Exploring Effective Speech Representation via ASR for High-Quality End-to-End Multispeaker TTS

Dawei Liu[1], Longbiao Wang[1(✉)], Sheng Li[2(✉)], Haoyu Li[3], Chenchen Ding[2], Ju Zhang[4], and Jianwu Dang[1,5]

[1] Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin, China
{daveliu,longbiao_wang}@tju.edu.cn
[2] National Institute of Information and Communications Technology, Kyoto, Japan
sheng.li@nict.go.jp
[3] National Institute of Informatics (NII), Tokyo, Japan
[4] Huiyan Technology (Tianjin) Co., Ltd., Tianjin, China
[5] Japan Advanced Institute of Science and Technology, Ishikawa, Japan

Abstract. The quality of multispeaker text-to-speech (TTS) is composed of speech naturalness and speaker similarity. The current multispeaker TTS based on speaker embeddings extracted by speaker verification (SV) or speaker recognition (SR) models has made significant progress in speaker similarity of synthesized speech. SV/SR tasks build the speaker space based on the differences between speakers in the training set and thus extract speaker embeddings that can improve speaker similarity; however, they deteriorate the naturalness of synthetic speech since such embeddings lost speech dynamics to some extent. Unlike SV/SR-based systems, the automatic speech recognition (ASR) encoder outputs contain relatively complete speech information, such as speaker information, timbre, and prosody. Therefore, we propose an ASR-based synthesis framework to extract speech embeddings using an ASR encoder to improve multispeaker TTS quality, especially for speech naturalness. To enable the ASR system to learn the speaker characteristics better, we explicitly feed the speaker-id to the training label. The experimental results show that the speech embeddings extracted by the proposed method have good speaker characteristics and beneficial acoustic information for speech naturalness. The proposed method significantly improves the naturalness and similarity of multispeaker TTS.

Keywords: Speech synthesis · End-to-end model · Speech embedding · Speech recognition

1 Introduction

In recent years, end-to-end speech synthesis [1–3] has achieved significant progress. An increasing number of researchers have started to explore how to

T. Mantoro et al. (Eds.): ICONIP 2021, CCIS 1517, pp. 110–118, 2021.
https://doi.org/10.1007/978-3-030-92310-5_13

synthesize high-quality speech using a small amount of target speakers' speech, just minutes or even seconds. The ultimate goal of multispeaker text-to-speech (TTS) tasks is to solve the above problem.

The most straightforward approach for multispeaker TTS is to fine-tune the pretrained model directly using target speakers' data [4,5], but it is limited by the size of the target speakers' data. Another practical approach is to use speaker embedding. Previous studies have trained speaker embedding networks jointly with the end-to-end TTS model [4,6]. This means that speaker embedding networks and the TTS model are trained on the same datasets. However, speaker embedding networks and the TTS model have different requirements for datasets: the former requires a vast number of speakers in the training data, whereas the latter requires high-quality training data for each speaker. Therefore, some researchers have proposed training the speaker embedding networks separately [7,8], and then, they can be trained on more data regardless of speech quality. Speaker verification or speaker recognition (SV/SR) systems are currently widely used to extract the speaker embedding for multispeaker TTS [7–9]. [8] extracted the d-vector from the SV system as the speaker embedding, and the model can synthesize unseen target speakers' speech with only seconds of reference audio. [7] investigated two state-of-the-art speaker embeddings (x-vector and learnable dictionary encoding) to obtain high-quality synthesized speech. [9] used the traditional SR network as the speaker encoder to extract the embedding for cross-lingual multispeaker TTS.

Although the current multispeaker TTS [7–9] mentioned above has made remarkable progress, it still has substantial room for improvement. First, the objective of SV/SR tasks is to discriminate the speakers. The speaker embeddings extracted by the SV/SR model can improve the speaker similarity in general but ignore the dynamic properties of speech. The lack of the dynamic information related with the speaker might damage the quality of multispeaker TTS. Second, the speaker embedding extraction methods are borrowed from SV/SR tasks, making the development of multispeaker TTS depend on SV/SR tasks. More novel and practical approaches should be explored.

Unlike the above drawbacks of using an SV/SR system, the speech representations extracted from automatic speech recognition (ASR) include relatively more complete speech information. In this study, (1) we propose novel speaker embeddings extracted from a transformer-based ASR system to improve the multispeaker TTS quality instead of using an SV/SR system. (2) This ASR system is specially trained since the ASR task eliminated the speaker's characteristics in the network. To compensate for the speaker information loss, we explicitly added the speaker-id to the label in training so that the system would preserve the speaker's characteristics. Experiments show the proposed method can effectively improve the naturalness of synthesized speech without any loss of similarity compared with the conventional SV-based method.

