論文の内容の要旨

Thesis Summary

Title:

Improving quality and flexibility of deep neural network based text-to-speech synthesis

(深層ニューラルネットワーク型テキスト音声合成における自然性と柔軟性の改良)

氏名 : 趙　禕

Although deep neural networks have been applied to text-to-speech systems and shown promising future, it still suffers from the degradation of synthesized speech quality and distortion of speaker identity. In this thesis, we focus on improving quality and flexibility in neural network based text-to-speech system.

We firstly propose a multi-speaker speech synthesis framework that is built on bidirectional long short-term memory networks to synthesize high-quality speech with limited database. In this framework, a speaker independent model is trained combined with speaker identity vector. This speaker independent model can be viewed as a transformation from linguistic features to generalized vocoder parameters. Then speaker dependent models for transforming speaker independent vocoder parameters to those of the target speaker are followed. The proposed framework can also be applied to speaker adaptation.

Further, in multi-speaker speech synthesis, both speaker adaptive training and speaker adaptation techniques need to learn speaker identity precisely. Therefore, we conduct a series of comparative studies on different speaker representations for controlling speaker identity. We focus on speaker identity control at the input layer of our proposed framework, and investigate different speaker representations like i-vector and speaker code when they are used as augmented input vector. We also propose two approaches to estimate a new speaker's code. The first one is estimating a new speaker's code from i-vector using Gaussian Mixture Model. The other is extracting the speaker codes extracted from mel-spectrogram features using recurrent neural networks.

However, traditional training criteria such as mean squared loss can lead to obvious mismatches between generated and natural acoustic parameters. It definitely decreases the quality of generated speech. In addition, vocoders built on speech related prior assumptions give rise to distortion especially in phase reconstruction. To alleviate these problems, we propose frameworks that incorporate either a conditional generative adversarial network (GAN) or its variant, Wasserstein GAN with gradient penalty, into multi-speaker speech synthesis that uses the WaveNet vocoder. We also extend the GAN frameworks and use the discretized mixture logistic

loss of a well-trained WaveNet in addition to mean squared error and adversarial losses as parts of objective functions. Experimental results show that proposed framework using back-propagated discretized-mixture-of-logistics loss achieves the highest subjective evaluation scores in terms of both quality and speaker similarity.

Besides, linguistic features such as prosodic boundaries affects speech naturalness a lot. Instead of traditional cascaded prediction, we propose a unified framework for Chinese prosodic boundary prediction. This framework can be considered to be a multi-task learning process. We also explore various text features and representations to predict prosodic boundaries without task specific knowledge or sophisticated feature engineering. The synthesized speech naturalness can be improved by increasing the accuracy of prosodic boundary prediction.