Investigating the effectiveness of various speaker embeddings for multi-speaker end-to-end speech synthesis system using small-sized speech data

Anonymous ACL submission

Abstract

In this paper, we investigated the effectiveness of incorporating various speaker embeddings into an end-to-end speech synthesis system, for generating a unseen speaker's voice with small-sized speech data. To do so, we adopted learned speaker embeddings from various tasks, such as voice conversion and speaker verification. By combining the speaker embeddings using additive attention mechanism to an autoregressive-based speech synthesis framework, we could evaluate the performance of these embedding methods. To further enhance the speaker similarity and speech quality, the post-net for the output spectrogram sequence is replaced by a post-filter network. Experimental results showed that the proposed speech synthesis system with speaker embedding is capable of generating fluent arbitrary speech utterances of a unseen speaker with only few speech utterances. Besides, the post-filter network is helpful for enhancing the speaker similarity and speech naturalness of the output speech.

1 Introduction

In recent years, end-to-end speech synthesis framework has achieved great results in the respect of speech naturalness and fluency, especially for synthesizing speech of a speaker whose speech corpus is available for training the voice model. If speech data of a speaker is not large enough for training, adapting the available voice models is also quite successful. Conventionally, voice models are adapted supervisedly, which require data alignment or text transcription for the adaptation data. Recently, speaker embedding-based method could alleviate the problem. It encodes the speaker information directly from speech data without transcription, and the speaker embeddings are jointly trained with end-to-end text-to-speech (TTS) model. By applying different speaker embeddings, the generated speech is perceived as the corresponding speaker. This method is also popular in voice conversion and vocoder model training. Besides, multi-speaker speech synthesis could also be constructed by using generative adversarial network (GAN) framework (Kameoka et al., 2018) or autoencoder-based methods (Qian et al., 2019; Chou et al., 2019) to learn speaker-related information. However, for generating speech of an unseen speaker, jointly trained speaker encoder is not able to obtain the speaker information in the training phase, therefore it still requires model adaptation or transform learning (Jia et al., 2018; Chien et al., 2021) after model training.

There are works that tried to tackle this problem in order to generate multi-speaker TTS system for unseen speakers (Chou et al., 2019; Cooper et al., 2020). In (Chou et al., 2019), they focused on encoding speaker and content information separately with specially designed layers for the sequence-to-sequence framework. By applying instance normalization (Huang and Belongie, 2017), which is widely used for style transformation in computer vision, to encode the speaker information and then the adaptive instance normalization of the decoder takes the target speaker information to generate speech utterances of the target speaker. In (Cooper et al., 2020), a zero-shot multi-speaker end-to-end text-to-speech system is implemented based on speaker verification task. Their speaker embedding is trained from an end-to-end speaker recognition model and inputted it at pre-net and the self-attention of an end-to-end TTS framework generated best results in both speaker verification and speaker similarity tasks. These two methods have a common idea that only few speech data from the target speaker are required. Therefore, in this paper, we investigate the performance of both speaker embeddings and tried to further enhance the speech quality and speaker similarity.

For constructing our end-to-end multi-speaker TTS system, there are two main trends that can be chosen. One is autogressive model which pre-

dicts output speech frames based on previous results, such as Tacotron2 (Wang et al., 2017; Stanton et al., 2018). Tacotron2 is a sequence-to-sequence model that encodes the input text to a textual embedding and then decodes it to corresponding Mel-spectrogram using attention alignment. Another trend is non-autoregressive model like FastSpeech (Ren et al., 2020) and Transformer TTS (Li et al., 2019). This method is much faster than autoregressive method and alleviate generation errors derived from the attention mechanism. Besides, it could also adopt additional acoustic features, which is helpful for generating more stable and expressive speech output (Łańcucki, 2021).

In this paper, we adopted the Tacotron2 as our baseline system due to its various applications of TTS systems (Stanton et al., 2018). To alleviate the problem of autoregressive model, in the decoding process, the forward attention is used instead of original local sensitive attention in Tacotron2 and the Bahdanau attention layer is added to improve frame repetition or skipping problem, and also accelerate the training process. For enhancing the speech quality of output speech, we aimed to post-processing the output spectrogram while not delaying the entire training process significantly. The organization of the paper is listed as follows. In Section 2, we describe our multi-speaker Tacotron system modified by incorporating speaker embeddings and a post-filter enhancement. Section 3 shows experimental setups and the objective and subjective evaluations. Section 4 concludes with our findings and future work.

