end-to-end model that performs speech synthesis and waveform generation in a single model. Unlike systems that require a separate vocoder, VITS can directly generate speech waveforms from text. Their experiments showed that VITS could produce high-quality speech.

For multilingual and low-resource speech processing, Pratap et al. [9] introduced the Massively Multilingual Speech (MMS) project, which provides speech models for a large number of languages. MMS includes text-to-speech models for more than one thousand languages, including Burmese. The

availability of a pretrained Burmese speech model is useful when only a limited amount of target speech data is available. In this study, the pretrained Burmese MMS-VITS model is further fine-tuned using Myeik speech data to generate Myeik speech.

2.4 Speech-to-Speech Translation

Speech-to-speech translation (S2ST) is one of the techniques that converts speech of one language to another language. One of the approaches is the cascaded method that combines automatic speech recognition (ASR), machine translation (MT), and text-to-speech synthesis (TTS). Each component performs its task-specific function and passes the output to the subsequent component. The cascade approach allows for separate training and evaluation of each component.

Recent research also proposed direct speech-to-speech translation models. For example, Jia et al. [2] proposed Translatotron, an end-to-end model that directly translates speech from source language to target language without any explicit intermediate transcription. Translatotron 2 [12] is the enhanced version of the first model that aims to achieve better quality of the translated speech and preserve the characteristics of the original speaker. The findings of [2] and [12] demonstrated that it is possible to implement direct speech-to-speech translation and eliminate the need for the pipeline that consists of ASR, MT, and TTS.

Large multilingual speech translation systems have also been introduced. SeamlessM4T [13] was developed to support different speech and text translation tasks within a multilingual model, including speech-to-speech translation. Such models can support many languages, but developing speech translation for low-resource languages remains difficult when suitable speech and translation data are limited.

In this study, a cascaded approach is used for Burmese-to-Myeik speech-to-speech translation. This approach allows the Burmese ASR, Burmese-to-Myeik MT, and Myeik TTS components to be developed and evaluated separately before they are connected. It also makes it possible to examine how errors from the ASR component affect the following MT component. To reduce this problem, the MT model is further trained using ASR-generated Burmese text together with clean Burmese text.

3. Proposed Methodology

In this study, we proposed the cascaded Burmese to Myeik speech-to-speech translation system. This system comprises of three parts: Burmese Automatic Speech Recognition (ASR), Burmese to Myeik Machine Translation (MT), Myeik Text-to-Speech (TTS). The input of the system consists of Burmese speech utterance, and the output is the Myeik speech utterance. The flow of the system is that the Burmese speech utterance is converted into Burmese text by the Burmese ASR model and then translated into Myeik text by the Burmese to Myeik MT model. Finally, the Myeik text is converted into the Myeik speech by the Myeik TTS model.

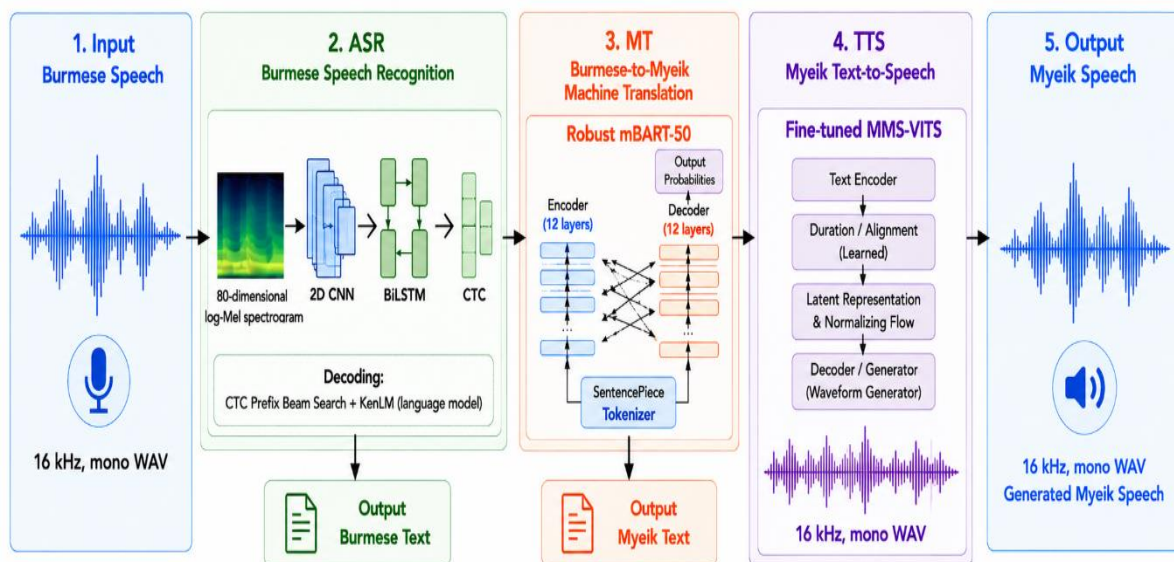


Fig. 1. Overall Architecture of the Proposed Burmese-to-Myeik Speech-to-Speech Translation System