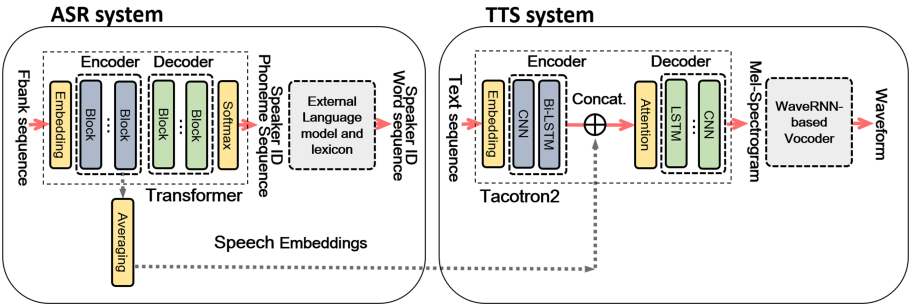

Fig. 1. Framework of our proposed model.

2 Exploring Effective Speech Representation via ASR for High-Quality End-to-End Multispeaker TTS

In this paper, we use ASR's encoder module to extract speech embeddings for multispeaker TTS. This approach avoids the problem encountered with speaker embeddings extracted by the SV/SR-based method that can only improve the speaker similarity but lack a positive influence on speech naturalness. Li et al. [10] used the TTS model to show that speaker information is relatively complete before the ASR decoder, but it decreases linearly as the layers of the encoder become deeper. This discovery indicates that although the ASR task eliminated the speaker characteristics, the ASR system still preserves them. In this paper, we explicitly feed the speaker-id to the label during transformer-based ASR model training so that we can more effectively preserve the speaker characteristics [11].

Although there have been existing works that integrate ASR and TTS models, our proposed method lies in none of these following categories. Recent studies have shown that jointly training ASR and TTS using cycle-consistency training [12] or autoencoders [13] can substantially improve ASR systems. [14] proposed a machine speech chain with semisupervised learning and improved ASR-TTS performance by training each other using only unpaired data. Tjandra et al.[15] used a SR model to let TTS synthesize unseen speakers' speech and further improve ASR.

The framework of the proposed method is shown in Fig. 1. The proposed system comprises two components: the transformer-based ASR model and the multispeaker TTS system based on Tacotron2 and the WaveRNN vocoder. We first train the Transformer-based ASR model, and then use it to extract speech embedding for multispeaker TTS. As described in Fig. 1, we extract 512-dimensional speech embeddings from the encoder of the transformer-based ASR model and concatenate these speech embedding to the outputs of the TTS encoder (512-dim); then, the augmented results (1024-dim) are input into the TTS attention module. The audio samples are available online[1].

[1] https://daveliuabc.github.io/multispeaker-demo/.

2.1 Transformer-Based End-to-End ASR Systems

We used the implementation of the transformer-based neural machine translation (NMT-Transformer) [16] in tensor2tensor[2] for all our experiments. The feature settings are the same as in our previous work [11].

We used 69 English phones[3] as the modeling unit. An external decoding process with lexicon and language models from LibriSpeech transcriptions is used to generate word-level recognition results. The speaker-id was explicitly added as the label during training [11,16]. We feed speaker-ids as the ground truth in training, and the combinations of speaker attributes (e.g., <SPK>) are inserted at the beginning of the label of the training utterances. The training labels are organized as "<SPK-1001> labels </S>". The network is trained to output them at the beginning of decoding automatically, so we do not have to prepare classifiers for these attributes.

2.2 Multispeaker TTS Systems

We based the end-to-end multispeaker TTS model architecture on Tacotron2 [3][4]. In multispeaker TTS, the input text sequences are converted to fixed-dimensional character embeddings, and then, the character embeddings pass through convolutional neural network (CNN) layers and the BLSTM layer to generate fixed-dimensional encoder outputs. We concatenate the embeddings extracted from the trained transformer-based ASR model with the fixed-dimensional output of the multispeaker TTS encoder and then input it to the location-sensitive attention module of the multispeaker TTS. The multispeaker TTS decoder can predict an 80-dimensional Mel-spectrogram. We used WaveRNN [17][5] as the multispeaker TTS vocoder, which converts the synthesized 80-dimensional Mel-spectrogram into time-domain waveforms.

3 Experimental Setup

3.1 Data Description

We trained the ASR model and the synthesizer of multispeaker TTS using 100 h of LibriSpeech [18] data (train-clean-100) and trained the vocoder using VCTK [19] datasets. All of them were trained separately. The LibriSpeech data (test-clean) were used to test the ASR model and multispeaker TTS.

