EJSIT
European Journal of Science, Innovation and Technology
ISSN: 2786-4936
www.ejsit-journal.com
Volume 4 | Number 3 | 2024

Toward Text-To-Speech in Low-Resource and Unwritten Languages by Leveraging Transfer Learning: Application in Viet Muong Closed Language Pair

Van-Dong Pham
Hanoi University of Mining and Geology, Vietnam

ABSTRACT

Text-to-speech systems require a lot of text and speech data to train models on. But with over 6,000 languages in the world, making text-to-speech systems for minority and low-resource languages is very difficult. Traditional text-to-speech has two parts: an acoustic model that predicts sounds from text and a vocoder that turns the sounds into waveforms. This paper proposes a text-to-speech system for languages with very little data to support minority languages. It uses three techniques: 1. Pre-training the acoustic model on languages with a lot of data, then fine-tuning on the low-resource language; 2. Using "knowledge distillation" to adapt the model to match a high-quality example voice; 3. Treating input text data for a minority language like Muong the same way as Vietnamese text data. We first learn linguistic features from Vietnamese speech data using a standard Tacotron 2 acoustic model. Then, we train the acoustic model on Muong speech data, starting from the weights of the Vietnamese model. The synthesized Muong speech has a naturalness score of 3.63 out of 5.0 and a Mel Cepstral Distortion of 5.133, based on 60 minutes of Muong data. These results show the effectiveness and quality of the Muong text-to-speech system, built with very little Muong language data.

Keywords: Computing methodologies → Speech Synthesis, Tacotron 2, low-resource languages, unwritten language, Muong speech, transfer learning

INTRODUCTION

Recently, text-to-speech (TTS) research has progressed in producing human-like and high-quality speech (van den Oord et al., 2016; Shen et al., 2018; Wang et al., 2017; Weiss et al., 2021; Yasuda, Wang, & Yamagishd, 2021). Moreover, to resolve the problem of labeled data resources, speech synthesis systems for low-resource and unwritten languages are being created and developed more regularly. There are approximately 2,982 languages that are not written among the living languages in the world. Unwritten languages have rarely been studied and have faced numerous difficulties, leading to their eventual disappearance.

Many investigations for low-resource languages have been conducted recently using a variety of methods, including applying speaker characteristics (Yang, Yeh, & Chien, 2022), modifying phonemic features (Do et al., 2022; Lux & Vu, 2022), and cross-lingual text-to-speech (Cai, Yang, & Li, 2023; Huang et al., 2022). Yuan-Jui Chen et al. introduced end-to-end TTS with cross-lingual transfer learning (Tu et al., 2019). The authors proposed a method to learn a mapping between source and target linguistic symbols because the model trained on the source language cannot be directly applied to the target language due to input space mismatches. By using this memorization mapping, pronunciation information can be kept throughout the transfer process. Sahar Jamal et al. (2022) used transfer learning for the experiments to take advantage of the low-resource scenario. The information obtained then trains the model with a significantly smaller collection of Urdu training data. The authors created standalone Urdu and learning systems using pre-trained Tacotron English and Arabic models as parent models. Marlene Staib et al. (2020) improved or matched the performance of many baselines, including a resource-intensive expert mapping technique, by swapping out

Tacotron 2's character input for a manageably small set of IPA-inspired features. This model architecture also enables the automated approximation of sounds that have not been seen in training. They demonstrated that a model trained on one language could produce intelligible speech in a target language even in the lack of acoustic training data. A similar approach (Wells & Richmond, 2021) is used in transfer learning, where a high-resource English source model is fine-tuned with either 15 minutes or 4 hours of transcribed German data. Data augmentation is a different approach that researchers apply to solve the low-resource language challenge (Comini et al., 2022; Huybrechts et al., 2021; Byambadorj et al., 2021). An innovative three-step methodology has been developed for constructing expressive style voices using as little as 15 minutes of recorded target data, circumventing the costly operation of capturing large amounts of target data. Firstly, Goeric Huybrechts et al. (2021) augment data by using recordings of other speakers whose speaking styles match the desired one. In the next step, they use synthetic data to train a TTS model based on the available recordings. Finally, the model is fine-tuned to improve quality.

Muthukumar and his colleagues have developed a technique for automatically constructing phonetics for unwritten languages (Muthukumar & Black, 2014). Speech synthesis may be improved by switching to a representation closer to spoken language than written language. Van-Dong Pham et al. (2022) developed a speech synthesis system for the Muong language, explicitly using an intermediate representation created by the Vietnamese language's automatic speech recognition system (ASR) and a machine translation model to translate from Vietnamese to the intermediate representation. Based on the syllable structure of the Muong language being similar to the Vietnamese language, the intermediate representation created from the ASR model works effectively (Phạm et al., 2022).

Using transfer learning techniques, we build a speech synthesis system for the Muong language, assuming that the input data for the acoustic model is Vietnamese phonemes. The acoustic model we use is the Tacotron 2 model (Shen et al., 2018), which converts phonemes to Mel- spectrogram features. We achieve high-fidelity and efficient speech synthesis by generating waveforms using the Hifigan model (Kong, Kim, & Bae, 2020) from the output of the Tacotron 2 model. The student model was initialized using the whole weight of the teacher model, the Tacotron 2 model, which was trained on 20 hours of Vietnamese data. The student models were trained on the Muong language dataset with different sizes. The speech synthesis model shows impressive results when trained with only one hour of Muong audio.

