Multi-speaker TTS system for low-resource language using cross-lingual transfer learning and data augmentation

Zolzaya Byambadorj*, Ryota Nishimura*, Altangerel Ayush‡, Kengo Ohta† and Norihide Kitaoka††
*Tokushima University, Tokushima, Japan
E-mail: {C501947001, nishimura}@tokushima-u.ac.jp
‡Mongolian University of Science and Technology
E-mail: a.altangerel@must.edu.mn
†National Institute of Technology, Anan College
E-mail: kengo@anan-nct.ac.jp
†† Toyohashi University of Technology
E-mail: kitaoka@tut.jp

***Abstract*—We propose a multi-speaker text-to-speech (TTS) system for use with low-resource languages when only a very small amount of target language data is available. We investigate the effects of using high-resource language datasets and augmented data during model training, and compare different strategies for fine-tuning the model. After using a combination of cross-lingual transfer learning, a small amount of target language data and augmented target language data for training our low-resource TTS model, we then fine-tuned the pre-trained model using the original and augmented target language data. Our experimental results show that by sequentially training the model with high-resource language data, target language data and augmented target language data, followed by gradual fine-tuning using the original and augmented target language data, our system was able to achieve the most natural speech after text-to-speech conversion, achieving a native speaker mean opinion score of 3.50 on a scale of 0 to 5.**

I. INTRODUCTION

Recently proposed TTS models based on deep learning techniques [1-4] are capable of synthesizing natural, human-like speech. These models require a large amount of speech data for training however, as well as substantial computational power, thus data sparsity is a challenge when developing advanced TTS systems for low-resource languages. Transfer learning is one method used to solve the problem of limited target language data when developing low-resource TTS systems. TTS models are trained with a large amount of a different type of speech data, and the model is then adapted using a small dataset of a particular type of speech in the same language [5, 6]. In [7], investigators fine-tuned a TTS model pre-trained with English, using only 2.5 hours of Sanskrit data. Even though only a limited amount of target language data was used, they obtained good TTS results. A hierarchical transfer learning strategy has also been used to train TTS models for low-resource languages [8]. First, the TTS model was trained using one high-resource language, and was then fine-tuned using the low-resource language. Partial network-based transfer learning from the pre-trained monolingual TTS to a multilingual TTS was then applied, and finally, from the pre-trained multilingual TTS to a multilingual, style-transfer TTS, demonstrating that a multi-stage, transfer learning strategy was effective for TTS in low-resource target languages. In [9], TTS models learned the mapping between symbols of high- and low-resource languages, showing that transferring knowledge from a high- to a low-resource language is effective for low-resource TTS.

Another approach that can be used when only a limited amount of speech data is available is combining the speech of many different speakers and using this data to train a multi-speaker TTS system. Several studies have shown that multi-speaker models trained with a small amount of data from different speakers can achieve better performance than speaker-dependent TTS models. In [10], researchers proposed using a general deep neural network (DNN) to model multiple speakers for TTS, improving the quality of the synthesized speech compared with speech synthesized from a single-speaker, DNN-based TTS model. In [11], speech data from one speaker was combining with data from other speakers to obtain a high-performance TTS system, demonstrating that multi-speaker models trained using a small amount of data from different speakers are more effective than speaker-dependent models trained with more data. In [12], researchers investigated the best strategy for training multi-speaker TTS model using an existing, speaker-imbalanced corpus.

In addition, some studies have shown that multilingual model training can increase the naturalness of low-resource language speech. In [13], it was reported that the naturalness of target language speech was improved by using additional foreign language speech data during training, while [14] demonstrated that learned phoneme embedding vectors were located closer together when the pronunciations of these phonemes were similar across multiple training languages, thus pre-trained TTS models, trained using both high- and low-resource language datasets, improved the performance of the low-resource language TTS model.

In this paper, we propose a method which can be used to create high-performance TTS models for low-resource languages, and use Mongolian as our target language. In addition to using cross-lingual transfer learning and a multi-speaker model to solve the problem of data scarcity when training TTS models for a low-resource language, as mentioned in our survey of related work, we also propose the use of data augmentation to increase the volume of target language training data.