2 Proposed Method

In this section, the proposed multi-speaker speech synthesis system is introduced. The main idea of the proposed system is: the speaker embedding is adopted for learning the speaker-level information such as timbre and speaking style, and thus it should be able to separate the speaker-level information from the original speech utterance while the language-level information is not altered. Then, we applied two attention mechanisms to input speaker embeddings for Tacotron framework integration. Finally, a post-filter network (Takamichi et al., 2014; Kaneko et al., 2017) is adopted instead of conventional post-net for better speech enhancement.

2.1 Speaker Embedding and Encoding process

In the encoding process, the input character sequences are firstly encoded by 3 convolution layers and a Bidirectional LSTM as the original Tacotron encoder. However, to alleviate the sequential information that LSTM has to carry for long sequences, we added a self-attention layer as another encoding output, which is helpful for capturing the long-term contextual information and thus should let decoding process find the mapping faster. This is inspired by the findings of previous works (Cooper et al., 2020; Yasuda et al., 2019). The encoder process is illustrated in fig. 1, Note that the speaker embedding is concatenated with both output sequences.

For the speaker embedding, as described in Section 1, we chose two speaker embeddings to evaluate the effectiveness for the unseen speaker problem. First is the LDE (Cooper et al., 2020), which is trained for speaker verification task. The other is the VAE-based method (Chou et al., 2019), which is trained for voice conversion task. From their results, LDE is not only able to discriminate different speakers better than conventional x-vector (Snyder et al., 2018), but also useful for zero-shot multi-speaker TTS system. For VAE-based speaker embedding, it also outperformed the x-vector in the voice conversion task, so we are interested in comparing these two methods in the same experimental setup.

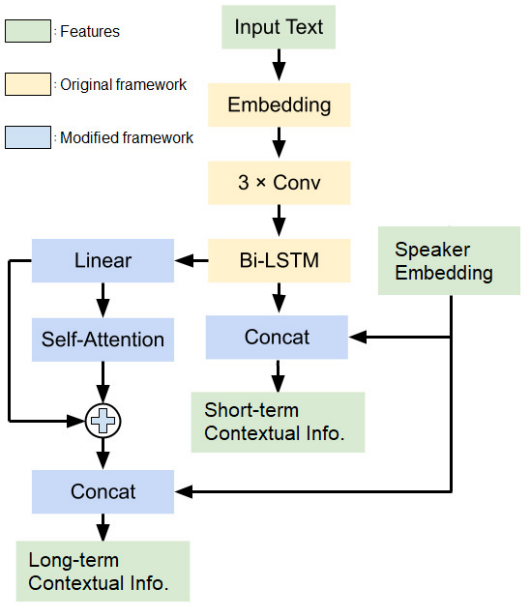

Figure 1: The modified framework for Tacotron encoder with speaker embedding.

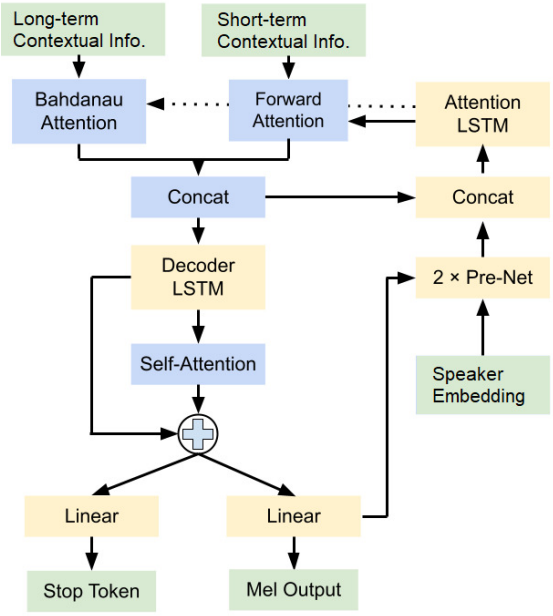

Figure 2: The modified framework for Tacotron decoder with speaker embedding.