METHODOLOGY

We apply the same two-component speech synthesis system as the Tacotron 2 model (Shen et al., 2018), however, the vocoder component we use is Hifigan (Kong, Kim, & Bae, 2020) instead of a modified version of WavNet (van den Oord et al., 2016).

Acoustic Model

The acoustic model has a sequence-to-sequence architecture. It consists of an encoder, which produces symbolic tokens as an internal representation of the input signal, and a decoder, which converts the symbolic tokens into a Mel-spectrogram. Mel-frequency spectrograms are related to linear-frequency spectrograms or short-time Fourier transforms (STFTs) magnitudes. Because it is stable to phase in each frame, this representation is smoother than waveform samples and simpler to learn using a squared error loss. The Mel-spectrogram feature simulates human ear sound perception, which is sensitive at low frequencies and less sensitive at high frequencies. At the same time, noise is reduced due to decreasing the signal's amplitude at high frequencies. These features, an 80-dimensional audio Mel-spectrogram with frames every 12.5 milliseconds, capture not just word pronunciation but also many aspects of human speech, such as volume, speed, and intonation.

The network is composed of an encoder and a decoder with location-sensitive attention. The encoder transforms a phoneme sequence into linguistic features, which the decoder uses to generate a Mel spectrogram. The decoder constructs a Mel spectrogram based on its autoregressive recurrent neural network by decoding the input sequence one frame at a time. Figure 1 shows the particular architecture model that we apply. PostNet was created to enhance the Mel-spectrogram generated by the decoder, an essential network component (Do et al., 2022).

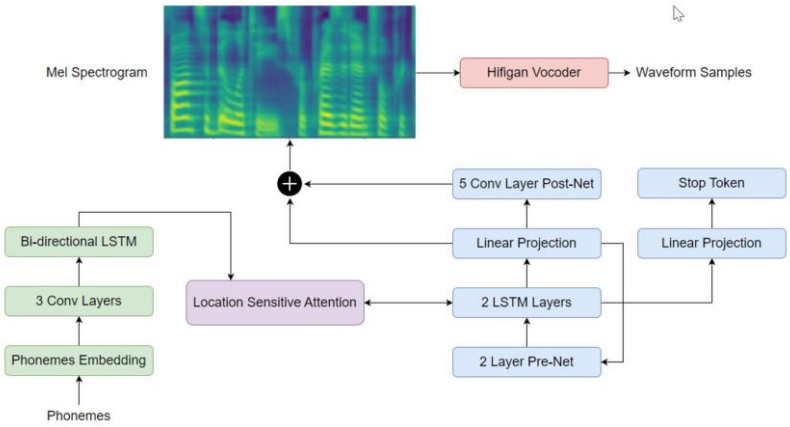

Figure 1: Block diagram of the speech synthesis system architecture

Hifigan Vocoder

HiFi-GAN (Kong, Kim, & Bae, 2020) consists of one generator and two multi-scale and multi-period discriminators. In order to increase training stability and model performance, the generator and discriminators are trained against each other with two additional losses. The generator takes a Mel-spectrogram as input and upsamples it using transposed convolutions until the length of the output sequence corresponds to the temporal resolution of raw waveforms.

A multi-period discriminator (MPD) is utilized for the discriminator, which is made up of several sub-discriminators, each of which handles a subset of the periodic signals of the input audio. Additionally, the multi-scale discriminator (MSD) suggested in Mel-GAN (Kumar et al., 2019) is utilized, which sequentially assesses audio samples at multiple levels to capture ongoing patterns and long-term relationships.

Grapheme-to-phoneme Conversion (G2P)

Muong and Vietnamese are phonologically monosyllabic languages. Muong is an unwritten language but has the same syllabic structure as Vietnamese (Van Dong et al., 2022; Phạm et al., 2022; Van Dong & Ha, 2022). In our processing system, Muong may be considered a variant of Vietnamese due to the similarities between the two languages. The syllable structure for Vietnamese and Muong languages is the same as below (Nguyen, Vu, & Luong, 2016):

$$Syllable = C_1 + [w] + V + C_2 + T \quad (1)$$

where C_1 is the initial consonant (onset), w is medial, V is a vowel, C_2 is the final consonant or semivowel (coda), and T is one of six tones of Vietnamese.

EXPERIMENTS

Dataset

Vietnamese data

We used approximately 20 hours of labeled Vietnamese audiobook data collected from open websites NgheAudio* and dtruyen†. The raw data consisted of long audio files (average duration of one hour) for each story chapter. After selecting the voiceover based on criteria such as clear voice and minimal noise, we chose Tran Van's voice for the story "Dai Mong Chu." The audio data underwent processing steps, including changing the sample rate to 22050 Hz, converting to a mono channel, and using pcm_s16le codec. The original audio files were segmented into smaller units based on signal segments containing inadequate voice, resulting in around 19,000 sentences of varying lengths. The duration distribution over the entire dataset after slicing into segments is shown in Figure 2.