The Burmese ASR model was constructed with 2D CNN-BiLSTM-CTC architecture. The pretrained mBART-50 model was fine-tuned with a Burmese-Myeik parallel text corpus for the MT model. In addition, both clean Burmese sentences and Burmese sentences recognized by ASR model were used for fine-tuning to enhance the performance of the MT model, which is suitable for the ASR recognition results with possible recognition errors. The pretrained Burmese MMS-VITS model was finetuned with Myeik speech utterances for the TTS model.

3.1 Dataset and Preprocessing

Three types of datasets were used for constructing the proposed system; Burmese speech utterances for the ASR model, Burmese-Myeik parallel text for the MT model, and Myeik speech utterances for the TTS model. In addition, each component of the system was trained with data preprocessed with different procedures since each part plays a different role in the speech-to-speech translation system.

For the ASR model, Burmese speech utterances were collected from two publicly available datasets: the OpenSLR-80 Burmese speech dataset [14] and the Myanmar Speech Dataset for ASR available through Hugging Face [15]. They were converted into 16kHz mono audio, and the utterances with the same content were removed. The text transcriptions were also normalized for training the ASR model. An 80-dimension log-Mel spectrogram was extracted from audio utterances and used for training the ASR model. The text transcriptions were divided into syllables and converted into label ids for CTC training. The data augmentation was also applied to increase the diversity of the training data.

A parallel Burmese-Myeik text corpus was prepared for the MT model with the help of Myeik speakers. We collected 2,936 sentence pairs of Burmese-Myeik texts. After cleaning and normalizing the text corpus, the text was divided into sub word units with the mBART-50 tokenizer. Unlike the ASR model, syllable segmentation was not applied to the MT model. The dataset was divided into training (2,348), validation (294), and test (294) sets.

For the TTS model, Myeik speech utterances and their text transcriptions were prepared. Although we collected speech utterances from various speakers at the beginning of the experiment, the final trained TTS model was built with data recorded from one speaker to make the speech style consistent. Speech utterances were converted into 16kHz mono audio, and text transcriptions were normalized before training the TTS model. Then, the pretrained MMS-VITS model was fine-tuned with the training data.

Table 1 - Summary of Datasets Used in the Proposed System

Component	Dataset	Data Type	Purpose
ASR	Burmese speech dataset	Speech + Burmese transcription	Burmese speech recognition
MT	2,936 Burmese-Myeik pairs	Parallel text	Burmese-to-Myeik translation
TTS	Myeik speech dataset	Myeik text + speech	Myeik speech generation

3.2 Burmese Automatic Speech Recognition

The first component of the proposed system is the Burmese ASR model, which converts the input Burmese speech into text. The model uses a 2D CNN-BiLSTM-CTC architecture. The input speech is first converted into an 80-dimensional log-Mel spectrogram, which represents the speech signal in both time and frequency dimensions.

The two-dimensional convolutional neural network (2D CNN) layers are applied to the log-Mel spectrogram to capture meaningful local features. The convolution operation can generally be represented as

$$y_i = b + \sum_{k=0}^{K-1} x_{i,s+k} + w_k \quad (1)$$

where y_i represents the output at position i , x represents the input feature, w_k represents the convolution filter weight, K is the kernel size, s is the stride, and b is the bias. The CNN layers learn local features from the input log-Mel spectrogram.

The extracted features are then passed to a two-layer Bidirectional Long Short-Term Memory (BiLSTM) network to learn the sequence information in the speech. The BiLSTM processes the feature sequence in both directions, allowing information from previous and following speech frames to be used together. The output at each time step is represented as

$$h_t = [\vec{h}_t; \overleftarrow{h}_t] \quad (2)$$

where $\vec{h}_t$ and $\overleftarrow{h}_t$ denote the forward and backward hidden states at time step t , respectively. By combining the information from both directions, the model can use the surrounding speech information when predicting the output syllables.

