Abstract

While recent neural sequence-to-sequence models have greatly improved the quality of speech synthesis, there has not been a system capable of fast training, fast inference and high-quality audio synthesis at the same time. In this thesis, we present a neural speech synthesis system capable of high-quality faster-than-real-time spectrogram synthesis, with low requirements on computational resources and fast training time. Our system consists of a teacher and a student network. The teacher model is used to extract alignment between the text to synthesize and the corresponding spectrogram. The student uses the alignments from the teacher model to synthesize mel-scale spectrograms from a phonemic representation of the input text efficiently. Both systems utilize simple convolutional layers. We train both systems on the english LJSpeech dataset. The quality of samples synthesized by our model was rated significantly higher than baseline models. Our model can be efficiently trained on a single GPU and can run in real time even on a CPU.