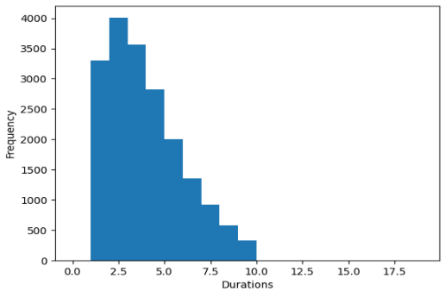

Figure 2: Duration histogram

The image illustrates that the Vietnamese audio dataset mainly consists of segments with lengths from 1s to 6s. To remove segments with background noise like music or ambient noise, the open-source inaSpeechSegmenter (Haldar & Mukhopadhyay, 2011) is used, resulting in a selection of clean audio tracks containing only the storyteller's voice. These selected segments are then labeled using an open-source Vietnamese Automatic Speech Recognition (ASR) model to obtain accurate labels for each audio segment, with WER ~ 10% on Vietnamese. To further improve accuracy, the Levenshtein distance algorithm is applied to correct any predicted label errors, while listening to the beginning and end of each long audio file before segmentation helps limit text space for comparison.

Table 1: Vietnamese and Muong dataset information

	Audiobook	**Muong recorded data**
Total duration	19 hours 58 minutes 30 seconds	4 hours 24 minutes 30 seconds
Total sentences	18885	1932
Total syllable	292841	62954
Total phonemes	1091384	307491
Distinctive syllable	3783	2934
Distinctive phone	44	44
Speaker name	Tran Van	Bui Viet Cuong
Speaker gender	Female	Male

* https://www.ngheaudio.org/truyen-audio-dai-mong-chu
† https://dtruyen.com/

The information on the entire audiobook data set is described in

Table 1. The number of distinctive phones here consists of 44 phonemes, including phones that represent silence (sil), end-of-sentence (eos), and padding (<pad>) used for shorter sentences within a batch during training.

Muong fine-tuning data

In the Muong language dataset of the ĐTĐLCN.20/17 project, Muong language data recorded by Bui Viet Cuong, a broadcaster from Hoa Binh Radio, was selected for transfer learning implementation. The details of the recorded dataset are described in Table 2. To investigate the relationship between the amount of training data and the quality of the synthesized speech output, we divided the high-quality recorded dataset into smaller training sets for fine-tuning purposes. The details of the smaller training sets are described in the table below:

Table 2: The Muong split data set

	M_15m	**M_30m**	**M_60m**
Total word	3581	7171	14458
Total phonemes	17559	35123	70477
Total syllable	1004	1333	1753
Distinctive phone	39	39	39
Num sentences	116	229	454
Total duration (min)	15	30	60

The training exercises are divided so that the maximum of the maximum coverage and the sentences are randomly taken. Looking at the above board, we can see the total number of phonemes increasing through the sets of M_15M, M_30M, and M_60M, corresponding to the data sets with a duration of 15 minutes, 30 minutes, and 60 minutes.

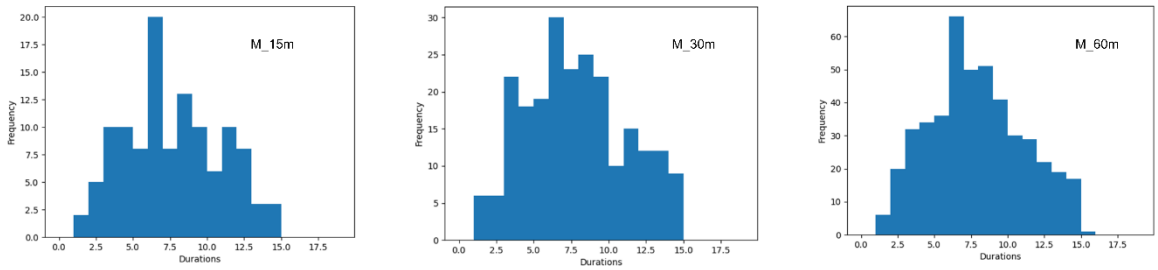

Figure 3: Duration distribution across the M_15m, M_30m, and M_60m datasets

In Figure 3, the duration is evenly distributed across the datasets and ranges from 1 to 15 seconds.

Practice Validate when training for all three training evaluations of 50 sentences, randomly taken from the Muong dataset and containing different from training data.

Training Procedure

First, we preprocess the input of the acoustic model, the Vietnamese text, into the phoneme representation. The process is carried out by applying the expression in section 2.3.

We used approximately 20 hours of Vietnamese audiobook data to train the acoustic model, which learns how to convert phoneme inputs into Mel spectrogram features. We use the Adam Optimization Algorithm as the neural Network optimization algorithm for the Acoustic Model. The parameters of the Adam optimizer are described in the table below:

Table 3: Parameter for optimizer

Optimization Hyperparameters	Value
Learning rate	0.0004
Weight_decay	0.000001
Grad_clip_threshold	1.0
Batch_size	16
Betas	(0.9, 0.999)
Eps	1e-08

The total number of training steps is 100,000 steps, and the model converges after approximately 50,000 steps.

Next, we trained the vocoder model on Vietnamese data using a pre-trained English model[‡]. The pre-trained model was trained on the English LJSPEECH dataset, which consists of approximately 24 hours of audio data, with 2.5 million training steps. The total number of training steps is 100,000 steps, and the model converges after approximately 20,000 steps.

Finetuning parameter table Hifigan model is described in
Table 3:

Table 3: Value of parameters when training Hifigan model

Hyperparameters	Value
Learning rate	0.0002
Learning rate decay	0.999
Optimizer	Adam
Batch_size	16
Betas (optimizer)	(0.9, 0.999)
Eps (optimizer)	1e-08
Sample rate (Hz)	22050

All models were trained on 1 GTX 2080 Ti GPU with a batch size of 16.