The BiLSTM output is then passed through a fully connected layer to obtain the probability distribution over the syllable labels. Connectionist Temporal Classification (CTC) [19] is used for training because the speech frames and their corresponding transcription are not directly aligned. For an input speech sequence x and a target transcription y , the CTC probability is defined as

$$P(y|x) = \sum_{\pi \in B^{-1}(y)} P(\pi|x) \quad (3)$$

Here, π denotes a possible alignment path, while B converts the path into the final label sequence by removing repeated labels and blank tokens. The probability of the target transcription is obtained by summing the probabilities of all valid paths that produce the same transcription. The CTC loss used during training is calculated as

$$\mathcal{L}_{CTC} = -\ln p(l|x) \quad (4)$$

During training, this loss is minimized so that the predicted syllable sequence becomes closer to the reference transcription.

Before training, the Burmese transcriptions are segmented into syllables, and each syllable is assigned a unique integer ID. A blank token is also included for CTC training. To provide more variation in the speech data, speed and volume changes are applied during data augmentation. After the model produces its predictions, CTC prefix beam search with a KenLM language model is used to generate the final Burmese transcription.

3.3 Burmese-to-Myeik Machine Translation

The second part of the proposed system is the Burmese-to-Myeik MT. In this work, the pretrained mBART-50 model [8] is utilized as the base translation model. As mBART-50 does not support separate language code for Myeik, the Burmese language code (**my_MM**) is used for both the source and target sides as Burmese and Myeik use the same Myanmar script.

Fine-tuning is carried out using the prepared parallel sentences of Burmese and Myeik. In the model, first each Burmese sentence is tokenized by the mBART-50 tokenizer. The encoder processes the given Burmese sentence in a sequence of sub words sub word tokens. As mBART-50 is based on the Transformer architecture, the attention mechanism can be represented as [16]

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (5)$$

In Equation (5), Q , K , and V denote the query, key, and value matrices, while d_k is the dimension of the key vectors. The decoder generates the Myeik translation sequentially, with each new token depending on the previously generated tokens. Given a Burmese input sentence X and its Myeik target sentence Y , the translation probability can be represented as

$$P(Y|X) = \prod_{t=1}^T P(y_t | y_{<t}, X) \quad (6)$$

where y_t represents the target token at time step t , $y_{<t}$ represents the previously generated target tokens, and T is the length of the target sequence. During fine-tuning, the MT model is trained by minimizing the negative log-likelihood of the correct target tokens:

$$L_{MT} = - \sum_{t=1}^T \log P(y_t | y_{<t}, X) \quad (7)$$

where L_{MT} represents the machine translation training loss. By minimizing this loss, the model is able to generate the correct Myeik tokens for the given Burmese input sentence.

Several saved checkpoints are evaluated on the BLEU score [17] and the checkpoint with the highest BLEU score is selected as an initial MT model.

3.3.1 Robust MT Fine-Tuning

In a cascaded speech-to-speech translation system, the input given to the MT model is not always clean written Burmese. The Burmese text generated by the ASR model may have missing, wrong, or substituted syllables. These recognition errors can affect the following translation process.

To make the MT model more suitable for this type of input, additional fine-tuning is performed using both clean Burmese sentences and ASR-generated Burmese sentences. In both cases, the corresponding correct Myeik sentence is used as the target. Therefore, the model learns from normal Burmese input as well as input containing errors like those produced by the ASR component.

For example, the training pairs can be represented as:

- 1) Clean Burmese $\rightarrow$ Correct Myeik
- 2) ASR-generated Burmese $\rightarrow$ Correct Myeik

The selected original mBART-50 checkpoint is used as the starting model for this additional training. The resulting model is used as the final MT component of the complete S2ST pipeline.

3.4 Myeik Text-to-Speech

The final component of the proposed system is the Myeik Text-to-Speech model, which converts the translated Myeik text into speech. During the earlier experiments, Tacotron2 [10] was trained using manually prepared Myeik speech dataset. However, the generated speech was mostly unclear, especially for sentences that were not included in the training data. Therefore, a pretrained speech synthesis model was used for the final TTS component.

The pretrained Burmese MMS-VITS model [9] was chosen as the base model. As Burmese and Myeik use the same Myanmar script, the pretrained Burmese model could be used as a good starting point for training the model on Myeik speech. The model was then fine-tuned using the prepared Myeik text-speech pairs.

VITS [11] uses latent variables to learn the relationship between the input texts and their corresponding speech. The difference between the posterior distribution obtained from speech and the prior distribution conditioned on text can be measured using the Kullback-Leibler (KL) divergence as

$$D_{KL}(q(z|x)||p(z|c)) = \mathbb{E}_{q(z|x)} \left[\log \frac{q(z|x)}{p(z|c)} \right] \quad (8)$$

where z represents the latent variable, x represents the speech information, c represents the text condition, $q(z|x)$ represents the posterior distribution, and $p(z|c)$ represents the prior distribution conditioned on the text. The KL divergence helps the model learn a relationship between the speech representation and the text-conditioned representation.

