Shared model for multi-source speech generation tasks

Mingyang Zhang

National University of Singapore, Singapore

mingyang.zhang@u.nus.edu

1. Introduction and motivation

Many speech technologies contain speech generating stage, such as text-to-speech (TTS), voice conversion (VC), speech enhancement (SE). Recent advances in deep learning based methods significantly improve the performance of these technologies [1, 2, 3, 4, 5, 6, 7, 8].

So far, even though various successful deep learning based speech processing methods have been proposed, most of the systems can achieve only one task. For each problem, the network architecture is designed for the targeted task only and involves a long period of tuning specifically for the problem. This procedure needs to be repeated for different tasks, and this restrict the powerful effect of the neural network. The question is can we create a unified deep learning model to solve tasks cross multiple speech technologies.

We see that theoretical differences between these technologies are currently becoming much smaller than their original narrow definitions. To give a few examples, the recent advanced high-performance VC systems gain from the use of the phone posteriorgram (that is, a continuous phone representation) of inputted speech [9]. There was also an attempt to use both the spectrum features and phone posteriorgram to further improve the performance of voice conversion [4]. We can also see similar trends for TTS. The end-to-end TTS system sometimes also uses phone-embedding vectors as the input instead of letter inputs [3, 10]. There was also an attempt to use a reference audio signal as the additional input for Tacotron to transfer the prosody of the reference audio into synthetic speech via a reference encoder [11].

Given the above trends, we strongly believe that we can construct one model shared for multi-task. We assume that the speech generation related tasks can be divided into two parts: an input encoder and an acoustic decoder. The difference among the different tasks is the input. For example, the input of TTS is text characters while that of VC and SE is acoustic features. The model can be thought of as an encoder-decoder model that supports multiple encoders. The role of multiple encoder networks is the frond-end processing of each type of input data and the role of a decoder network is to predict acoustic features required for waveform generation. Our initial work starts with the joint training model for TTS&VC [12].

2. Joint training framework for TTS&VC

Inspired by the success of end-to-end TTS models, we adopt architectures similar to Tacotron for the encoders and decoder. More specifically, we have two encoders that encode different inputs and a shared decoder that predicts the acoustic features, followed by the generation of high-quality waveform signals based on WaveNet, a generative model for raw audio waveforms [13].The other contributions of our work are as follows. First, to achieve better TTS performance with a small amount of training data, we adapt a pre-trained TTS model to a target speaker. Second, for voice conversion, we train a many-to-one conversion to increase the size of training data while restricting the use of parallel data.

Our proposed multi-source Tacotron model is illustrated in Figure 1. It consists of a TTS input encoder, a VC input encoder, and a dual attention mechanism-based acoustic decoder followed by a WaveNet vocoder.

Both TTS input and VC input encoders have the same architecture, which includes a pre-net and a CBHG network. Each encoder transforms the corresponding input sequences into a fixed dimension state vector and a set of encoder output vectors.

The decoder of our model consists of a pre-net, an attention RNN layer, and a decoder RNN layer. Since a character embedding sequence and mel spectrogram have different time scales and we have to cope with the asynchronous input sequences, we use a dual attention mechanism. Two independent attention mechanisms Att^t and Att^v are used for transforming the outputs of the TTS and VC input encoders into context vectors, respectively.

Networks with multi-source inputs can often be dominated by one of the inputs [14]. In our proposed framework, since mapping from a source mel spectrogram to target mel spectrogram is much easier than mapping from a character embedding to a target mel spectrogram, the model will be dominated by the mel spectrogram input. To alleviate this problem, one of the following input types is randomly chosen during training: character embedding only, source mel spectrogram only, or both of the inputs to ensure that we only give specific input information to the decoder. To achieve this, we introduce a random masker for indicating which input to use during the training. Using the masker, we set the context vector that belongs to the unused input types to zero.

We jointly train the multi-source model with two inputs by using the dual attention mechanism. This mechanism allows the model to extract information from both character embedding and mel spectrogram inputs, even when one of them is absent, or the two of them are not time aligned. Given the different kinds of input to our proposed framework, we can choose which task should be achieved by setting the masker. If we use only the character embedding input, the system becomes a TTS model. If we use only the source mel spectrogram input, the system becomes a VC model. If we use both of the inputs, we can see this as a hybrid model of TTS and VC.

3. Results discussion

The results for speech quality and speaker similarity are shown in Figure 2 and Figure 3. We evaluated our model and compared it with the following systems.

- **TTS**: Stand-alone model of adapted TTS system
- **VC**: Stand-alone many-to-one VC model using same source speakers and target speaker

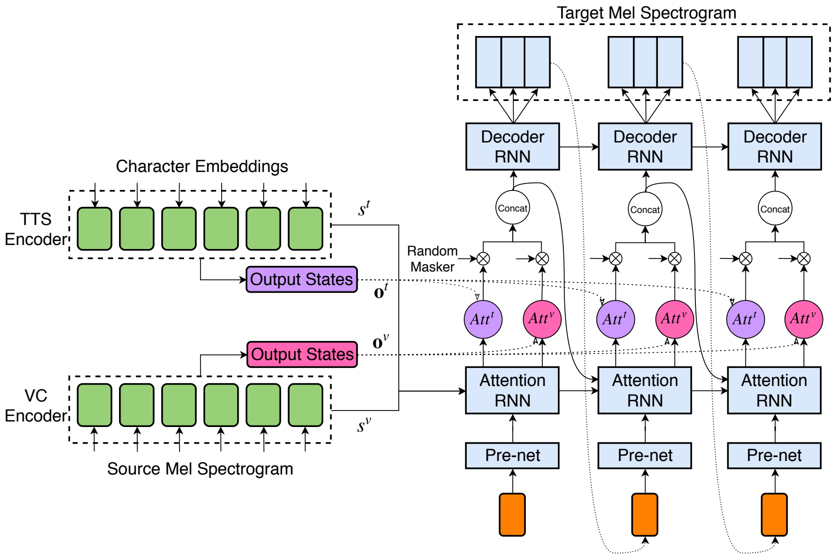

Figure 1: *Model architecture of our proposed multi-source sequence-to-sequence model for training TTS & VC simultaneously. Random maskers are applied to all decoder steps.*