While some studies have used the transfer learning method to transfer knowledge from a large amount of speech data to a TTS model targeting a particular type of speech in the same language, e.g., emotional or whispered speech, etc., others have transferred knowledge from high-resource languages to TTS models targeting low-resource languages. In this study, we employ the latter strategy to train a multi-speaker model with a small amount of target language data from seven speakers. Augmented data was also generated using the original target language data from these seven speakers. First, we pre-trained the TTS model with multilingual data from two high-resource languages and with our low-resource target language data. We then fine-tuned the pre-trained model with the low-resource target language data. Augmented target language data was used for both pre-training and fine-tuning of the TTS model. Finally, we used only the original target language speech data to fine-tune the model further. Our experimental results show that this synthesis of training data, combined with gradual fine-tuning, improved the naturalness of the TTS model's output speech.

The contributions of this paper are as follows:

1. We construct a TTS model which can obtain reasonable results when using only approximately 4 minutes of speech data from each of 7 speakers, a total amount of target language data of about 30 minutes.
2. We investigate the use of high-resource language data to improve the performance of a low-resource TTS model, both alone and when combined with target language data.
3. We propose the use of augmented target language data and investigate how to most effectively use this data.
4. We propose a method of gradual fine-tuning to improve the performance of the low-resource TTS model.

The rest of this paper is organized as follows. A description of the TTS system used in our experiments, the datasets used for training, and a description of each of the methods evaluated are provided in Section 2. In Section 3, we explain our evaluation method, describe our experiments, and report the results of our subjective evaluations. Our conclusions are then presented in Section 4.

II. METHODS

A. TTS system

We used a Tacotron 2-based [2], multi-speaker TTS model with additional speaker embedding, using the same hyperparameters as in [2], except for the addition of a loss function for guided attention loss, which supports faster convergence, and a reduction factor (r = 2), which represents the number of frames to generate at each decoding step. Table I shows the hyper-parameters used in all models. We used a batch size of 64 for all of the models, except the final fine-tuned, multi-speaker models trained with only the original target language data. Fig. 1 shows an overview of our TTS system. Since it is also an issue in low-resource language scenarios to train the x-vector speaker encoder and neural vocoder, we used pre-trained models for both. We used the pre-trained x-vector for speaker embedding provided by Kaldi [15], and the speaker embeddings were concatenated with each encoder state. The pre-trained Parallel WaveGan neural vocoder trained on the LibriTTS dataset [16], provided by [17], was used to generate the waveforms in all of our experiments.

We used multilingual data to pre-train the TTS model. Since our TTS model takes phoneme input representations as input, overlapping phonemes are shared across languages.

TABLE I
HYPER-PARAMETERS AND NETWORK ARCHITECTURES

Feature extraction	
Sampling rate	24 kHz
Window size	85.3 ms (2048 pt)
Shift size	12.5 ms (300 pt)
Acoustic feature	log-mel spectrogram 80 dim
Encoder	
# phoneme embedding dimension	512
# CNN layers	3
# CNN filters	512
CNN filter size	5
# BLSTM layers	1
# BLSTM units	512
Decoder	
# LSTM layers	2
# LSTM units	1024
# prenet layers	2
# prenet units	256
# postnet layers	5
# postnet filters	512
Postnet filter size	5
# Speaker embedding dimension	512
Attention	
# Dimensions in attention	128
# Filters in attention	32
Filter size in attention	31
Sigma in guided attention loss	0.4
Reduction factor (r)	2
Optimization and minibatch	
Dropout rate	0.5
Zoneout rate	0.1
Learning rate	0.001
Optimization method	Adam with $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-6}$
Batch size	32 / 64
# Epochs	200

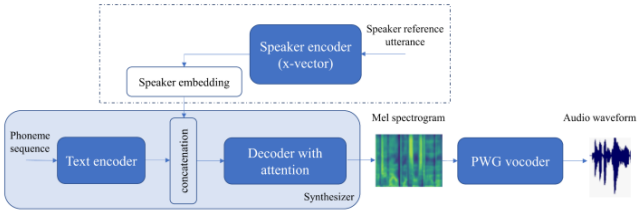

Fig. 1. Overview of the base TTS system

B. *Datasets*

We used English and Japanese as our high-resource languages, while Mongolian was the low-resource target language in our experiments. Multi-speaker datasets were used for both the high- and low-resource languages. We used a portion of both the LibriTTS [16] and JVS [18] corpora, in English and Japanese, respectively, as our high-resource, multi-speaker datasets. We randomly selected speakers and used approximately 24 hours of read speech data from the LibriTTS corpus and 10 hours of read speech data from the JVS corpus to create the high-resource language datasets used in all of our experiments. The data used in our experiments consisted of speech from 121 English speakers and 59 Japanese speakers, respectively. The sampling rate of each audio file was 24 kHz.