[2] https://github.com/tensorflow/tensor2tensor.
[3] We train the phone-level ASR system to extract the phonetic posteriorgram (PPG) feature for TTS in the future.
[4] https://github.com/CorentinJ/Real-Time-Voice-Cloning.
[5] https://github.com/mkotha/WaveRNN.

3.2 ASR Models

Two ASR models required for embedding extraction were trained on the same LibriSpeech train-clean-100 but with different multitask training methods following [11,16]. These models and their performance on test-clean are as follows:

1. ASR_{ori}: trained using the original label. Rescored with an external trigram language model from all LibriSpeech transcripts, the WER% was approximately 9.0% after language model rescoring.
2. ASR_{spk}: trained using multitask training with the speaker-id and label. The WER% was approximately 9.0% after language model rescoring.

We randomly selected seven speakers from train-clean-100 and test-clean and randomly selected 30 voices for each speaker. We used ASR_{ori} and ASR_{spk} models to extract speech embeddings and used uniform manifold approximation and projection (UMAP) to visualize the extracted speech embeddings. The visualization results are shown in Fig. 2. Through visualization, it can be found that the proposed ASR_{spk} model can extract effective speaker information not only for seen speakers (train-clean-100) but also for the unseen speakers (test-clean). There were only 251 speakers in the training set (train-clean-100), so the speaker information contained in the unseen speaker's speech embeddings extracted using the ASR_{spk} model was encouraging.

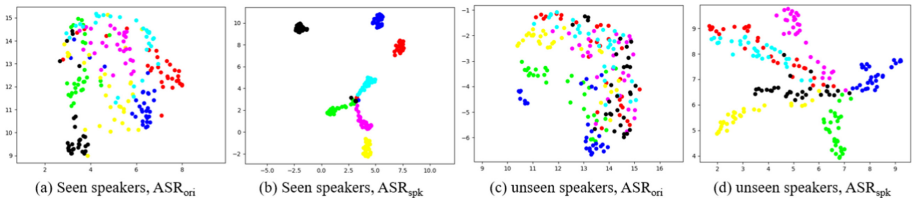

Fig. 2. Different E2E ASR models' speech embedding distributions by UMAP on selected data (seen from the training set, unseen from the testing set).

3.3 Multispeaker TTS System

We trained the synthesizer and vocoder separately, and used the same synthesizer and vocoder in all the experiments. We trained the synthesizer based on the original LibriSpeech train-clean-100 datasets and embeddings from the above two models (ASR_{ori} and ASR_{spk}).

We trained the vocoder based on VCTK datasets. We refer to the TTS systems according to the different ASR embedding sources: **ASR_{ori}-TTS** and **ASR_{spk}-TTS**. Simultaneously, we trained the synthesis model in [8](See Footnote 4) on the original LibriSpeech train-clean-100 datasets as a baseline that was referred to **Baseline(replica [8])**. The speaker encoder of the model maps a sequence of Mel-spectrograms to a d-vector and uses a generalized end-to-end SV loss [20,21].

3.4 Evaluation Metrics

The evaluations for the multispeaker TTS task comprise subjective and objective evaluations. The subjective evaluation metrics use the mean opinion score (MOS) for naturalness and the differential MOS (DMOS) [22] score for similarity. As metrics for the objective evaluation, acc% is used for SV, which is the ratio of the number of testing pairs identified as the same speaker over the total number of testing pairs, and the word recognition error rate (WER%) for the ASR task. All experiments were conducted on public datasets.

Subjective Evaluation. The same 25 listeners provided the MOS and DMOS scores. The listeners come from a professional team, and all of them have been learning English for more than ten years, while nineteen of them have majored in English. They used headphones for listening tests. For the naturalness evaluation, all listeners completed 56 audio tasks. Additionally, for the similarity evaluation, all listeners completed 100 pairs of audio tasks. The definitions of MOS and DMOS are as follows:

1. The MOS evaluates the naturalness of synthesized speech and reference audio from the target speakers with rating scores from 1 to 5, where 1 is the poorest result to understand, and 5 is the best result, with 0.5-point increments for each level.
2. The DMOS is used to evaluate the similarity between synthesized audio and reference audio subjectively: 1 (from a different speaker, sure), 2 (from a different speaker, not sure), 3 (from the same speaker, not sure), and 4 (from the same speaker, sure).

