conTTS: Text-to-Speech Application using a Continuous Vocoder

Mohammed Salah Al-Radhi, Tamás Gábor Csapó, Géza Németh, {malradhi,csapot,nemeth}@tmit.bme.hu

1. Introduction

- Text-to-Speech (TTS)
  - generates speech from the corresponding text
  - used in educational, telecommunication and multimedia applications
- Vocoder
  - category of speech codec that analyzes and synthesizes human voice
- End-to-end TTS
  - WaveNet, Tacotron, char2wav used to predict speech parameters directly from graphemes or phonemes
  - quality and naturalness of synthesized speech greatly improved, comparable with human recordings

2. Problem and Objective

- Problem
  - end-to-end systems require a large amount of speech data from one speaker to obtain good quality
  - slower performance
- Aim
  - achieve high-quality TTS
  - assist people with speech disorders
- Hypothesis
  - parametric vocoding is central to the success of state-of-the-art TTS systems

3. Methods

- Continuous vocoder
  - continuous F0 model to decrease the disturbing effect of creaky voice
  - no voiced/unvoiced decision
  - Kalman smoothing-based interpolation
  - MVF to model the voiced/unvoiced characteristics of sounds
  - Cheaptrick algorithm based on Mel-Generalized Cepstral analysis (MGC)

4. Objective evaluation

- Data: from CMU-ARCTIC
  - AIWB (Scottish English, male) and SLT (American English, female)
  - 25 sentences from each speaker were taken randomly to be synthesized.
- Error Metrics
  - Mel-cepustum distortion (MCD), root mean square error (RMSE), validation loss between valid and train sets, and correlation (CORR) measures the degree to which reference and generated data are close.

5. Perceptual evaluation

- Multi-Stimulus test with Hidden Reference and Anchor MUSHRA
  - 13 participants (6 males, 7 females) with engineering background
  - rate from 0 (highly unnatural) to 100 (highly natural)
  - samples: [https://malradhi.github.io/conTTS/](https://malradhi.github.io/conTTS/)

6. Discussion and Conclusion

- It is a text to speech application based on a speech analysis and synthesis system, that is lightweight and easy to use.
- It focused on the task of sequence modeling based on continuous vocoder.
- Experimental results demonstrated that the proposed RNN models can improve the naturalness of the speech synthesized significantly over the DNN model.
- Future work might be focused on voice conversion ensuring compatibility with a non-parallel dataset.