Joint training framework for text-to-speech and voice conversion using multi-source Tacotron and WaveNet

Mingyang Zhang, Xin Wang, Fuming Fang, Haizhou Li, Junichi Yamagishi

NUS
National University
of Singapore

NII
Inter-University Research Institute Corporation /
Research Organization of Information and Systems
National Institute of Informatics

Outlines

- Introduction
- Motivations & Proposal
- Tacotron
- Proposed joint training framework and training procedures:
  - Framework of the joint model
  - Training Procedures
- Experiments
- Conclusion

Introduction

❖ Text-to-speech

TTS is a technology that synthesizes natural-sounding human-like speech from text.

- Traditional approaches:

  waveform concatenation approach, statistical parametric approach

- Neural-network-based end-to-end approaches:

  Tacotron, Deep Voice, Char2wav, …

❖ Voice conversion

Voice conversion is a technology that modifies the speech of a source speaker and makes their speech sound like that of another target speaker without changing the linguistic information.

- Spectrum mapping:

  GMM, DNN, LSTM, VAE, GAN, …

Motivations

- Even though various successful methods have been proposed for TTS and voice conversion, most of the systems can achieve only one task.

- Both TTS and voice conversion can be benefited from each other.
  - Using phone posteriorgram to achieve high quality voice conversion[1].
  - Using reference audio signal to transfer the prosody of the reference audio into synthetic speech[2].

- Both TTS and voice conversion can be divided into two parts: an input encoder and an acoustic decoder.

[1] L. Sun, K. Li, H. Wang, S. Kang and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," 2016 IEEE International Conference on Multimedia and Expo (ICME), Seattle, WA, 2016, pp. 1-6.
[2] R. J. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stan- ton, 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.

Proposal

- Given the above motivations, we proposed to construct one model shared for both the TTS and voice conversion tasks.

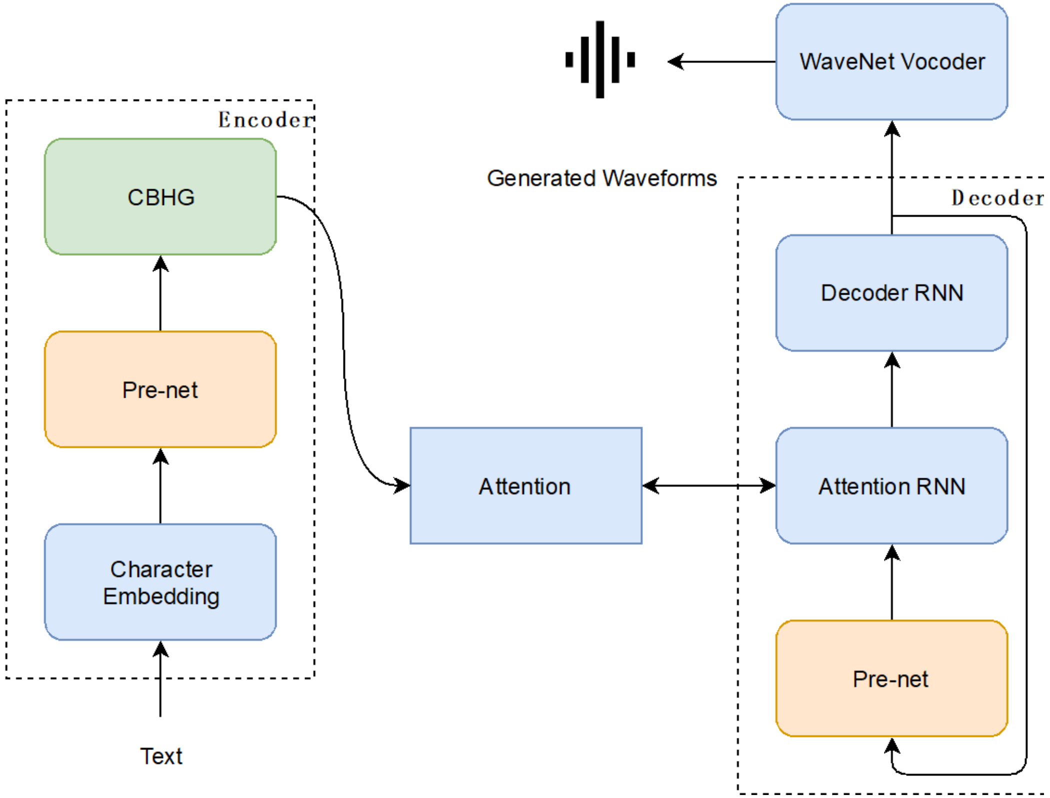
- The model can be thought of as an encoder-decoder model that supports multiple encoders.