The training loss and validation loss during training of the acoustic model on Vietnamese data are shown in Figure 4.

Figure 4: Training loss and validation loss of pre-trained TTS model

The figure above shows that the model starts to converge from step 50k, as the loss on the validation set does not change significantly from this point. The darker line represents the

[‡] https://github.com/jik876/hifi-gan

smoothed curve with a smoothing value of 0.9, while the lighter line is the actual loss curve. The evaluation is based on this actual loss curve.

The loss curves during training with the HiFiGAN model are shown in the figure below:

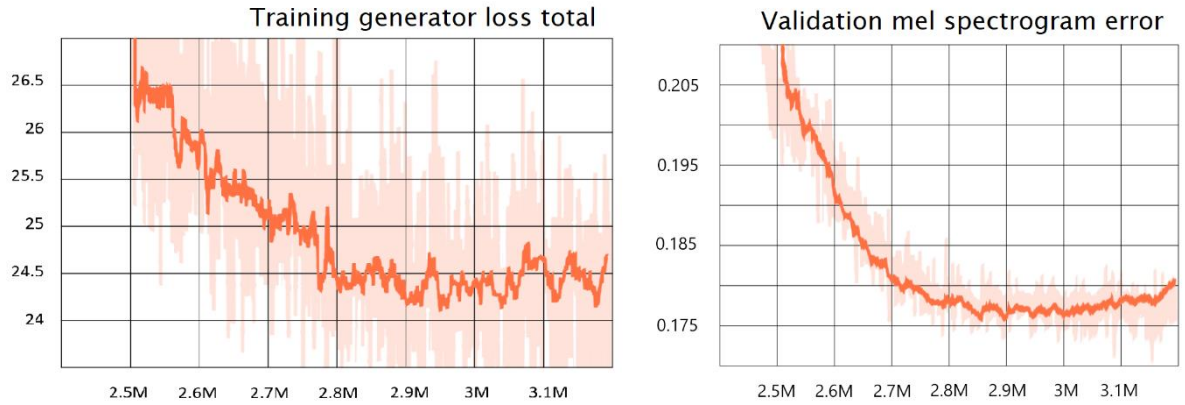

Figure 5: Training loss and validation error of Hifigan model

From the figure above, it can be seen that the vocoder model converges after about 300,000 steps during fine-tuning, and the total number of fine-tuning steps is close to 700,000 steps

Finetuned TTS model on Muong Datasets

After obtaining the pre-trained Tacotron 2 model, including the acoustic model and vocoder model, we performed finetuning on three different Muong language datasets from Hoa Binh province with different durations: M_15m, M_30m, M_60m, as described in section 3.1.

For the acoustic model, we finetune using a learning rate of 1e-04, and for the vocoder model, the learning rate is 2e-04. The process of training the Hifigan vocoder in the Muong language is similar to that of the Vietnamese language, which both use pre-trained English language and differ only in the languages used. Below are the training loss and validation loss plots during the finetuning process of the Tacotron 2 model on the Muong language datasets.

The loss curves during the training of the acoustic model on the M_15m dataset are shown in the following figure:

Figure 6: Training loss and validation loss of M_15m

The acoustic model is fine-tuned on the Muong language dataset M_15m for about 26,000 steps, and converges after approximately 20,000 steps.

Similarly, the loss curves during fine-tuning the acoustic model on the remaining two datasets, M_30m and M_60m, are shown in the figure below:

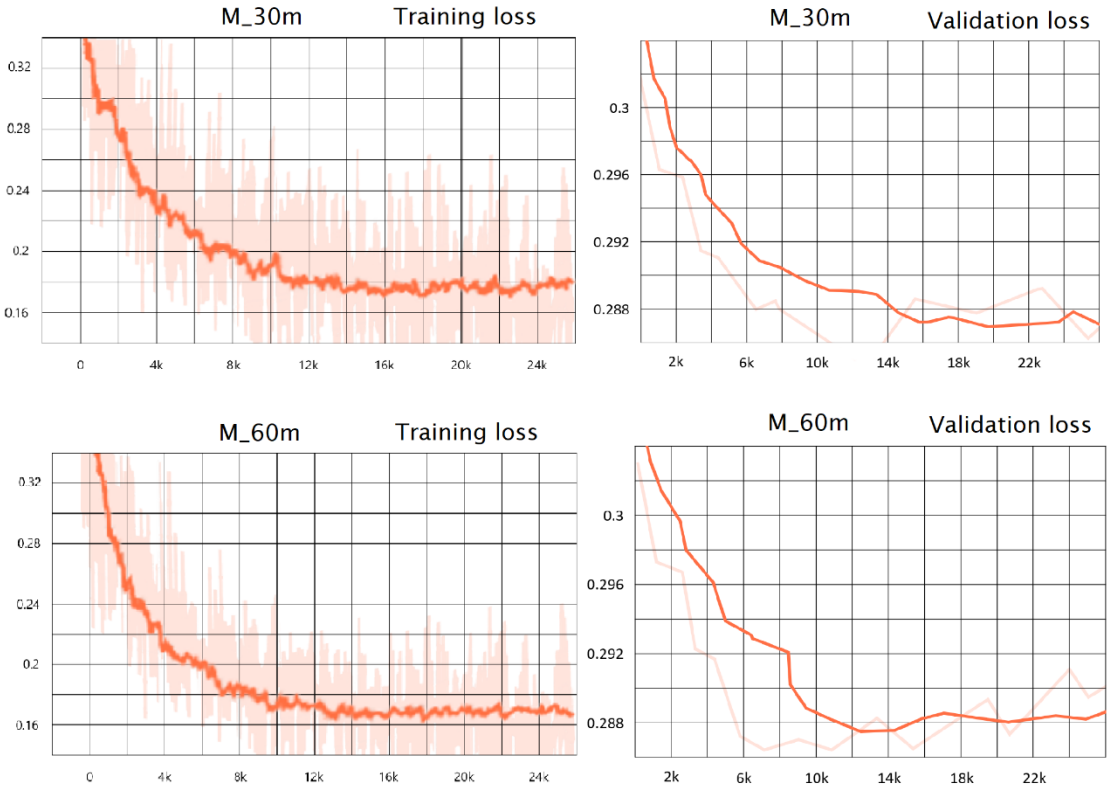

Figure 7: Training loss and validation loss of M_30m and M_60m

Both datasets are fine-tuned for about 26,000 steps, with the model converging after about 16,000 steps for M_30m and about 12,000 steps for M_60m.

Thus, all models converge when fine-tuning the Tacotron 2 model on different Muong language datasets, as shown in the above figures. However, to truly assess the quality of the synthesized Muong speech, let's proceed to the next section: Evaluation.

Evaluation

To examine the model's effectiveness when finetuning pre-trained models on different durations of Muong language datasets, we used 50 in-domain test sentences and 50 out-of-domain test sentences. Details of the two test sets are described in the following table:

Table 5: The specifications of the in-domain and out-domain test sets

	In-domain test set	**Out-domain test set**
Number sentence	50	50
Total word	1680	465
Total phonemes	8632	2063
Phonemes vocab	39	39
Total duration (min)	6.33	1.67

The in-domain test set is randomly selected from the recorded Muong dataset, ensuring that all phonemes are represented. The in-domain set consists of sentences collected from news sources, including newspapers, radio broadcasts, and current affairs. On the other hand, the

out-domain test set comprises daily conversation sentences, primarily short phrases that also cover all phonemes for a comprehensive assessment.

The Mean Opinion Score (MOS) was evaluated by a cohort of 50 Muong Hoa Binh native speakers. This cohort was balanced in terms of gender, with 25 males and 25 females participating in the study. The average age of the participants was 23.33 years old. In terms of educational attainment, half of the participants, 25 in number, held university degrees, while the remaining 25 had high school diplomas.

As part of the evaluation process, each participant was instructed to listen to 20 sentences comprising two sets. The first set included ten in-domain sentences covering topics like news, current affairs, and broadcasting. The second set consisted of 10 out-of-domain sentences reflecting daily communication scenarios. Each set of 10 sentences was randomly selected from a larger pool of 50 test sentences to ensure a diverse representation of linguistic contexts.

For the quantitative evaluation, we utilize the MCD DTW[§] (Mel Cepstral Distortion with Dynamic Time Warping) score, which measures the difference between two sequences of Mel cepstra. The smaller the score, the better the quality of the synthesized speech. While it is not a perfect metric to assess synthetic speech quality, it can be useful when combined with other measures. The MCD DTW score is calculated between the synthesized audio file and the original audio file, and the final score is averaged over 50 pairs for each set.

Table 6

	Test in-domain		Test out-domain	
	MOS	MCD (DTW)	MOS	MCD (DTW)
Ground Truth	4.36 ± 0.21	0.0	4.31 ± 0.22	0.0
M_15m	3.09 ± 0.45	6.875 ± 0.127	2.88 ± 0.45	7.125 ± 0.235
M_30m	3.27 ± 0.30	5.622 ± 0.214	3.08 ± 0.44	6.890 ± 0.161
M_60m	3.63 ± 0.36	5.133 ± 0.091	3.35 ± 0.36	6.521 ± 0.143

The MOS was used to evaluate the subjective quality of the speech samples from the different models. In the table provided, we observe a trend of improvement in MOS scores as we increase the training duration. This implies that with more training, the subjective quality of the synthesized speech increases.

For the in-domain test:

- Ground Truth: As the reference point for natural speech, the Ground Truth yielded the highest MOS score (4.36 ± 0.21).
- M_15m: With a MOS score of 3.09 ± 0.45, this model received the lowest score of the three, implying that the quality of the synthesized speech was not as good as the others.
- M_30m: An improvement from the M_15m model is seen with a MOS score of 3.27 ± 0.30. This suggests that additional training time improved the subjective quality of the synthesized speech.
- M_60m: This model achieved the highest MOS score (3.63 ± 0.36) among the synthesized models, indicating that the quality of the speech generated was the most appreciated by listeners, albeit not quite reaching the level of the natural speech.

For the out-of-domain test:

- Ground Truth: Again, the Ground Truth demonstrated the highest MOS score (4.31 ± 0.22).
- M_15m: The M_15m model had the lowest MOS (2.88 ± 0.45), suggesting that its synthesized speech was perceived as less satisfactory.

[§] https://github.com/SandyPanda-MLDL/ALGAN-VC-Generated-Audio-Samples

- M_30m: An increase in MOS is observed with a score of 3.08 ± 0.44, indicating a better speech quality perception compared to M_15m.
- M_60m: Mirroring the in-domain test, M_60m achieved the highest MOS score among the models (3.35 ± 0.36), though it still fell short of the natural speech.