We also prepared a multi-speaker dataset for the Mongolian target language using speech from 7 speakers (two males and five females). Read speech data from each speaker included news stories, audiobooks and a Mongolian translation of the Bible. Most of the speech data was not high quality. We manually split each audio file into individual sentences and up-sampled each utterance to 24kHz. Our overall Mongolian speech data corpus included approximately 12 hours of read speech from the Bible and between 25 minutes to 1 hour of news and audiobook speech, however a total of only about 30 minutes of Mongolian speech data, about 4 minutes from each of our seven speakers, was included in the target language dataset used in our experiments.

We also generated approximately 15 hours of augmented target language data from the 30 minutes of multi-speaker, Mongolian speech data selected for use in our experiments. This was done by changing the pitch and speed of the original speech using the Sox utility [19]. The number of semitones of shift when changing the pitch was between -2.5 and 2.5, at steps of 0.5, while the ratio of the speed of the augmented speech to the speed of the original speech was within the range of 0.7 to 1.55 times that of the original speech, at steps of 0.05. Thus, we generated 27 variations of each target language utterance.

C. *Proposed methods*

We investigated the effects of transfer learning, data augmentation and various fine-tuning strategies on the low-resource target language TTS model. First, we created several different models by training them using the various datasets described in Section II-B, and fine-tuned some of them. Table II shows the pre-trained models we created (Model name), the datasets used to train each model (Stage 1) and the datasets used for fine-tuning each model (Stages 2 and 3). The first model, denoted as M1, was trained using only the original target language dataset. Model M2 was trained using the two high-resource, monolingual datasets in the first stage, then fine-tuned with the original target language dataset in the second stage. Model M3 was trained using the two high-resource, monolingual datasets and the original target language dataset, then fine-tuned with the same target language dataset. Models M4 and M5 were trained using the two high-resource, monolingual datasets, the original target language dataset and the augmented target language dataset. Model M4 was then fine-tuned with the original target language dataset in the second stage. To adapt the pre-trained model, we also used all target language data, consisting of the original target language dataset and the augmented target language dataset. As a result, we believe that it may yield further gains in performance. Therefore, model M5 was fine-tuned using our proposed gradual fine-tuning method, using the original and augmented target language dataset in the second stage and only the original target language dataset in the third stage.

TABLE II

DATASETS USED FOR TRAINING EACH MULTI-SPEAKER MODEL IN STAGE ONE, AND FOR FINE-TUNING THEM IN STAGES TWO AND THREE

Model name	1st stage	2nd stage	3rd stage
	Datasets		
M1	MN	-	-
M2	EN, JP	MN	-
M3	EN, JP, MN	MN	-
M4	EN, JP, MN, AD	MN	-
M5	EN, JP, MN, AD	MN, AD	MN

MN: original Mongolian dataset (7 speakers)
EN: portion of LibriTTS corpus (121 speakers)
JP: portion of JVS corpus (59 speakers)
AD: augmented data generated from original Mongolian dataset

III. EXPERIMENTS

A. *Evaluation*

We only evaluated the naturalness of the synthesized speech generated by the TTS models, at various stages, using a mean opinion score (MOS) subjective test. Six native Mongolian-speaking subjects were asked to rate the naturalness of the synthesized speech output using a scale of 1 to 5 in 0.5-point increments (1 = Bad, 2 = Poor, 3 = Fair, 4 = Good, 5 = Excellent). The tests were conducted in a lab environment. We used different transcriptions to synthesize the speech for each evaluation. One utterance was generated for each of the seven speakers represented in our target language dataset, thus seven utterances were generated from each of the four TTS models for each evaluation. As a result, each listener evaluated a total of 35 synthesized utterances and ground truth for naturalness.

