Budapest University of Technology and Economics,
Department of Telecommunications and Media Informatics, Budapest, Hungary

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
- **Vocoders**
  - 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)

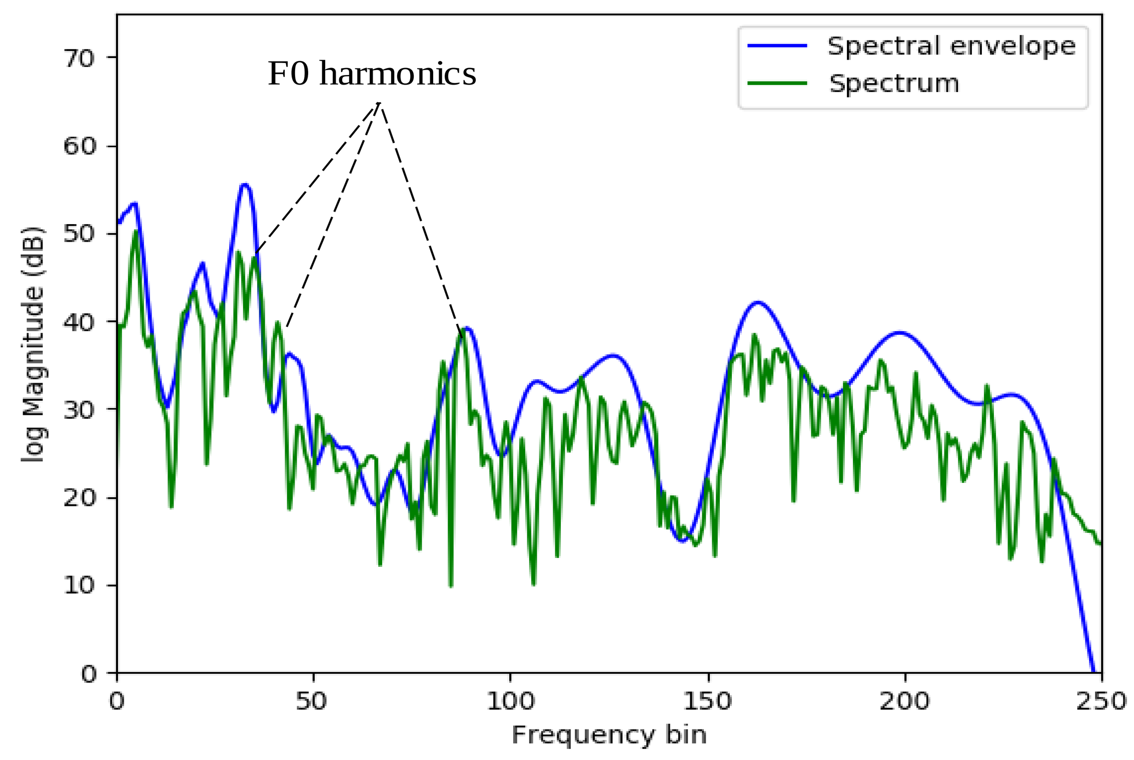

Figure 1. *Within a speech frame, the continuous vocoder measures the F0 and the MGC representation of the spectral envelope.*

- **Acoustic modelling**
  - we have used the feed-forward DNN and RNNs based statistical parametric speech synthesis
    - LSTM, BLSTM, and GRU topologies
  - 4 feed-forward hidden layers each consisting of 1024 units and performs a non-linear function
    - followed by a single RNN layer with 385 units used to train the continuous parameters
  - Festival toolkit is used to extract the linguistic features,
    - encode information regarding the phone identity, and syllable stress features