For the final fine-tuning, speech data from a single speaker were used to maintain consistency in voice and speaking style. Each training sample consists of a Myeik text sentence and its corresponding speech recording. The Myeik text is processed by the tokenizer, while the corresponding speech recordings are prepared at a sampling rate of 16 kHz. The pretrained MMS-VITS model [9] is then fine-tuned on the Myeik speech dataset for 3,000 training steps.

After the fine-tuning process, the trained model is used to generate Myeik speech. The translated Myeik text produced by the MT model is given as input to the TTS model, which generates the corresponding Myeik speech waveform. During inference, the translated Myeik sentence is passed to the fine-tuned MMS-VITS model, which generates the corresponding speech waveform. This model is used as the final TTS component in the complete Burmese-to-Myeik S2ST system.

3.5 Complete Speech-to-Speech Translation Pipeline

The three developed models are integrated to form the complete Burmese-to-Myeik speech-to-speech translation system. In the integrated system, the result from each model is passed to the next model for further processing.

First, a Burmese speech utterance is given as an input to the ASR model. The trained 2D CNN-BiLSTM-CTC model recognizes the speech and gives the corresponding Burmese text as output. The recognized Burmese text is then fed into the fine-tuned robust mBART-50 model, which translates the Burmese text into Myeik text. Next, the translated Myeik text is passed to the fine-tuned MMS-VITS model. The TTS model converts the translated text into Myeik speech, which is the final output of the system. The complete process can be represented as follows:

Burmese Speech → CNN-BiLSTM-CTC → Burmese Text → mBART-50 → Myeik Text → MMS-VITS → Myeik Speech

Since the three models work sequentially, errors from one model can also affect the following model. Recognition errors produced by the ASR model may affect the Myeik translation generated by the MT model. Therefore, the robustly fine-tuned mBART-50 model is used in the final pipeline to make the MT component more suitable for ASR-generated Burmese text.

4. Experimental Setup and Evaluation

The ASR, MT, and TTS components were first tested individually before they were combined into the complete speech-to-speech translation system. Each component was evaluated using a method suitable for its task. CER and SER were used for the ASR model, while BLEU was used for MT evaluation. The complete system was then tested by passing the output of one component to the next component.

4.1 ASR Experimental Setup

The Burmese ASR dataset was split into training, validation, and test sets with a ratio of 80:10:10. The CNN-BiLSTM-CTC model was trained with a batch size of 16, while a batch size of 8 was used during validation and testing. Checkpoints were saved at different epochs during training so that their performance could be evaluated and compared.

The saved ASR checkpoints were evaluated on the validation data, and the model at epoch 40 was selected for the final evaluation. For decoding, CTC prefix beam search was used together with a KenLM language model [18]. The beam size was set to 60, with a language model weight (alpha) of 0.7 and a word insertion weight (beta) of 1.75.

The final ASR model was evaluated on an unseen test set containing 295 Burmese speech samples. The recognition performance was measured using CER and SER by comparing the ASR output with the reference transcription. CER is calculated as

$$CER = \frac{S_c + D_c + I_c}{N_c} \quad (9)$$

In Equation (9), S_c , D_c , and I_c denote character substitutions, deletions, and insertions, respectively, while N_c represents the total number of characters in the reference transcription. Similarly, SER is calculated as

$$SER = \frac{S_s + D_s + I_s}{N_s} \quad (10)$$

For SER, S_s , D_s , and I_s represent syllable substitutions, deletions, and insertions, while N_s is the total number of syllables in the reference transcription. The settings used for the final ASR evaluation are presented in Table 2.

Table 2 – Final ASR Evaluation Settings

Setting	Value
Selected Checkpoint	Epoch 40
Number of Test Samples	295
Decoding Method	CTC Prefix Beam Search + KenLM
Beam Size	60
KenLM Alpha	0.7
KenLM Beta	1.75

4.2 MT Experimental Setup

The Burmese-to-Myeik parallel dataset consists of a total of 2936 sentence pairs. Dataset pairs are then divided into 2348 for training, 294 for validation, and 294 for testing the final performance. The pretrained mBART-50 model [8] is fine-tuned on the prepared dataset using a batch size of 1 due to GPU memory limitations. Gradient accumulation is set to 4, and the model is trained for five epochs with a learning rate of 3×10^{-5} .

The model checkpoints were stored every 500 steps during the training process, which saw the BLEU score evaluated for each model to identify the most accurate checkpoint in terms of translation quality. The model at step 1000 was found to be the best-performing one and was selected as the initial Burmese-to-Myeik MT model. This model was also used as a starting point for the robust MT fine-tuning procedure.