B. *Results*

First, we investigated the effects of training TTS models with the various datasets described in Section II-B: a limited amount of target language data only, high-resource language data only, with both target language data and high resource language data, and with all of these datasets plus the augmented target language

data, as well as fine-tuning all but the first model with the limited amount of target language data. Table III shows the results of these evaluations. The M1 model, trained with only a small amount of multi-speaker, target language data, failed to synthesize intelligible speech. Note that, although the naturalness MOS of the M3-1st model, trained using the two high-resource language datasets and the same target language dataset during the first stage, was not high, it was able to synthesize speech. Therefore, the use of high-resource language data during training appeared to improve the performance of the low-resource TTS model. The naturalness MOS of the M2-2nd model was higher than that of the M3-1st model before fine-tuning, but after fine-tuning with the target language dataset, the performance of the M3-2nd model was better than before and outperformed the M2-2nd model. This indicates that our fine-tuning method improved the linguistic knowledge learned by the model during training. On the other hand, by comparing our 2nd stage results for the M2 and M3 models, we can see that by also using the small target language dataset to train the M3 model during Stage 1, we improved the naturalness of the output speech. The M4-1st model, which used the high-resource language datasets, the small target language dataset and the augmented target language data, achieved the best performance of the evaluated models, even before fine-tuning. The results in Table III show that transferring knowledge from the high-resource language datasets, the use of augmented target language data and fine-tuning with a small amount of target language data all improved the performance of the low-resource TTS model. We also observed that using the small target-language dataset when pre-training the model significantly impacted performance for further adjustment of the model through fine-tuning.

TABLE III
NATURALNESS MOS RESULTS WITH 95% CONFIDENCE INTERVALS FOR MULTI-SPEAKER TTS MODELS TRAINED WITH DIFFERENT DATASETS, WITH OR WITHOUT 2ND STAGE FINE-TUNING

Model	Naturalness MOS
Ground Truth	4.94 ± 0.077
M1	Failed
M2-2nd	2.40 ± 0.227
M3-1st	1.83 ± 0.171
M3-2nd	2.65 ± 0.197
M4-1st	**3.18 ± 0.097**

Second, we evaluated the effect of various fine-tuning methods on our M4 model, which had generated the most natural speech. Table IV shows the results for the M4 and M5 TTS models, the latter of which was identical to the M4 model before fine-tuning using the original target language data supplemented with augmented target language data, and then fine-tuned a second time using only the original target language data. The naturalness results for these models at various stages are very close, possibly because the models were trained with a sufficient amount of target language data due to the use of augmented data. Although the difference in the naturalness results for the various stages of the M4 and M5 models are small, all of fine-tuned models outperformed the M4-1st model, indicating that even though the model was trained with target language data and augmented target language data, fine-tuning the model with target language data improved performance. The M5-3rd model, using fine-tuning with original and augmented target language data, followed by further fine-tuning with the original target language data (i.e., gradual fine-tuning), achieved the best performance in our evaluation.

TABLE IV
NATURALNESS MOS RESULTS WITH 95% CONFIDENCE INTERVALS FOR MULTI-SPEAKER TTS MODELS TRAINED WITH ALL DATASETS, WITH OR WITHOUT FINE-TUNING USING VARIOUS METHODS

Model	Naturalness MOS
Ground Truth	4.95 ± 0.049
M4-1st	3.14 ± 0.047
M4-2nd	3.43 ± 0.135
M5-2nd	3.30 ± 0.133
M5-3rd	**3.50 ± 0.080**

IV. CONCLUSIONS

We proposed a method of training and fine-tuning multi-speaker TTS models for low-resource languages with a very limited amount of target language data available. In this study, we used only about 30 minutes of target language speech data in total, collected from 7 speakers, with an average of approximately 4 minutes of speech data per speaker. We used cross-lingual transfer learning and augmented data to resolve the issue of data scarcity. We also investigated the effect of using a combination of high-resource language data, limited target language data and augmented target language data during training, as well as various fine-tuning strategies using target language data. Although only a small amount of multi-speaker, target language data was used, we achieved a reasonable level of synthesized speech naturalness by using cross-lingual transfer learning and data augmentation. In the future, we will investigate additional methods of building multilingual TTS models for use in low-resource scenarios.