2.2 Decoding process

In the decoding process, as shown in fig. 2 in order to let two outputs from the encoder could be combined together, we used forward attention for the short-term contextual information, which can align the sequence faster, and Bahdanau attention for long-term one since it is flexible for selecting suitable segments from the entire sequence. The speaker embedding is also used in the decoding process, which is also fed into pre-net as previous study suggested (Chien et al., 2021). Finally, a self-attention layer is added for alleviating the long-term dependencies to sequentially propagate information over long distances (Lin et al., 2017).

2.3 Post-net and Post-filter

Originally in the Tacotron framework, a post-net is consisting of 5 convolution layers to predict residuals for improving the reconstructed Mel-spectrogram. However, inspired from (Kaneko et al., 2017), we substituted a post-filter network for post-net to see if the speech quality could be further improved. Here, we adopted the Diffwave (Kong et al., 2020), which is a non-autoregressive model based on Markov chain to gradually convert a simple distribution (e.g., white noise) into complex distributions (e.g., waveform), and it served as the post-filter and for our system. The output Mel-spectrogram from decoder is served as the input for training DiffWave model. Finally, a original WaveNet vocoder is used for generating waveform.

3 Experiments

3.1 Data and experiment setup

In our implementation, we used AISHELL-3 (Shi et al., 2020) as the training corpus, which consists of 218 speakers. In the training process, we randomly select 173 speakers (which is about 75% of the corpus) and 100 utterances for each speaker as the training set. We down-sampled the sampling rate to 22,050 Hz and extract 80-dims Mel-spectrogram as the acoustic feature. The frame size and step size are 1,024 and 256 samples, respectively. Pitch range is set from 20 to 8,000 Hz.

For speaker embedding training, we followed the provided parameters in (Chou et al., 2019; Cooper et al., 2020) and set the speaker embedding size as 128. The speaker embedding for each speaker has been obtained by averaging the speaker embeddings calculated for all utterances of each speaker in the training set. For multi-speaker Tacotron TTS model training, both encoder outputs are set to 128-dims. In the decoding process, the dimension of the previous frame is increased from 80 to 256 before entering pre-net, therefore we also increased the dimension of the speaker embedding to 256 and use Softsign activation function before feeding into pre-net.

Finally, for the post-filter training, the decoded Mel-spectrogram is inputted into DiffWave model training process. We constructed a pseudo-parallel corpus that has the same textual information as the training set in order to let DiffWave learns the diffusion probabilistic model for generating speech utterance similar to original speech utterances. Table 1 shows the training parameters for each modules used in this study. For the speaker embedding training, LDE is denoted as $Spkr_{sv}$ and VAE is denoted as $Spkr_{vc}$, respectively.

In order to compare both models in the same situation, the batch size and total training step are set to equal. Tacotron baseline is the original Tacotron framework without self-attention while speaker embedding is inputted into the same place as described in Section 2. Tacotron + $Spkr_{vc/sv}$ is the proposed systems with different speaker embeddings.

These two systems are also used the same training setup to evaluate the effectiveness of self-attention. However, during model training, the Tacotron baseline, without self-attention, could not be trained properly, therefore is neglected for fur-

ther evaluations and also showed that self-attention is necessary when speaker embedding is trained jointly. Finally, the post-filter is trained based on the original paper, however, since the goal of the post-filter is to generate speech utterances as similar to the natural speech as possible. Therefore, we set it to 320k based on our preliminary evaluations, which is slightly smaller than the setup of the original paper.