Objective Evaluation. The value of acc% from the ResCNN-based SV system [23], which was trained on VoxCeleb2 datasets, was used to evaluate similarity as the objective evaluation of the multispeaker TTS task. Every model provided 90 pairs of audio for testing seen speakers and 200 pairs for testing unseen speakers.

4 Experimental Results

4.1 Subjective Evaluation

The experimental results are listed in Table 1. The embedding extracted by ASR contained relatively complete speech information, such as speaker information and timbre. Therefore, compared with the baseline (replica [8]), all the proposed systems (ASR_{ori}-TTS and ASR_{spk}-TTS) achieved significant improvement in naturalness, whereas the similarity did not decrease (at least for ASR_{ori}-TTS). In Table 1, the randomly selected reference audio sometimes contains plenty of prosody and emotion, leading to a slightly higher MOS score for the unseen speaker than for the seen speaker. The work in [8] obtained similar experimental results on the same dataset. The naturalness' improvement of synthesized speech leads to a DMOS score of unseen speakers that is slightly higher than that of seen speakers. Table 1 shows that ASR_{ori}-TTS has slightly better speaker similarity

Table 1. Naturalness and similarity for multispeaker TTS. (95% confidence interval)

	Naturalness (MOS)		Similarity (DMOS)	
	Seen	Unseen	Seen	Unseen
Ground truth	4.53 ± 0.26	4.51 ± 0.23	3.38 ± 0.29	3.58 ± 0.84
Baseline (replica [8])	2.60 ± 0.57	3.11 ± 0.31	1.89 ± 0.24	2.01 ± 0.29
ASR_{ori}-TTS	3.12 ± 0.52	3.47 ± 0.27	1.90 ± 0.21	2.10 ± 0.23
ASR_{spk}-TTS	3.57 ± 0.25	3.82 ± 0.25	1.96 ± 0.22	1.93 ± 0.28

Table 2. SV Performance (acc%) as the objective evaluation of multispeaker TTS.

	Seen	Unseen
Ground truth	100%	100%
Baseline (replica [8])	57.78%	18.00%
ASR_{ori}-TTS	65.56%	26.00%
ASR_{spk}-TTS	91.11%	52.00%

than the baseline (replica [8]). For ASR_{spk}-TTS, the seen speakers achieve a higher DMOS score. This result proves the effectiveness of explicitly feeding the speaker-id as the label during training. The DMOS score on unseen speakers obtained by ASR_{spk}-TTS may reflect the shortcoming where we did not give speaker information expected by the system.

4.2 Objective Evaluation

The experimental results are listed in Table 2. The experimental results show that the proposed ASR_{spk}-TTS model achieved the best results, effectively surpassing the baseline model (replica [8]) for both seen and unseen speakers. The low scores for the baseline (replica [8]) were caused by the small number of speakers in the training datasets, which caused the speaker encoder network of the baseline [8] to fail to learn useful speaker embedding. Although the proposed ASR_{spk}-TTS model achieved good results, there was still a gap between seen and unseen speakers. The reason is that there are only 251 speakers in the training set, which may have caused a problem in the proposed method's construction of the speaker embedding space. Moreover, the slight drop in similarity scores in the subjective evaluation in Table 1 may have been caused by this.

4.3 Further Analysis

The proposed ASR_{ori} and ASR_{spk} models almost achieved the best and second-best performance (highlighted in gray and light gray, respectively) on both objective and subjective tasks compared with the baseline (replica [8]). In Subsect. 3.2, we also noticed that the recognition performance of these two models is almost identical. As pointed out in previous work [10], speaker information is

relatively complete before the ASR decoder, but the ASR task eliminated the speaker's characteristics in the network. To compensate for this fact, we explicitly added the speaker-id to the label in training so that the system would learn the speaker's characteristics. For this reason, the current task (extracting the speech embedding) can benefit from it. This is an interesting topic that merits an in-depth investigation in the future.

In real applications, TTS is integrated with ASR systems for complex tasks, such as speech-to-speech translation or dialogue systems, such as Amazon Alex, Microsoft Cortana, Apple Siri, and Google Translation. Inspired by this approach, the proposed method saves the development cost of training additional SV/SR systems on data containing many speakers.

The speech chain [24] and motor theory indicated that human speech production and perception functions evolve and develop together, sharing the same speech gestures in speech communication [25,26]. ASR and TTS are the inverse tasks of each other, and this paper reveals a close relation between ASR and TTS, which can help us design the next generation of speech applications. Human can recognize linguistic information meanwhile can preserve the speaker information, vice versa. In the current situation, however, either ASR or SV/SR cannot replicate this function. This is a topic worth investigating in the future.