REFERENCES

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R.J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, R.A. Saurous, "Tacotron: Towards end-to-end speech synthesis," *Interspeech 2017,* pp. 4006–4010, 20-24 August 2017, Stockholm

[2] J. Shen, R. Pang, R.J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R.A. Saurous, Y. Agiomyrgiannakis, Y., Wu, "Natural TTS synthesis by conditioning Wavenet on mel spectrogram predictions," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783, 15-20 April 2018, Canada

[3] W. Ping, K. Peng, A. Gibiansky, S.O. Arik, A. Kannan, S. Narang, J. Raiman, J. Miller, "Deep Voice 3: Scaling text-to-speech with convolutional sequence learning," *6th International Conference on Learning Representations (ICLR)*, pp. 1094–1099, April 30-May 3, 2018, Vancouver, Canada

[4] J. Sotelo, S. Mehri, K. Kumar, J.F. Santos, K. Kastner, A. Courville, Y. Bengio, "Char2wav: End-to-end speech synthesis,"

5th International Conference on Learning Representations (ICLR), 24-26 April 2017, Toulon, France

[5] N. Tits, K. El Haddad, T. Dutoit, "Exploring transfer learning for low-resource emotional TTS," *Proceedings of SAI Intelligent Systems Conference*, pp. 52–60, 5-6 September 2019, London, United Kingdom

[6] B. Bollepalli, L. Juvela, P. Alku, "Lombard speech synthesis using transfer learning in a Tacotron text-to-speech system," *Interspeech 2019*, pp. 2833–2837, 15-19 September 2019, Graz, Austria

[7] A. Debnath, S.S. Patil, G. Nadiger, R.A. Ganesan, "Low-Resource End-to-End Sanskrit TTS using Tacotron2, WaveGlow and Transfer Learning," *2020 IEEE 17th India Council International Conference (INDICON)*, pp. 1-5, 11-13 December 2020, New Delhi, India

[8] A. Kurniawati, M. Adriani, J. Wisnu, "Hierarchical Transfer Learning for Multilingual, Multi-Speaker, and Style Transfer DNN-Based TTS on Low-Resource Languages," *IEEE Access Journal*, vol. 8, pp. 179798-179812, 2020

[9] T. Tu, Y.J. Chen, C.C. Yeh, H.Y. Lee, "End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning," *Interspeech 2019*, pp. 2075–2079, 15-19 September 2019, Graz, Austria

[10] Y. Fan, Y. Qian, F.K. Soong, L. He, "Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4475-4479, 19-24 April 2015, Brisbane, Australia

[11] J. Latorre, J. Lachowicz, J. Lorenzo-Trueba, T. Merritt, T. Drugman, S. Ronanki, V. Klimkov, "Effect of data reduction on sequence-to-sequence neural TTS" *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7075–7079, 12-17 May 2019, Brighton, UK

[12] H.T. Luong, X. Wang, J. Yamagishi, N. Nishizawa, "Training multi-speaker neural text-to-speech systems using speaker-imbalanced speech corpora," *Interspeech 2019*, pp. 1303–1307, 15-19 September 2019, Graz, Austria

[13] M. de Korte, J. Kim, E. Klabbers, "Efficient neural speech synthesis for low-resource languages through multilingual modeling," *Interspeech 2020*, pp. 2967–2971, 25-29 October 2020, Shanghai, China

[14] Y. Lee, S. Shon, T. Kim, "Learning pronunciation from a foreign language in speech synthesis networks," 2018

[15] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5329–5333, 15-20 April 2018, Canada

[16] H. Zen, V. Dang, R. Clark, Y. Zhang, R.J. Weiss, Y. Jia, Z. Chen, Y. Wu, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," *Interspeech 2019*, pp. 1526-1530, 15-19 September 2019, Graz, Austria

[17] T. Hayashi, Kan-bayashi/ParallelWaveGan, https://github.com/kan-bayashi/ParallelWaveGAN, 2019

[18] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, H. Saruwatari, "JVS corpus: free Japanese multi-speaker voice corpus," *arXiv:1908.06248*

[19] SoX: Sound eXchange audio manipulation tool. http://sox.sourceforge.net