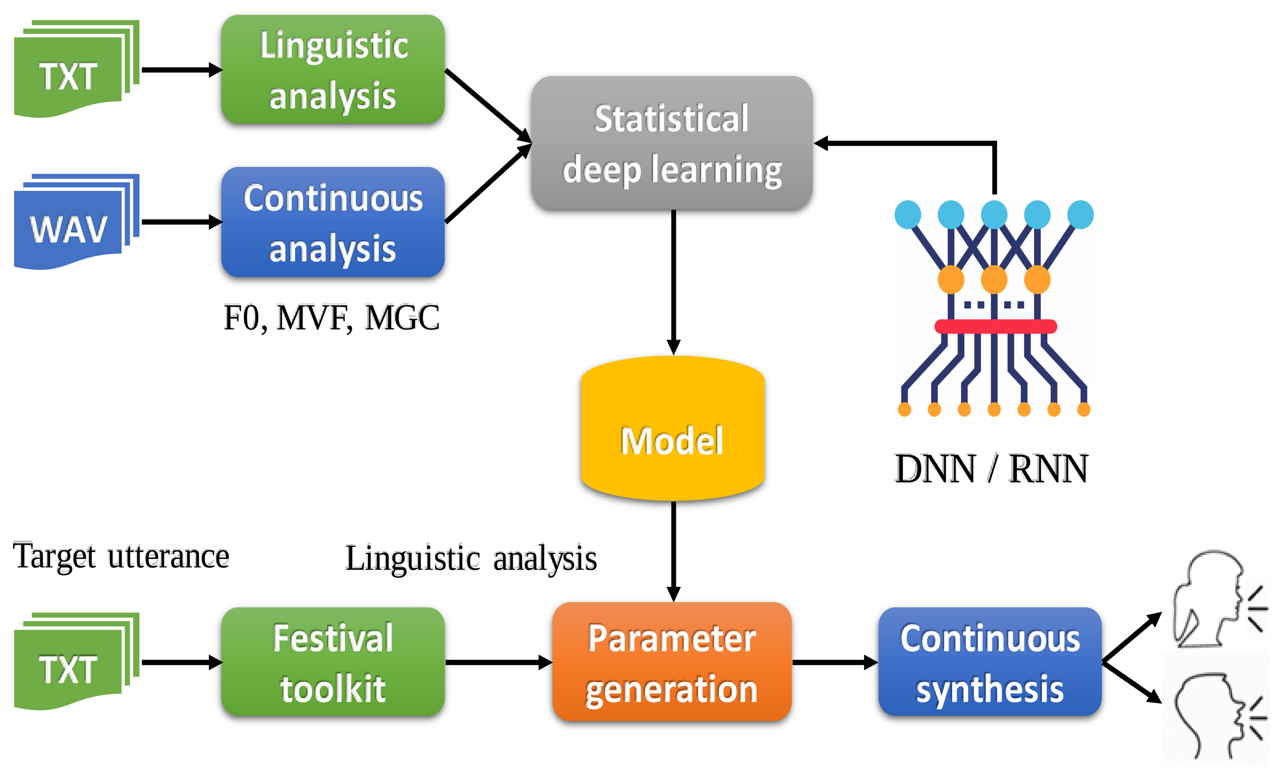

Figure 2. *Schematic overview of the conTTS system flow.*

https://github.com/malradhi/merlin

4. Objective evaluation

- Data: from CMU-ARCTIC
  - AWB (Scottish English, male) and SLT (American English, female)
  - 25 sentences from each speaker were taken randomly to be synthesized.
- Error Metrics
  - Mel-cepstrum 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.

Table 1. *Objective measures for all training systems based on synthesized speech signal using continuous vocoder. Lower value indicates better performance except for the CORR.*

Systems	MCD (dB)		$RMSE_{MVF}$ (Hz)		$RMSE_{F0}$ (Hz)		CORR		Validation error	
	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
DNN	4.923	4.592	0.044	0.046	17.569	22.792	0.727	0.803	1.543	1.652
LSTM	4.825	4.589	0.046	0.047	17.377	23.226	0.732	0.793	1.526	1.638
GRU	4.879	4.649	0.046	0.047	17.458	23.337	0.731	0.791	1.529	1.643
BLSTM	4.717	4.503	0.042	0.044	17.109	22.191	0.746	0.809	1.517	1.632

Table 2. *Parameters and excitation type of applied vocoders.*

Vocoder	Parameter per frame
Continuous	F0: 1 + MVF: 1 + MGC: 24
WORLD	F0: 1 + Band aperiodicity: 5 + MGC: 60
STRAIGHT	F0: 1 + Aperiodicity: 1024 + Spectrum: 1024

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/

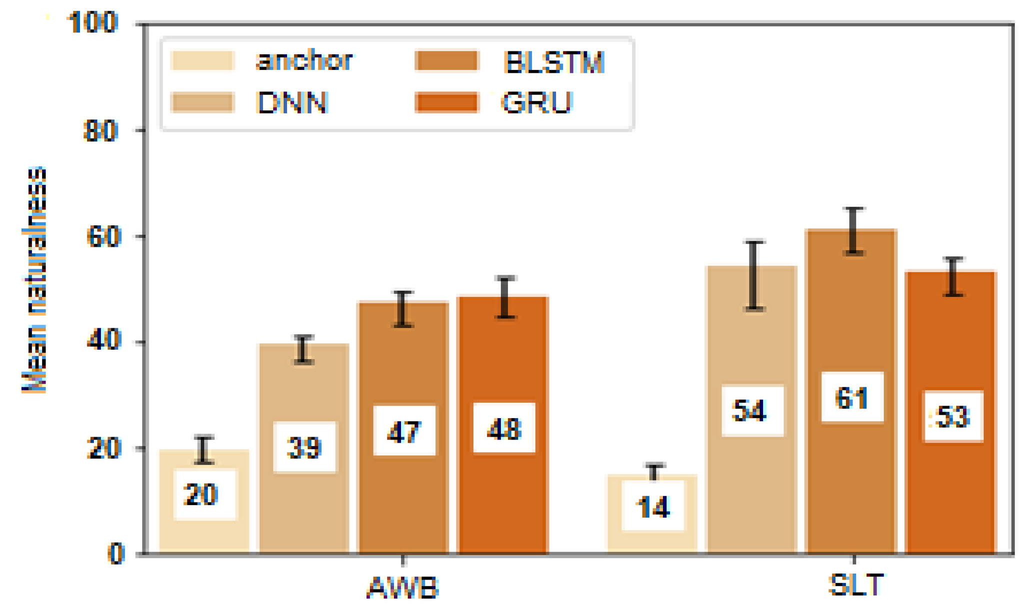

Figure 3. *Results of the MUSHRA listening test for the naturalness question. Error bars show the bootstrapped 95% confidence intervals. The score for the reference (natural speech) is not included.*

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.

Acknowledgements

The research was partly supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 825619 (AI4EU), and by the National Research Development and Innovation Office of Hungary (FK 124584 and PD 127915).