5 Conclusion

This paper proposed a novel method to extract more effective speech representations from a transformer-based ASR model to improve the naturalness and similarity of multispeaker TTS. Compared with the traditional method, the proposed method does not rely on an individual SV/SR system. To enable the ASR system to learn more speaker characteristics, we explicitly added the speaker-id to the training label. Experiments showed that the proposed method almost achieved the best performance on both objective and subjective tasks. Because TTS is always integrated with ASR systems for complex tasks, such as a multispeaker speech chain, the proposed method reduces the development cost caused by integrating an additional SV/SR model.

Acknowledgment. This work was supported in part by the National Natural Science Foundation of China under Grant 61771333, NICT International Funding, and JSPS KAKENHI Grant No. 21K17837. We thank Prof. Zhenhua Ling of the University of Science and Technology of China for useful discussions.

References

1. Arik, S., et al.: Deep voice: real-time neural text-to-speech. In: Proceedings of ICML, pp. 264–273 (2017)
2. Ren, Y., et al.: Fastspeech: fast, robust and controllable text to speech. In: Advances in Neural Information Processing Systems (2019)
3. Shen, J., et al.: Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions. In: Proceedings of ICASSP, pp. 4779–4783 (2018)

4. Chen, Y., et al.: Sample efficient adaptive text-to-speech. In: Proceedings of ICLR (2019)
5. Kons, Z., et al.: High quality, lightweight and adaptable TTS using LPCNet. In: Proceedings of INTERSPEECH, pp. 176–180 (2019)
6. Nachmani, E., et al.: Fitting new speakers based on a short untranscribed sample. In: Proceedings of ICML, pp. 5932–5940 (2018)
7. Cooper, E., et al.: Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings. In: Proceedings of ICASSP, pp. 6184–6188 (2020)
8. Jia, Y., et al.: Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In: Advances in Neural Information Processing Systems, pp. 4480–4490 (2018)
9. Chen, M., et al.: Cross-lingual, multi-speaker text-to-speech synthesis using neural speaker embedding. In: Proceedings of INTERSPEECH, pp. 2105–2109 (2019)
10. Li, C., et al.: What does a network layer hear? Analyzing hidden representations of end-to-end ASR through speech synthesis. In: Proceedings of ICASSP, pp. 6434–6438 (2020)
11. Li, S., et al.: Improving transformer-based speech recognition systems with compressed structure and speech attributes augmentation. In: Proceedings of INTERSPEECH, pp. 1408–1412 (2019)
12. Hori, T., et al.: Cycle-consistency training for end-to-end speech recognition. In: Proceedings of ICASSP, pp. 6271–6275 (2019)
13. Karita, S., et al.: Semi-supervised end-to-end speech recognition using text-to-speech and autoencoders. In: Proceedings of ICASSP, pp. 6166–6170 (2019)
14. Tjandra, A., et al.: Listening while speaking: speech chain by deep learning. In: Proceedings of ASRU, pp. 301–308 (2017)
15. Tjandra, A., et al.: Machine speech chain with one-shot speaker adaptation. In: Proceedings of INTERSPEECH, pp. 887–891 (2018)
16. Vaswani, A., et al.: Attention is all you need. CoRR abs/1706.03762 (2017)
17. Kalchbrenner, N., et al.: Efficient neural audio synthesis. In: Proceedings of ICML, pp. 3775–3784 (2018)
18. Panayotov, V., et al.: Librispeech: an ASR corpus based on public domain audio books. In: Proceedings of ICASSP, pp. 5206–5210 (2015)
19. Yamagishi, J., et al.: CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92) (2019). https://doi.org/10.7488/ds/2645
20. Wan, L., et al.: Generalized end-to-end loss for speaker verification. In: Proceedings of ICASSP, pp. 4879–4883 (2018)
21. Paul, D., et al.: Speaker conditional WaveRNN: towards universal neural vocoder for unseen speaker and recording conditions. In: Proceedings of INTERSPEECH (2020)
22. Lorenzo-Trueba, J., et al.: The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods. In: Odyssey 2018 The Speaker and Language Recognition Workshop (2018)
23. Zhou, D., et al.: Dynamic margin softmax loss for speaker verification. In: Proceedings of INTERSPEECH (2020)
24. Denes, P., Pinson, E.: The Speech Chain, 2nd edn. Worth Publisher, New York (1993)
25. Kashino, M.: The motor theory of speech perception: its history, progress and perspective (Japanese). Acoust. Sci. Tech. **62**(5), 391–396 (2006)
26. Liberman, A., Mattingly, I.: The motor theory of speech perception revised. Cognition **21**, 1–36 (1985)