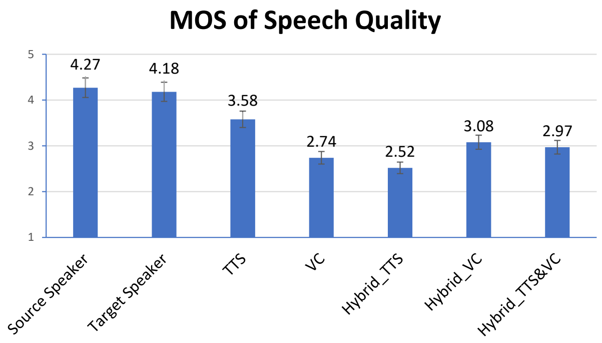

Figure 2: *MOS results with 95% confidence intervals for speech quality*

- **Hybrid TTS**: Proposed model with only text input
- **Hybrid VC**: Proposed model with only source speaker's speech input
- **Hybrid TTS & VC**: Proposed model with both text and source speaker's speech inputs

It was observed that our proposed model worked for both the TTS and VC tasks. We can see that the hybrid VC system outperformed the VC stand-alone system in terms of both speech quality and speaker similarity. This indicates that our proposed model improved the performance of VC. However, the MOS results for the hybrid TTS system were worse than those for the TTS stand-alone system. We can hypothesize several reasons for this. First, the current multi-source model might still be over-fitting to the VC task. Second, it might not have sufficient parameters for doing both the TTS and VC tasks. We may need to increase the number of parameters especially for the TTS task. Third, random selection may not be the best strategy for the maskers of the input encoders.

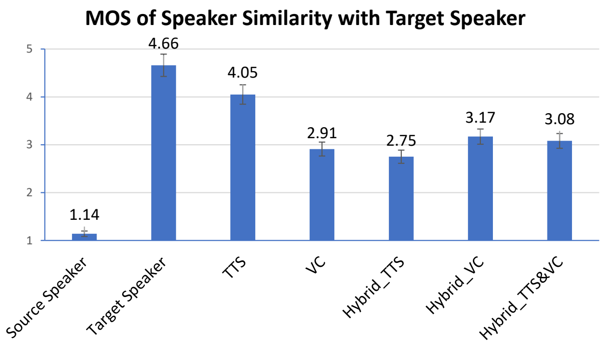

Figure 3: *MOS results with 95% confidence intervals for speaker similarity.*

4. Future works

From the experimental results we can see that even though our shared model can achieve both TTS&VC tasks, it still not sufficient for both tasks. A better seq2seq network architecture may need to be conducted, such as Tacotron 2 [15]. To alleviate the over-fitting problem, we need to investigate a better training strategy and scheduling of the maskers for the joint training stage, and a better attention mechanism of the decoder. We may also extend our work for other speech technologies, such as SE.

5. Acknowledgements

I express sincere gratitude to my advisors, Haizhou Li. I thank my internship mentor, Junichi Yamagishi. This work was partially supported by JST CREST Grant Number JPMJCR18A6 (VoicePersonae project), Japan and by MEXT KAKENHI Grant Numbers (16H06302, 17H04687, 18H04120, 18H04112, 18KT0051), Japan. This research is also supported by Ministry of Education, Singapore AcRF Tier 1 NUS Start-up Grant FY2016.

6. References

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[2] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio, "Char2wav: End-to-end speech synthesis," 2017.

[3] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: 2000-speaker neural text-to-speech," *Proc. ICLR*, 2018.

[4] J. Zhang, Z. Ling, L. Liu, Y. Jiang, and L. Dai, "Sequence-to-sequence acoustic modeling for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 3, pp. 631–644, March 2019.

[5] B. Sisman, M. Zhang, S. Sakti, H. Li, and S. Nakamura, "Adaptive wavenet vocoder for residual compensation in gan-based voice conversion," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, Dec 2018, pp. 282–289.

[6] Y. Zhou, X. Tian, H. Xu, R. K. Das, and H. Li, "Cross-lingual voice conversion with bilingual phonetic posteriorgram and average modeling," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 6790–6794.

[7] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5024–5028.

[8] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Monaural speech enhancement using deep neural networks by maximizing a short-time objective intelligibility measure," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5059–5063.

[9] L.-J. Liu, Z.-H. Ling, Y. Jiang, M. Zhou, and L.-R. Dai, "Wavenet vocoder with limited training data for voice conversion," in *Proc. Interspeech 2018*, 2018, pp. 1983–1987. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1190

[10] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhou, "Close to human quality tts with transformer," *arXiv preprint arXiv:1809.08895*, 2018.

[11] R. J. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," *CoRR*, vol. abs/1803.09047, 2018.

[12] M. Zhang, X. Wang, F. Fang, H. Li, and J. Yamagishi, "Joint training framework for text-to-speech and voice conversion using multi-source tacotron and wavenet," *arXiv preprint arXiv:1903.12389*, 2019.

[13] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[14] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1933–1941.

[15] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.