For robust fine-tuning, both clean Burmese sentences and ASR-generated Burmese sentences are included as source inputs, while their corresponding correct Myeik sentences are used as the target outputs. The combined robust training set contains 4,640 training examples. The selected mBART-50 model is then fine-tuned on this dataset for three epochs. The learning rate is set to 1×10^{-5} and Adafactor is used as the optimizer. A batch size of 1 with gradient accumulation of 8 is used, giving an effective batch size of 8. FP16 and gradient checkpointing are also used during this training stage.

BLEU [17] is used to evaluate translation performance. The BLEU score is calculated as

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^N w_n \log p_n\right) \quad (11)$$

where BP is the brevity penalty, p_n is the modified precision for n -grams, w_n is the weight assigned to each n -gram, and N is the maximum n -gram order.

The original MT model is evaluated using clean Burmese sentences. The robust MT model is evaluated using ASR-predicted Burmese sentences to examine its performance when the input contains ASR errors. This comparison helps evaluate whether the robust MT model can handle errors produced by the ASR model. The main training settings for the original MT model and the robust MT fine-tuning are summarized in Table 3.

Table 3 – Training Settings of the Burmese-to-Myeik MT Models

Setting	Original MT	Robust MT
Starting Model	Pretrained mBART-50	Original MT checkpoint-1000
Training Data	Clean Burmese-Myeik pairs	Clean + ASR-generated Burmese input
Training Examples	2,348	4,640
Epochs	5	3
Optimizer	AdamW	Adafactor

Setting	Original MT	Robust MT
Learning Rate	3×10^{-5}	1×10^{-5}
Batch Size	1	1
Gradient Accumulation	4	8
Effective Batch Size	4	8
FP16	Enabled	Enabled
Gradient Checkpointing	Enabled	Enabled
Evaluation	Every 500 steps	End of each epoch

4.3 TTS Experimental Setup

For the final Myeik TTS experiment, the pretrained Burmese MMS-VITS model [9] is fine-tuned using the prepared single-speaker Myeik speech dataset. Each sample contains a Myeik text sentence and its corresponding speech recording. The speech recordings are prepared at a sampling rate of 16 kHz before being used for model training.

The MMS-VITS model is fine-tuned for a total of 3,000 training steps. During training, the model is evaluated every 250 steps, and checkpoints are also saved every 250 steps. The training process does not use automatic best-checkpoint selection. Therefore, after completing the 3,000 training steps, the final saved model is used as the TTS model in the proposed system. The main settings used for Myeik TTS fine-tuning are summarized in Table 4.

Table 4 – Myeik TTS Fine-Tuning Settings

Setting	Value
Base Model	Pretrained Burmese MMS-VITS
Training Data	Single-speaker Myeik text-speech pairs
Sampling Rate	16 kHz
Speaker	Speaker 1
Training Steps	3,000
Evaluation Interval	Every 250 steps
Checkpoint Interval	Every 250 steps
Maximum Saved Checkpoints	4
Load Best Model at End	False
Model Used for Final System	Final model after 3,000 steps

During inference, the Myeik text is processed by the tokenizer and given to the fine-tuned MMS-VITS model [9]. The model then generates the corresponding Myeik speech waveform. For the complete S2ST system, the Myeik translation produced by the MT model is directly used as the input to this TTS model.

The generated speech is evaluated based on its intelligibility and naturalness. The TTS model is first tested separately using Myeik text sentences and is then evaluated as part of the complete Burmese-to-Myeik speech-to-speech translation pipeline.

4.4 End-to-End S2ST Evaluation

After evaluating the individual components, the complete Burmese-to-Myeik speech-to-speech translation system is tested by connecting the ASR, MT, and TTS models. In this evaluation, Burmese speech is given as the input to the system and the generated Myeik speech is obtained as the final output.

The Burmese speech is first processed by the ASR model to obtain the recognized Burmese text. Instead of using the correct Burmese transcription, the actual ASR-generated text is passed to the robustly fine-tuned mBART-50 model. This allows the MT component to be evaluated under the same input condition that occurs in the complete system. The generated Myeik translation is then passed to the fine-tuned MMS-VITS model to produce the final Myeik speech.

For the ASR-to-MT evaluation, the generated Myeik translations are compared with the reference Myeik sentences using the BLEU score [17]. A total of 293 matched test samples is used for this evaluation. In addition, randomly selected samples are passed through the complete ASR-MT-TTS pipeline to examine the final speech output.

The final Myeik speech is evaluated through listening-based evaluation to examine its intelligibility and naturalness. This evaluation is used to determine whether the complete system can successfully convert Burmese speech into understandable Myeik speech.

5. Results and Discussion

This section presents the experimental results of the individual ASR, MT, and TTS components and the integrated Burmese-to-Myeik speech-to-speech translation system. The performance of each component is first discussed separately, followed by the results obtained when the components are connected.