In summary, the MOS scores demonstrate a noticeable improvement in the subjective quality of synthesized speech with increased training duration from 15 minutes to 30 minutes and then to 60 minutes. However, there is still a noticeable gap between the models and the natural speech, suggesting room for further improvement.

The Mel Cepstral Distortion (MCD) measured using Dynamic Time Warping (DTW) provides a quantitative metric that compares the difference between the synthesized speech and the natural reference speech. Lower MCD DTW values indicate a closer match to the natural reference speech, implying better synthesized speech quality.

The data table shows a clear trend of decreasing MCD DTW values as we move from the M_15m model to the M_60m model for both in-domain and out-domain tests. This suggests that the quality of the synthesized speech improves with increased training duration, becoming more similar to natural speech.

In the in-domain tests:

- The M_15m model exhibited the highest MCD DTW value (6.875 ± 0.127), suggesting its synthesized speech is most divergent from natural speech among the three models.
- The M_30m model showed an improvement over the M_15m model, with a lower MCD DTW value (5.622 ± 0.214). This indicates that its synthesized speech is closer to natural speech than the M_15m model.
- The M_60m model had the lowest MCD DTW value (5.133 ± 0.091) among the three models, suggesting its synthesized speech is closest to natural speech.

In the out-domain tests:

- Once again, the M_15m model had the highest MCD DTW value (7.125 ± 0.235), suggesting its synthesized speech is most divergent from natural speech among the three models.
- The M_30m model had a lower MCD DTW value (6.890 ± 0.161) compared to the M_15m model, indicating that its synthesized speech is closer to natural speech.
- Consistent with the in-domain tests, the M_60m model had the lowest MCD DTW value (6.521 ± 0.143), indicating that its synthesized speech is closest to natural speech in the out-domain context as well

The M_60m model achieved the best performance in terms of MCD DTW for both in-domain and out-domain tests, indicating that the synthesized speech can approach the quality of natural speech more closely with increased training duration.

We can see that the MOS scores increase when the training data increases and the MCD (DTW) scores decrease. When training with only 60 minutes of data, the quality of the synthesized audio is approximately the same as the original signal.

MOS Analysis by ANOVA

Applying two-way ANOVA, called ANOVA5 in our research, provides the means to test three distinct null hypotheses for the in-domain test set. The hypotheses for our ANOVA5 analysis, which considers two independent variables—TTS_System and Subject (Muong volunteers), are as follows:

- Null Hypothesis (H0) - TTS System: There is no significant variance in the mean of MOS attributable to the difference between the TTS systems being evaluated. In other words, the TTS system used does not significantly affect the MOS scores.

- Null Hypothesis (H0) - Subject: There is no significant variation in MOS scores across different Muong volunteers who are evaluating the synthesized speech. This implies that the subjectivity of the listeners does not significantly influence the MOS scores.
- Null Hypothesis (H0) - Interaction effect: There is no significant interaction effect between the TTS systems and the subjects on the resulting MOS scores. This means that the TTS system's combined effect and the volunteers' subjectivity do not significantly affect the MOS scores.

In our ANOVA6 analysis, we are considering two independent variables: TTS_System and Sentences. The null hypotheses for this analysis are as follows:

- Null Hypothesis (H0) - TTS System: There is no substantial difference in the Mean Opinion Scores (MOS) that can be ascribed to variations between the evaluated TTS systems. Essentially, the type of TTS system employed does not have a significant impact on the MOS.
- Null Hypothesis (H0) - Sentences: There is no considerable variation in MOS scores across different sentences used in the evaluation process. This suggests that the specific sentences chosen for the evaluation do not exert a significant influence on the MOS.
- Null Hypothesis (H0) - Interaction Effect: There is no noteworthy interaction effect between the TTS systems and the sentences on the derived MOS scores. This implies that the combined influence of the TTS system and the sentences used in the evaluation do not significantly alter the MOS.

Table 7: ANOVA Results for in-domain MOS Test

ANOVAs	Factor	df	f	p	η2
ANOVA5	TTS_System	3	116.321	0.000	0.162
	Subject	49	1.292	0.086	0.034
	TTS_System * Subject	49	0.789	0.968	0.061
ANOVA6	TTS_System	1	122.822	0.000	0.170
	Sentences	49	0.842	0.773	0.022
	TTS_System * Sentences	49	0.935	0.694	0.070

Table show the results of an ANOVA5 analysis of the Mean Opinion Scores (MOS) based on the hypotheses stated earlier.

- The first hypothesis being tested is whether there is a significant difference in MOS between TTS systems. The analysis shows that the factor "TTS_System" has a significant effect ($F = 116.321$, $p < 0.001$, $\eta2 = 0.162$), indicating that there is a significant difference in MOS between TTS systems.
- The second hypothesis being tested is whether there is a significant difference in MOS across different subjects. The analysis shows that the factor "Subject" does not have a significant effect ($F = 1.292$, $p = 0.086$, $\eta2 = 0.034$), indicating that there is no significant difference in MOS across different subjects.
- The third hypothesis being tested is whether there is an interaction effect between TTS systems and subjects on MOS. The analysis shows that there is no significant interaction effect between "TTS_System" and "Subject" on MOS ($F = 0.789$, $p = 0.968$, $\eta2 = 0.061$), indicating that the effect of TTS systems on MOS does not depend on the subject.

These results suggest that the MOS scores are affected by the TTS systems used but not by the subjects listening to the synthesized speech. These findings could be useful in improving the overall performance of TTS systems by identifying the specific factors that affect MOS scores and addressing them accordingly.