Model	Batch size	Total step
$Spkr_{vc}$	32	1M
$Spkr_{sv}$	32	1M
Tacotron baseline	64	99k
Tacotron + $Spkr_{vc/sv}$	64	99k
Post-filter	16	320k

Table 1: Model training parameters

3.2 Evaluations

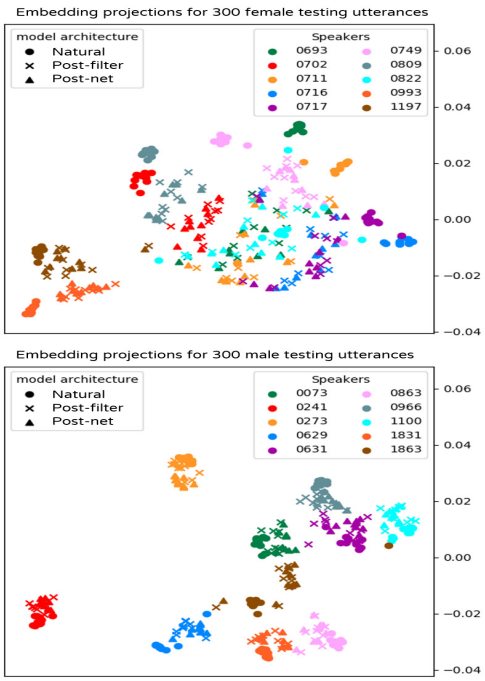

Figure 3: Visualization of speaker embeddings ($Spkr_{vc}$ is used here) that are generated w/ post-net and w/ post-filter. upper plot shows female distributions and lower plot shows male ones.

First evaluation is comparing the performance of two speaker embeddings by calculating the distortion between the generated speech and the synthesized speech, the Mel-cepstral distortion (MCD) is used. Here, we randomly select 5 unseen speakers from the testing set and used 10 utterances for each speaker. MCD is an objective measurement to

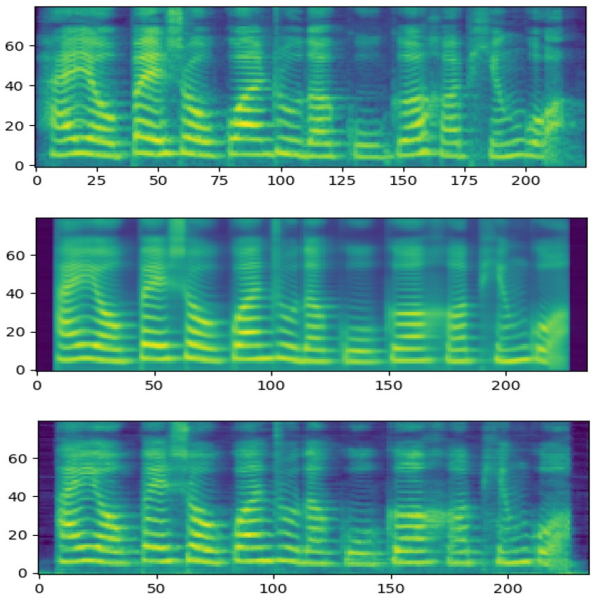

Figure 4: Original Mel-spectrogram (top figure), Mel-spectrogram generated from Tacotron + $Spkr_{vc}$ (middle figure), and generated from Tacotron + $Spkr_{vc}$ + Post-filter (bottom figure).

evaluate the acoustic feature distance between the generated speech sequences and the forced aligned natural speech sequences.

From the results from Table 2, the system using $Spkr_{vc}$ slightly outperformed the one using $Spkr_{sv}$. We also conduct subjective listening tests for evaluating speaker similarity and speech naturalness. Here, a 5-point mean opinion score (MOS) tests were held and 19 native Mandarin speakers were participated. The same set of 50 testing utterances are used here. For speaker similarity test, each participant is presented with several original sentences uttered by the target speaker and asked to score the speaker similarity of generated speech utterances. Table 3 shows that the system with $Spkr_{vc}$ generated speech utterances with better speaker similarity. Note that for comparing speech quality, only $Spkr_{vc}$ is used since it outperformed $Spkr_{sv}$ in speaker similarity test. These results imply that using the transfer learning of speaker verification task could neglect some useful information for generating natural speech while the VAE-based method, which is a generative model, is more suitable for TTS task.

For evaluating the post-filter, we used the Resemblyzer tool to visualize the output speaker embeddings of the testing 300 utterances from 10 unseen male and female speakers. In fig. 3, one can see that post-filter indeed generate similar embeddings

to the natural ones than the post-net. Besides, female speakers also have denser results compared to males. This could be the result of more female speakers in the AI-SHELL corpus (around 80%).