5.1 ASR Results

During ASR training, several saved checkpoints were evaluated using the validation set. Although epoch 20 obtained the lowest validation loss, epoch 40 achieved the lowest SER among the evaluated checkpoints. Since SER directly represents the syllable recognition performance, epoch 40 was selected for the following decoding experiments.

Table 5 – Validation Results of ASR Checkpoints

Epoch	Validation Loss	Validation SER
15	1.9152	0.4256
20	1.8495	0.4028
25	1.9294	0.3997
30	1.9692	0.3835
35	2.0255	0.3863
40	2.0437	0.3831

After selecting epoch 40, different decoding methods were tested. Greedy decoding achieved an SER of 39.33% and a CER of 30.02% on the test set. CTC beam search with a beam size of 10 was also tested without a language model and slightly reduced the SER to 39.02% and the CER to 29.75%.

For language-model-based decoding, KenLM [18] was added to the CTC prefix beam search. The decoding parameters were tuned using the validation set. First, different alpha values were tested while keeping the beam size at 10. An alpha value of 0.7 obtained the lowest validation SER and was selected. Next, beta was varied while keeping alpha fixed at 0.7, and beta = 1.75 obtained the lowest validation SER among the tested values.

After selecting alpha and beta, different beam sizes were evaluated on the validation set. As shown in Table 6, increasing the beam size generally reduced both SER and CER. Among the evaluated beam sizes, a beam size of 60 achieved the lowest validation SER and CER.

Table 6 – Effect of Beam Size on ASR Validation Performance

Beam Size	Validation SER	Validation CER
5	0.2903	0.2239
10	0.2744	0.2107
20	0.2685	0.2054
30	0.2637	0.2016
40	0.2618	0.1997
50	0.2604	0.1986
60	0.2591	0.1979

Based on these validation experiments, the final decoding configuration used CTC prefix beam search with KenLM [18], a beam size of 60, alpha = 0.7, and beta = 1.75. The selected epoch 40 model was then evaluated on the unseen test set containing 295 speech samples. The final model achieved a CER of 21.26% and an SER of 27.29%.

Table 7 – Final performance of the Burmese ASR Model

Evaluation Metric	Result
CER	21.26%
SER	27.29%

The results show that the use of beam search and the KenLM language model improved the recognition performance compared with greedy decoding. However, recognition errors are still present in the ASR output. These errors include incorrect, missing, and substituted syllables. Since the recognized Burmese text is passed directly to the MT component, these errors can also affect the following Burmese-to-Myeik translation.

5.2 MT Results

Several checkpoints of the original mBART-50 model are evaluated using the validation dataset and BLEU score [17]. Among the evaluated checkpoints, checkpoint 1000 achieved the highest validation BLEU score of 14.24. Therefore, this checkpoint is selected as the original MT model and is also used as the starting point for robust fine-tuning. The selected checkpoint was then evaluated separately on the clean Burmese-to-Myeik test set and achieved a BLEU score of 13.53.

For the integrated ASR-to-MT evaluation, the actual Burmese text generated by the ASR model is used as the input to the robustly fine-tuned MT model. A total of 293 matched samples is evaluated, and the model achieves a BLEU score of 13.09.

Table 8 - BLEU Scores of the Original MT Model Checkpoints

Checkpoint	BLEU Score
500	10.99
1000	14.24
1500	11.85
2000	12.72
2500	11.47

During the original MT fine-tuning, the validation loss reached its lowest value around step 1000. The saved checkpoints were also evaluated using BLEU on the validation set. Checkpoint 1000 achieved the highest validation BLEU score of **14.24** and was therefore selected for the following experiments. The selected checkpoint was then evaluated on the clean test set and achieved a BLEU score of **13.53**.

The selected checkpoint 1000 was further fine-tuned using both clean Burmese sentences and ASR-generated Burmese sentences. This additional training was performed to make the MT model more suitable for the imperfect Burmese text produced by the ASR component. The robust MT model was trained for three epochs, resulting in 1,740 training steps. The saved robust MT checkpoints were evaluated on the validation set, and checkpoint 1740 achieved the highest validation BLEU score of 14.94. Therefore, checkpoint 1740 was selected as the final robust MT model.

To evaluate the MT model under the actual condition of the proposed S2ST system, Burmese sentences generated by the ASR model were used as the input to the robust MT model. A total of 293 matched test samples were evaluated. The generated Myeik translations were compared with the corresponding reference Myeik sentences, resulting in a BLEU score of **13.09**.