- 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.

- Inspired by the success of end-to-end TTS models, we adopt architectures similar to Tacotron for the encoders and decoder.

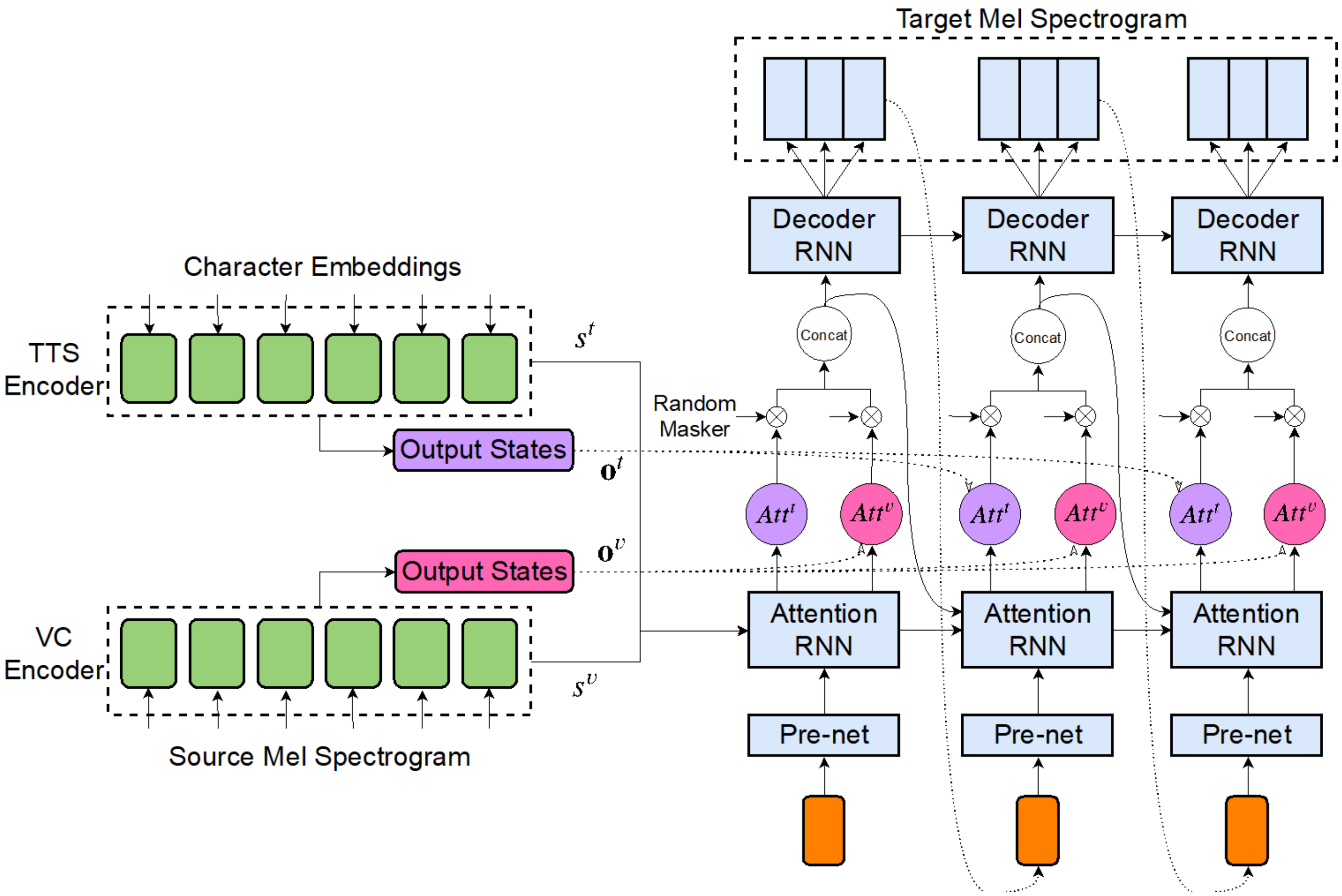
Tacotron

Tacotron[1] is an end-to-end text-to-speech (TTS) system that synthesizes speech directly from characters. The architecture of Tacotron is a sequence-to-sequence model with an attention mechanism .