In the two-way ANOVA6, with the factors being the TTS_System and Sentences. The ANOVA results for the MOS variable show that both the TTS_System and Sentences factors have significant effects on the MOS measurements, as well as a significant interaction between the two factors:

- The TTS_System factor has a significant effect on the MOS measurements ($F = 122.822$, $p < 0.001$, $\eta 2 = 0.170$), indicating that the choice of TTS system has a significant impact on the MOS scores.
- The analysis shows that the factor "Sentences" does not have a significant effect ($F = 0.842$, $p = 0.773$, $\eta 2 = 0.022$), suggesting that there is no significant difference in MOS across different sentences.
- The analysis shows that there is no significant interaction effect between "TTS_System" and "Sentences" on MOS ($F = 0.935$, $p = 0.694$, $\eta 2 = 0.070$), indicating that the effect of TTS systems on MOS does not depend on the sentence.

Both the TTS_System and Sentences factors have significant effects on the MOS measurements, and their interaction is also significant. This suggests that the choice of the TTS system and the quality of sentences used in the study are important factors to consider in predicting MOS scores.

For the out-domain test set, we followed a similar methodology as implemented for the in-domain test set. The results derived from the two-way ANOVA7 and ANOVA8 analyses are showcased in

Table8.

Table 8: ANOVA Results for out-domain MOS Test

ANOVAs	Factor	df	f	p	η2
ANOVA7	TTS_System	3	121.343	0.000	0.168
	Subject	49	0.975	0.523	0.026
	TTS_System * Subject	49	1.029	0.394	0.077
ANOVA8	TTS_System	1	135.433	0.000	0.184
	Sentences	49	1.334	0.062	0.035
	TTS_System * Sentences	49	1.079	0.254	0.080

The results of the two-way ANOVA analyses for the out-domain test set are presented as follows:

For ANOVA7, where TTS_System and Subject are the independent variables:

- The main effect of the TTS system was found to be significant ($F = 121.343$, $p < .001$, $\eta 2 = .168$), suggesting a significant difference in Mean Opinion Score (MOS) between the different TTS systems.
- The effect of the Subject factor was insignificant ($F = 0.975$, $p = .523$, $\eta 2 = .026$), indicating no significant difference in MOS scores across different subjects who listened to the synthesized speech.
- There was also no significant interaction effect between TTS_System and Subject on the MOS scores ($F = 1.029$, $p = .394$, $\eta 2 = .077$).

For ANOVA8, where TTS_System and Sentences are the independent variables:

- The main effect of TTS_System was significant ($F = 135.433$, $p < .001$, $\eta 2 = .184$), suggesting a significant difference in MOS scores between different TTS systems.
- The effect of Sentences was not significant ($F = 1.334$, $p = .062$, $\eta 2 = .035$), implying that the Sentences in the evaluation do not significantly influence the MOS scores.
- There was also no significant interaction effect between TTS_System and Sentences on the MOS scores ($F = 1.079$, $p = .254$, $\eta 2 = .080$).

The results indicate that while the TTS system does significantly affect MOS scores, there is no substantial effect from the variable Subject or Sentences, nor any significant interaction effects between the independent variables.

Table 9: ANOVA Results for in/out domain MOS Test

ANOVAs	Factor	df	f	p	η2
ANOVA_In_Out_Domain	TTS_System	3	261.869	0.000	0.164
	Domain	1	27.276	0.000	0.007
	TTS_System *Domain	3	3.922	0.008	0.003

As indicated in
Table 9, a two-way ANOVA analysis was conducted to scrutinize the impact of the Text-to-Speech (TTS) System and the specific domain, as well as the interplay between them, on the resultant variable. Remarkable effects were identified for both the TTS System ($F = 261.869$, $p < 0.001$) and the domain ($F = 27.276$, $p < 0.001$) on the resultant variable, denoting a significant fluctuation in the response when varying TTS Systems and domains. Moreover, a noteworthy interaction effect emerged between the TTS System and the domain ($F = 3.922$, $p = 0.008$), inferring that the influence of the TTS System on the response variable is conditional on the domain. Put differently, the proficiency of diverse TTS Systems could differ based on the specific domain.

CONCLUSIONS

In this paper, applying the transfer learning technique to the Muong language using a pre-trained model on Vietnamese, a closely related language, with Tacotron 2 as the model has demonstrated promising results. With just one hour of Muong audio data, the model was able to generate natural-sounding speech with a relatively good MOS score of 3.89. This indicates the clear benefits of using transfer learning in the context of low-resource languages, as it saves time on data labeling and reduces the effort needed for manual annotation.

One significant aspect of this research is that TTS for the Muong language is not widely available, making this study one of the pioneering efforts in this area. The successful application of a limited amount of Muong data to create Muong TTS holds significant importance, as it contributes to preserving the Muong language's cultural heritage and opens up possibilities for extending TTS applications to other minority languages in Vietnam.

However, it should be noted that some limitations still need further exploration. For example, the applicability of transfer learning to languages that are not from the same language family, such as pre-trained models on English transferred to the Muong language, requires investigation. Additionally, the accuracy of pronouncing different phonemes between two languages during fine-tuning with limited data sources is also an area that requires further research. Addressing these issues in future studies will help advance the understanding and application of transfer learning in TTS for under-resourced languages.