Table 9 – MT Evaluation Results

Evaluation	Input	Samples	BLEU
Original MT	Clean Burmese test text	294	13.53
ASR-to-Robust MT	ASR-generated Burmese text	293	13.09

The two BLEU scores were obtained under different input conditions and therefore should not be directly compared. The original MT model achieved a BLEU score of **13.53** using clean Burmese test sentences. In contrast, the robust MT model achieved a BLEU score of **13.09** using Burmese sentences generated by the ASR model. Therefore, the robust MT result represents the translation performance under actual ASR input conditions rather than a direct improvement over the original MT result.

5.3 TTS Results

The Myeik TTS component was tested using both training and unseen Myeik text sentences. During the earlier experiments, Tacotron2 [10] was trained using the prepared Myeik speech data. However, most of the generated speech was unclear or contained noise. Some training sentences could be recognized, but the model did not perform well on unseen sentences. Therefore, Tacotron2 was not used in the final S2ST system.

To improve the speech generation result, the pretrained Burmese MMS-VITS model [9] was fine-tuned using the single-speaker Myeik speech dataset. After fine-tuning, the final MMS-VITS model was able to generate understandable Myeik speech for both training and unseen sentences. Compared with the earlier Tacotron2 experiments, the generated speech was clearer and the expected Myeik words could be recognized more easily. Therefore, the fine-tuned MMS-VITS model was selected as the final TTS component.

Although the generated speech is understandable, some outputs still have unnatural speaking rhythm and timing, and the speech quality is not completely natural. These problems may be related to the limited amount of single-speaker Myeik speech available for fine-tuning. Increasing the amount and variety of Myeik speech data may help improve the naturalness and pronunciation of the generated speech in future work.

For the complete S2ST system, the Myeik text generated by the robust MT model is directly passed to the fine-tuned MMS-VITS model. The generated Myeik speech is then used as the final output of the proposed system. A listening-based evaluation is also used to evaluate the intelligibility and naturalness of the final generated speech.

5.4 End-to-End S2ST

The final experiment evaluates the complete Burmese-to-Myeik speech-to-speech translation pipeline. Burmese speech is first converted into Burmese text by the ASR model. The recognized text is translated into Myeik using the robustly fine-tuned mBART-50 model, and the resulting Myeik text is then converted into speech using the fine-tuned MMS-VITS model.

Twenty samples are used to examine the output of the complete pipeline. The experiments show that the system can successfully process Burmese speech through all three components and produce Myeik speech as the final output. In many cases, the main meaning of the original Burmese sentence is maintained in the generated Myeik output.

However, the final output depends on the performance of all three components. Recognition errors from the ASR model can change the Burmese text passed to the MT model. These errors may then affect the Myeik translation, and the TTS model generates speech based on the translated text. Therefore, an error produced at an earlier stage can influence the final speech output.

The robust MT fine-tuning is used to make the MT model more suitable for ASR-generated Burmese sentences that may contain recognition errors. The selected robust MT model, checkpoint-1740, achieved a BLEU score of 13.09 when evaluated using 293 actual ASR-generated Burmese test sentences. However, recognition and translation errors are still present, and the generated Myeik speech can also contain unnatural rhythm or timing.

The final generated Myeik speech was generally understandable, although some outputs contained unnatural rhythm and timing. These problems can come from the TTS model itself, while incorrect words in the final speech can also result from errors produced by the ASR or MT components. This shows the effect of error propagation in the cascade system.

Overall, the experimental results show that the developed ASR, MT, and TTS components can be connected to form a complete Burmese-to-Myeik speech-to-speech translation system. However, further improvements in speech recognition accuracy, translation quality, and speech naturalness are needed to improve the overall quality of the system.

5.5 Limitations

Although the proposed system can perform Burmese-to-Myeik speech-to-speech translation, several limitations remain. First, the amount of available training data is limited, especially for Burmese-Myeik parallel text and Myeik speech. The limited amount of data can affect the translation quality of the MT model and the naturalness of the generated Myeik speech.

Second, the cascaded structure causes errors from one component to affect the following components. Recognition errors from the ASR model may result in incorrect Burmese text, which can then affect the Myeik translation produced by the MT model. Although the MT model is further fine-tuned using ASR-generated Burmese sentences, some translation errors remain.

Another limitation is related to the TTS component. The final MMS-VITS model is fine-tuned using single-speaker Myeik speech data. The generated Myeik speech is reasonably intelligible, but some sentences still contain unnatural rhythm, timing, and pronunciation. A larger and more consistent Myeik speech dataset may help improve the quality of the generated speech.

Finally, the current system accepts recorded speech and processes it step by step through the ASR, MT, and TTS models. It operates on individual speech utterances and does not perform translation in real time.