[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.

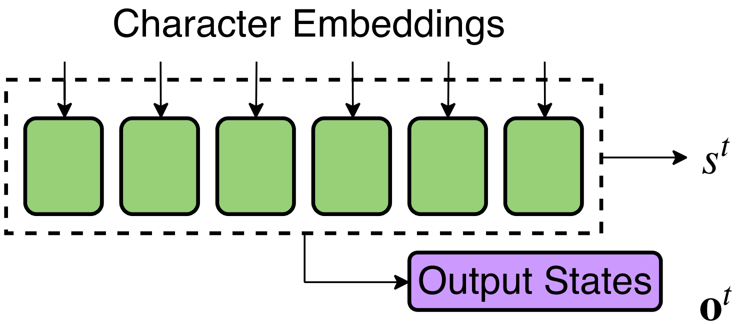
Proposed joint training framework



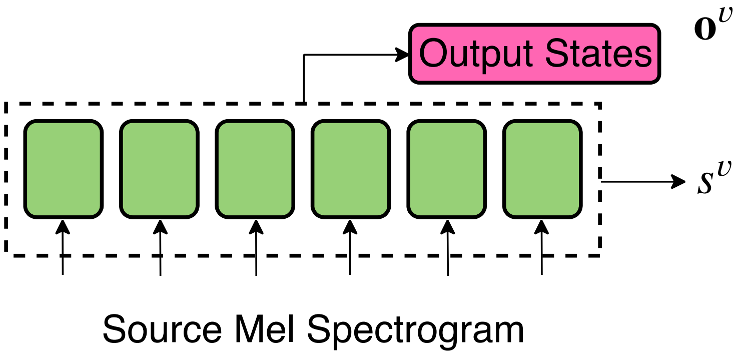
Proposed joint training framework

Two Encoders:

Encoders have the same architecture with Tacotron, which includes a pre-net and a CBHG network.