From these results, it could be a promising finding that incorporating speaker embedding trained from the voice conversion task is more helpful than the one from the speaker verification task for generating speech utterances of unseen speakers with only few speech data. Fig. 4 also shows an example of post-filter and post-net. One can see that the generated Mel-spectrogram is more detailed using post-filter. Demo wave files are showed in our page[1].

Embedding	Post-filter	MCD [db]
$Spkr_{vc}$	Yes	9.62±0.42
	No	10.16±0.53
$Spkr_{sv}$	Yes	9.29±0.85
	No	10.29±0.72

Table 2: Mel-cepstral distortion (with 95% confidence interval). Note that conventional post-net is used when post-filter is not.

Embedding	Post-filter	Similarity	Quality
$Spkr_{vc}$	Yes	3.75±0.71	3.70±0.5
	No	2.70±0.41	2.67±0.35
$Spkr_{sv}$	Yes	3.51±0.32	N/A
	No	2.31±0.32	

Table 3: The subjective MOS results (with 95% confidence interval).

4 Conclusion

In this study, we constructed a multi-speaker end-to-end Mandarin TTS system by integrating different speaker embeddings. We further enhanced speaker similarity and speech naturalness by using a post-filter network. This system is capable for generating speech utterances of unseen speakers using only few speech utterances. In our experiments, usually 5 utterances are enough. We also discovered that the speaker embedding trained from generative method is more suitable for TTS task comparing to the discriminative method. Future work will focus on augmenting additional features such as prosody or articulatory features for improving performance of the personalized TTS system.

[1] https://bit.ly/3uwCleM

References

Chung-Ming Chien, Jheng-Hao Lin, Chien-yu Huang, Po-chun Hsu, and Hung-yi Lee. 2021. Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech. In *ICASSP 2021*, pages 8588–8592. IEEE.

Ju-chieh Chou, Cheng-chieh Yeh, and Hung-yi Lee. 2019. One-shot voice conversion by separating speaker and content representations with instance normalization. *arXiv preprint arXiv:1904.05742*.

Erica Cooper, Cheng-I Lai, Yusuke Yasuda, Fuming Fang, Xin Wang, Nanxin Chen, and Junichi Yamagishi. 2020. Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6184–6188. IEEE.

Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510.

Ye Jia, Yu Zhang, Ron J Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, et al. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *arXiv preprint arXiv:1806.04558*.

Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo. 2018. Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 266–273. IEEE.

Takuhiro Kaneko, Hirokazu Kameoka, Nobukatsu Hojo, Yusuke Ijima, Kaoru Hiramatsu, and Kunio Kashino. 2017. Generative adversarial network-based postfilter for statistical parametric speech synthesis. In *ICASSP2017*, pages 4910–4914. IEEE.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2020. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*.

Adrian Łańcucki. 2021. Fastpitch: Parallel text-to-speech with pitch prediction. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6588–6592. IEEE.

Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. 2019. Neural speech synthesis with transformer network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6706–6713.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. 2019. Autovc: Zero-shot voice style transfer with only autoencoder loss. In *International Conference on Machine Learning*, pages 5210–5219. PMLR.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. 2020. Aishell-3: A multi-speaker mandarin tts corpus and the baselines. *arXiv preprint arXiv:2010.11567*.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333. IEEE.

Daisy Stanton, Yuxuan Wang, and RJ Skerry-Ryan. 2018. Predicting expressive speaking style from text in end-to-end speech synthesis. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 595–602. IEEE.

Shinnosuke Takamichi, Tomoki Toda, Graham Neubig, Sakriani Sakti, and Satoshi Nakamura. 2014. A postfilter to modify the modulation spectrum in hmm-based speech synthesis. In *ICASSP2014*, pages 290–294. IEEE.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

Yusuke Yasuda, Xin Wang, Shinji Takaki, and Junichi Yamagishi. 2019. Investigation of enhanced tacotron text-to-speech synthesis systems with self-attention for pitch accent language. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6905–6909. IEEE.