6. Conclusion

This paper presented a cascaded Burmese-to-Myeik speech-to-speech translation system consisting of ASR, MT, and TTS components. A CNN-BiLSTM-CTC model was developed for Burmese speech recognition, while a pretrained mBART-50 model was fine-tuned for Burmese-to-Myeik translation. The MT model was further fine-tuned using both clean Burmese sentences and ASR-generated Burmese sentences to make it more suitable for the output of the ASR model. For Myeik speech generation, a pretrained Burmese MMS-VITS model was fine-tuned using Myeik speech data.

In the final evaluation, the ASR model obtained a CER of 21.26% and an SER of 27.29%. For machine translation, the selected original MT model obtained a BLEU score of 13.53 on the clean Burmese test set. The selected robust MT model achieved a BLEU score of 13.09 when actual ASR-generated Burmese sentences were used as input. The fine-tuned MMS-VITS model was able to generate reasonably intelligible Myeik speech, although some unnatural rhythm and timing were still present.

The experimental results show that the three components can be integrated to translate Burmese speech into Myeik speech. However, errors produced by the ASR model can affect the following translation and speech generation stages. Future work can focus on collecting additional Burmese-Myeik parallel sentences and Myeik speech recordings. Further improvements to ASR accuracy and TTS speech naturalness can also help improve the overall system.

7. References

- [1] P. S. Krishna Paleti, B. Saluru, S. Vatturi, B. Yohoshiva and N. P. Challa, "Cascaded Speech-to-Speech Translation System Focusing on Low-Resource Indic Languages," *2024 3rd Edition of IEEE Delhi Section Flagship Conference (DELCON)*, New Delhi, India, 2024, pp. 1-6, doi: 10.1109/DELCON64804.2024.10867248.
- [2] Y. Jia et al., "Direct Speech-to-Speech Translation with a Sequence-to-Sequence Model," in *Proc. Interspeech*, 2019, pp. 1123–1127.
- [3] A. N. Mon, W. P. Pa, and Y. K. Thu, "Building HMM-SGMM Continuous Automatic Speech Recognition on Myanmar Web News," in *Proceedings of the 15th International Conference on Computer Applications (ICCA 2017)*, Yangon, Myanmar, pp. 446–453, 2017.
- [4] Chit, K.M.M., Lin, L.L. (2021). Exploring CTC Based End-To-End Techniques for Myanmar Speech Recognition. In: Vasant, P., Zelinka, I., Weber, G.W. (eds) *Intelligent Computing and Optimization. ICO 2020. Advances in Intelligent Systems and Computing*, vol 1324. Springer, Cham. https://doi.org/10.1007/978-3-030-68154-8_87
- [5] T. M. Oo, Y. K. Thu, K. M. Soe, and T. Supnithi, "Statistical Machine Translation between Myanmar and Myeik," in *Proceedings of the 12th International Conference on Future Computer and Communication (ICFCC 2020)*, Yangon, Myanmar, Feb. 26–28, 2020, pp. 36–45.
- [6] T. M. Oo, Y. K. Thu, K. M. Soe and T. Supnithi, "Neural Machine Translation between Myanmar (Burmese) and Dawei (Tavoyan)," *2020 IEEE Conference on Computer Applications (ICCA)*, Yangon, Myanmar, 2020, pp. 1-7, doi: 10.1109/ICCA49400.2020.9022813.
- [7] Y. Liu et al., "Multilingual Denoising Pre-training for Neural Machine Translation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, 2020.
- [8] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan, "Multilingual Translation with Extensible Multilingual Pretraining and Finetuning," *arXiv preprint arXiv:2008.00401*, 2020.
- [9] V. Pratap et al., "Scaling Speech Technology to 1,000+ Languages," *Journal of Machine Learning Research*, vol. 25, no. 97, pp. 1–52, 2024.
- [10] J. Shen et al., "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783, 2018.
- [11] J. Kim, J. Kong, and J. Son, "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 5530–5540, 2021.
- [12] Jia, Ye et al. "Translatotron 2: High-quality direct speech-to-speech translation with voice preservation." *International Conference on Machine Learning* (2021).
- [13] L. Barrault et al., "SeamlessM4T—Massively Multilingual & Multimodal Machine Translation," *arXiv preprint arXiv:2308.11596*, 2023.
- [14] Y. M. Oo et al., "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, 2020, pp. 6328–6339.
- [15] Chuu Htet Naing, "Myanmar Speech Dataset for ASR," Available: <https://huggingface.co/datasets/chuuhhtetnaing/myanmar-speech-dataset-for-asr>
- [16] A. Vaswani et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [17] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.
- [18] K. Heafield, "KenLM: Faster and Smaller Language Model Queries," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pp. 187–197, 2011.
- [19] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, 2006, pp. 369–376.