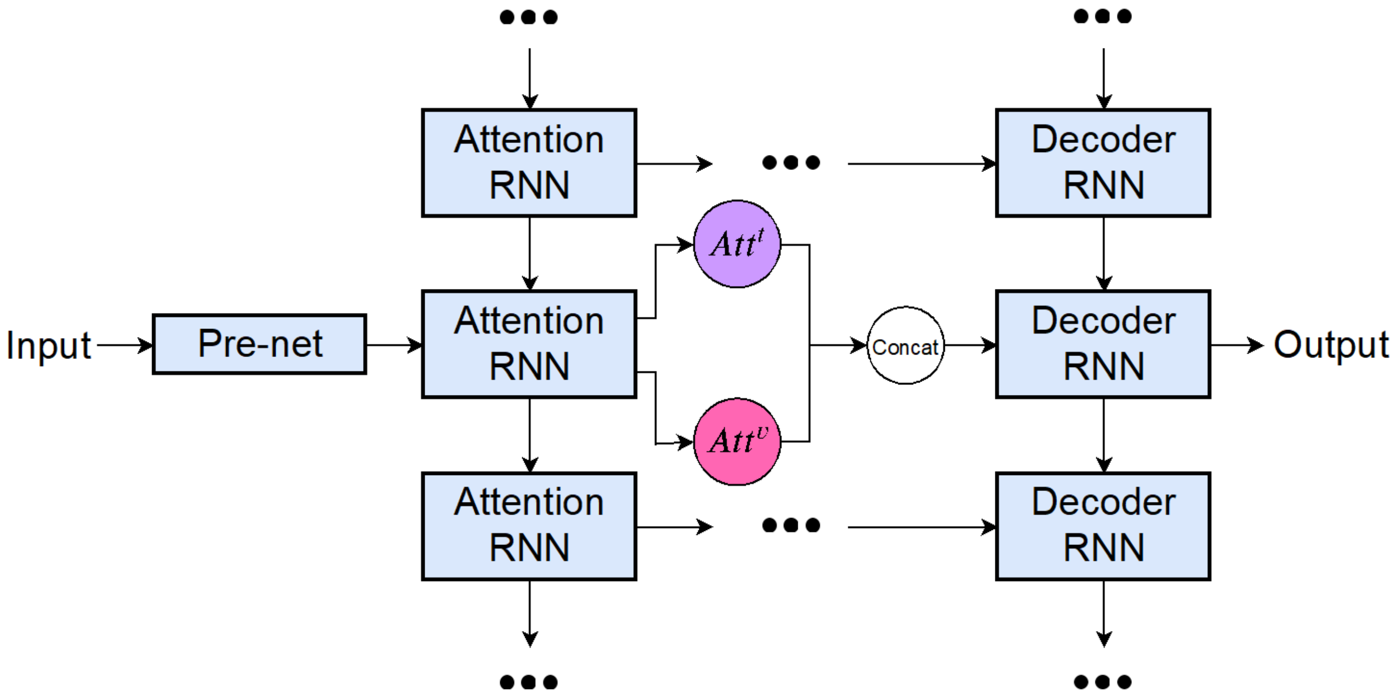
TTS encoder



VC encoder

Proposed joint training framework

Dual Attention-based Decoder

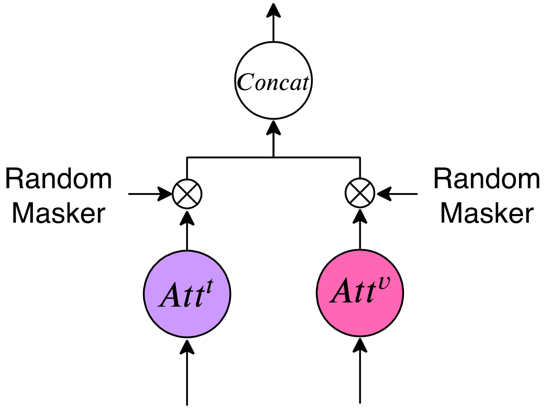


Proposed joint training framework

Random Selection of Input Encoders

- Networks with multi-source inputs can often be dominated by one of the inputs[1]. In our proposed framework, the model will be dominated by the mel spectrogram input.

- We introduce a random masker for indicating which input to use during the training. One of the following input types is randomly chosen during training: character embedding only, source mel spectrogram only, or both of the inputs.



[1] C. Feichtenhofer, A. Pinz, and A. Zisserman, “Convolutional two-stream network fusion for video action recognition,” in Pro- ceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1933–1941.

Training Procedures

Even with the random masker, if we train the model from the scratch, still only the mapping from spectrogram to spectrogram works. To prevent this from happening, we conduct the following step-by-step training:

1. We first train the two tasks' models separately and then use the encoder from each trained model to initialize the encoders of the multi-source model.
   - The TTS stand-alone model is adapted from a pre-trained TTS.
   - The VC stand-alone model is a many-to-one VC trained with parallel utterances from multiple source speakers and a target speaker.

2. Then, we jointly train the multi-source model with two inputs by using the dual attention mechanism and the random masker.

Experiment

Data Usage

- Pre-trained TTS: LJ speech database, approximately 24 hours
- Adaptation of TTS: 500 utts of a female speaker, SLT, from CMU ARCTIC
- Many-to-one VC: 500 parallel utts of two male speakers, BDL, RMS, as the source, same 500 utts of the female speaker, SLT, as the target.
- WaveNet Vocoder: speaker dependent WaveNet vocoder (SLT)

Experimental Setup

We evaluated our model and compared it with the following systems:

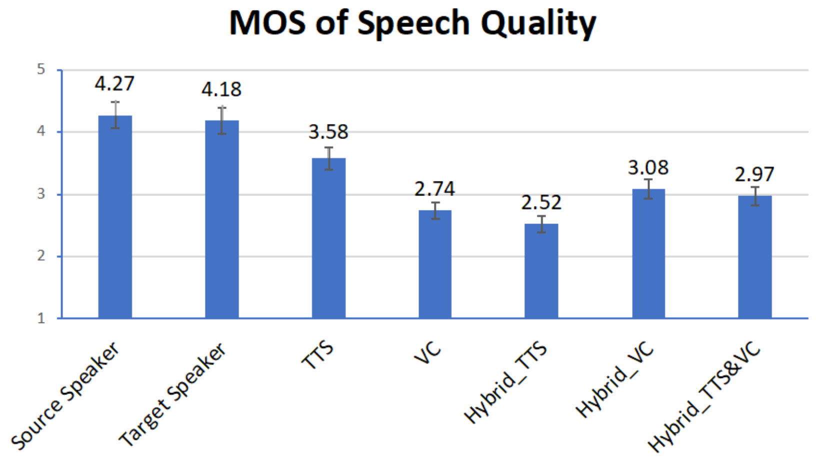
- TTS: Stand-alone model of adapted TTS system

  (Tacotron architecture, text characters as input)
- VC: Stand-alone many-to-one VC model

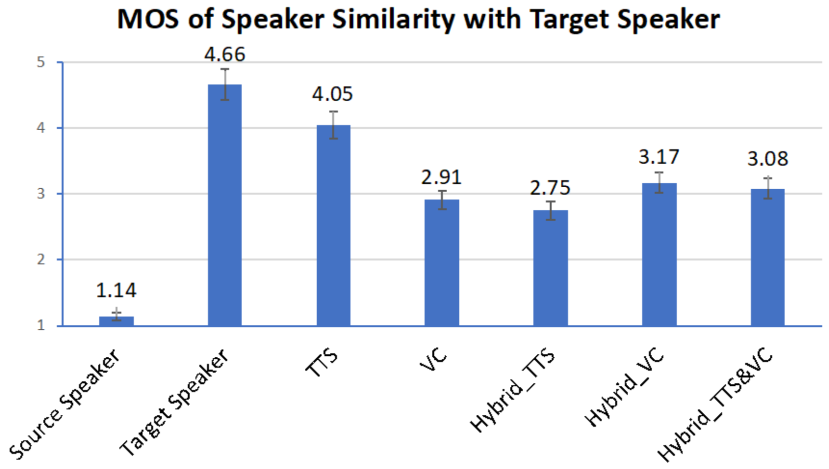
  (Tacotron architecture, Mel sprectrogram as input)
- Hybrid TTS: Proposed model with only text input
- Hybrid VC: Proposed model with only source speaker’s speech input
- Hybrid TTS&VC: Proposed model with both text and source speaker’s speech inputs

All the audio samples were generated by the same WaveNet vocoder.

Experiment



MOS results with 95% confidence intervals for speech quality of different models



MOS results with 95% confidence intervals for speech similarity of different models

In total, 97 evaluators participated in the test and produced a total of 18480 MOS scores. Accordingly, each speech sample received 10 quality and 10 similarity scores.

Experiments

Discussion:

- The current multi-source model might still be over-fitting to the VC task.

- 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.

- Random selection may not be the best strategy for the maskers of the input encoders. Better scheduling of the maskers needs to be investigated.

Experiment

Text: “Ah, it was sweet in my ears.”						
Source	Target	TTS stand alone	VC stand alone	Hybrid TTS	Hybrid VC	Hybrid TTS & VC
🔊	🔊	🔊	🔊	🔊	🔊	🔊

Audio Samples Page: https://nii-yamagishilab.github.io/hybrid-TTS-VC/

Conclusion

- We proposed a joint model for both the TTS and VC tasks.

- Given text characters as input, the model conducts end-to-end speech synthesis. Given the spectrogram of a source speaker, the model conducts sequence-to-sequence voice conversion.

- The experimental results showed that our proposed model achieved both TTS and VC tasks and improved the performance of VC compared with the stand-alone model.

- Our future work will be to investigate a better method for the maskers of the input encoders and a more appropriate training algorithm.

THANK YOU! ☺

Any Questions?