Tacotron: End-to-end high quality speech synthesis

YUXUAN WANG[1,a)]

Abstract: Text-to-speech synthesis system typically consists of multiple stages, such as a text analysis frontend, an acoustic model and an audio synthesis module. Building these components often requires extensive domain expertise and may contain brittle design choices. In this talk, I will describe recent advances on end-to-end neural speech synthesis modeling at Google.
I will start from introducing Tacotron, our first generation end-to-end model that synthesizes speech directly from characters. Given ¡text, audio¿ pairs, the model can be trained completely from scratch with random initialization. Tacotron greatly simplifies TTS pipeline and outperforms a production parametric system in terms of mean opinion score (MOS). To further improve audio quality, I will describe Tacotron 2, which combines Tacotron with a modified WaveNet model acting as a vocoder. Tacotron 2 achieves a MOS of 4.53 comparable to a MOS of 4.58 for professionally recorded speech. In addition to audio quality, prosodic modeling is also a core problem for speech synthesis. In the end, I will discuss style token, an unsupervised method for style modeling and control with end-to-end models like Tacotron.

Keywords: Speech synthesis, Deep neural network, ent-to-end learning, Wavenet

1. Introduction

Text-to-speech synthesis system typically consists of multiple stages, such as a text analysis frontend, an acoustic model and an audio synthesis module. Building these components often requires extensive domain expertise and may contain brittle design choices. In this talk, I will describe recent advances on end-to-end neural speech synthesis modeling at Google.

I will start from introducing Tacotron, our first generation end-to-end model that synthesizes speech directly from characters. Given ¡text, audio¿ pairs, the model can be trained completely from scratch with random initialization. Tacotron greatly simplifies TTS pipeline and outperforms a production parametric system in terms of mean opinion score (MOS). To further improve audio quality, I will describe Tacotron 2, which combines Tacotron with a modified WaveNet model acting as a vocoder. Tacotron 2 achieves a MOS of 4.53 comparable to a MOS of 4.58 for professionally recorded speech. In addition to audio quality, prosodic modeling is also a core problem for speech synthesis. In the end, I will discuss style token, an unsupervised method for style modeling and control with end-to-end models like Tacotron.

Fore more details, see [1], [2], [3].

References

[1] Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R. and Saurous, R. A.: Tacotron: Towards End-to-End Speech Synthesis, *Proc. Interspeech 2017*, pp. 4006–4010 (online), DOI: 10.21437/Interspeech.2017-1452 (2017).

[2] Wang, Y., Skerry-Ryan, R., Xiao, Y., Stanton, D., Shor, J., Battenberg, E., Clark, R. and Saurous, R. A.: Uncovering Latent Style Factors for Expressive Speech Synthesis, *NIPS Workshop on Machine Learning for Audio Signal Processing (ML4Audio)* (2017).

[3] Shen, J., Pang, R., Weiss, R. J., M.Schuster, Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R. J., Saurous, R. A., Agiomyrgiannakis, Y. and Wu, Y.: Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions, *submitted to ICASSP* (2018).

[1] Google Inc, USA
[a)] yxwang@google.com