ACKNOWLEDGMENTS

This work was supported by the Vietnamese national science and technology project: “Re-search and development automatic translation system from Vietnamese text to Muong speech, apply to unwritten minority languages in Vietnam” (Project code: ĐTĐLCN.20/17).

REFERENCES

Byambadorj, Z., Nishimura, R., Ayush, A., Ohta, K., & Kitaoka, N. (2021). Text-to-speech system for low-resource language using cross-lingual transfer learning and data augmentation. In *EURASIP Journal on Audio, Speech, and Music Processing*.

Cai, Z., Yang, Y., & Li, M. (2023). Cross-lingual multi-speaker speech synthesis with limited bilingual training data. *Computer Speech & Language*, *77*, 101427. https://doi.org/10.1016/j.csl.2022.101427

Comini, G., Huybrechts, G., Ribeiro, M. S., Gabrys, A., & Lorenzo-Trueba, J. (2022). Low-data? No problem: low-resource, language-agnostic conversational text-to-speech via F0-conditioned data augmentation. In *Interspeech*.

Do, P., Coler, M., Dijkstra, J., & Klabbers, E. (2022). Text-to-Speech for Under-Resourced Languages: Phoneme Mapping and Source Language Selection in Transfer Learning. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, European Language Resources Association, Marseille, France, 16–22.

Haldar, R., & Mukhopadhyay, D. (2011). Levenshtein distance technique in dictionary lookup methods: An improved approach. *arXiv preprint arXiv:1101.1232*.

Huang, W. P., Chen, P. C., Huang, S. F., & Lee, H. Y. (2022). Few-shot cross-lingual tts using transferable phoneme embedding. *arXiv preprint arXiv:2206.15427*.

Huybrechts, G., Merritt, T., Comini, G., Perz, B., Shah, R., & Lorenzo-Trueba, J. (2021). Low-resource expressive text-to-speech using data augmentation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 6593–6597.

Jamal, S., Rauf, S. A., & Majid, Q. (2022). Exploring Transfer Learning for Urdu Speech Synthesis. In *Proceedings of the Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia within the 13th Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, 70–74.

Kong, J., Kim, J., & Bae, J. (2020). HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (NIPS'20), Curran Associates Inc., Red Hook, NY, USA.

Kumar, K., Kumar, R., De Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., ... & Courville, A. C. (2019). MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis. In *Neural Information Processing Systems*.

Lux, F., & Vu, N. T. (2022). Language-Agnostic Meta-Learning for Low-Resource Text-to-Speech with Articulatory Features. *CoRR* abs/2203.03191. https://doi.org/10.48550/arXiv.2203.03191

Muthukumar, P. K., & Black, A. W. (2014). Automatic discovery of a phonetic inventory for unwritten languages for statistical speech synthesis. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2594–2598.

Nguyen, Q. B., Vu, T. T., & Luong, C. M. (2016). The Effect of Tone Modeling in Vietnamese LVCSR System. *Procedia Comput. Sci.*, *81*, 174–181. https://doi.org/10.1016/j.procs.2016.04.046

Phạm, V. Đ., Do, T. N. D., Mac, D. K., Nguyen, V. S., Nguyen, T. T., & Tran, D. D. (2022). How to generate Muong speech directly from Vietnamese text: Cross-lingual speech synthesis for close language pair. *Journal of Military Science and Technology, 81,* 138–147. https://doi.org/10.54939/1859-1043.j.mst.81.2022.138-147

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., ... & Wu, Y. (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE*

international conference on acoustics, speech and signal processing (ICASSP), IEEE, 4779–4783.

Staib, M., Teh, T. H., Torresquintero, A., Mohan, D. S. R., Foglianti, L., Lenain, R., & Gao, J. (2020). Phonological features for 0-shot multilingual speech synthesis. *ArXiv Prepr. ArXiv200804107*.

Tu, T., Chen, Y. J., Yeh, C. C., & Lee, H. Y. (2019). End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning. *ArXiv Prepr. ArXiv190406508*.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... & Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Van Dong, P., & Ha, V. T. H. (2022). Speech translation for Unwritten language using intermediate representation: Experiment for Viet-Muong language pair. *J. Mil. Sci. Technol., CSCE6*, 65–76.

Van Dong, P., Thanh, N. T., Do Dat, T., Ha, V. T. H., & Mai, D. T. (2022). Computational linguistic material for Vietnamese speech processing: applying in Vietnamese text-to-speech. *International Journal of Advanced Research in Computer Science*, *13*(6), 49-54.

Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., ... & Saurous, R. A. (2017). Tacotron: Towards end-to-end speech synthesis. *ArXiv Prepr. ArXiv170310135*.

Weiss, R. J., Skerry-Ryan, R. J., Battenberg, E., Marioryad, S., & Kingma, D. P. (2021). Wave-Tacotron: Spectrogram-free end-to-end text-to-speech synthesis. In *ICASSP*. Retrieved from https://arxiv.org/abs/2011.03568

Wells, D., & Richmond, K. (2021). Cross-lingual transfer of phonological features for low-resource speech synthesis. In *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*. https://doi.org/10.21437/SSW.2021-28

Yang, L. J., Yeh, I. P., & Chien, J. T. (2022). Low-Resource Speech Synthesis with Speaker-Aware Embedding. In *ISCSLP International Symposium on Chinese Spoken Language Processing*.

Yasuda, Y., Wang, X., & Yamagishd, J. (2021). End-to-end text-to-speech using latent duration based on vq-vae. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 5694–5698.