High Quality Continuous Vocoder in deep recurrent neural network based speech synthesis

Mohammed Salah Al-Radhi
malradhi@tmit.bme.hu

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

1. Introduction

- Speech synthesis is the computer generated human speech.
- **Key factors for quality degradation of speech synthesis**
  - Parametric vocoder (speech analysis & synthesis)
  - Acoustic modeling accuracy
  - Over-smoothing (sounds muffled)
- **Vocoder problems**
  - buzziness
  - real-time processing

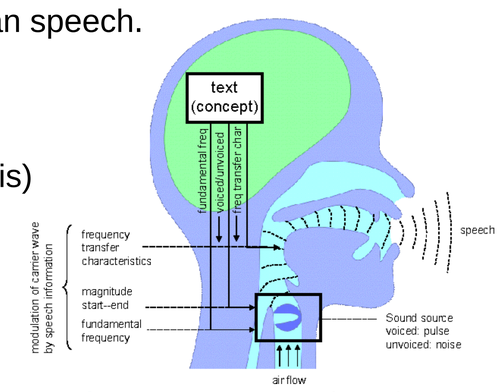

Figure 1. *Human speech synthesis.*

- **Hypothesis**
  - Continuous vocoder can be improved by introducing a better NLP solutions with low level synthesis.
- **Goal of this paper**
  - Build a deep learning for Text-to-Speech (TTS) synthesis using feedforward and recurrent neural networks as an alternative to hidden Markov models (HMMs) which often generate over-smoothing and muffled synthesized speech.
  - Develop a voice conversion system with simple model based neural network to convert the speech signal of a source speaker (e.g. male) into that of a target speaker (e.g. female).

2. Methods

- **Feed-Forward Deep Neural Network (FFD-NN)**
  - 6 feed-forward hidden lower layers of 1024 units each, performs
    - non-linear function of the previous layer's representation,
    - linear activation function at the output layer.
    - $y_k(x) = f\left(\sum_{j=0}^{M_i} W_{ij}x_j + b_i\right)$
  - minimize mean squared error function between target y and prediction output $\hat{y}$
    - $E = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
  - applied a hyperbolic tangent activation function
    - lower error rates and faster convergence
- **Recurrent Neural Network (RNN)**
  - 4 feed-forward hidden lower layers of 1024 units each, followed by a single top layer with 512 units as:
    - Long short-term memory (LSTM)
    - Bidirectional LSTM (B-LSTM)
    - Gated recurrent unit (GRU)
  - The iterative process of, for example, the Bi-LSTM can be defined as
    - $\vec{h}_t = f\left(W_{x\vec{h}}x_t + W_{\vec{h}\vec{h}}\vec{h}_{t-1} + b_{\vec{h}}\right)$
    - $\overleftarrow{h}_t = f\left(W_{x\overleftarrow{h}}x_t + W_{\overleftarrow{h}\overleftarrow{h}}\overleftarrow{h}_{t-1} + b_{\overleftarrow{h}}\right)$
    - $y_t = W_{\vec{h}y}\vec{h}_t + W_{\overleftarrow{h}y}\overleftarrow{h}_t + b_y$
- **Voice Conversion based DNN**
  - consists of feature processing, training and conversion-synthesis steps.
  - MVF, contF0, and MGC parameters are extracted from source and target voices
    - using the analysis function of the Continuous vocoder.
  - FF-DNN is applied to construct the conversion phase.
  - Dynamic Time Warping (DTW) algorithm is applied to map the training features
    - source speaker to the corresponding of the target speaker.

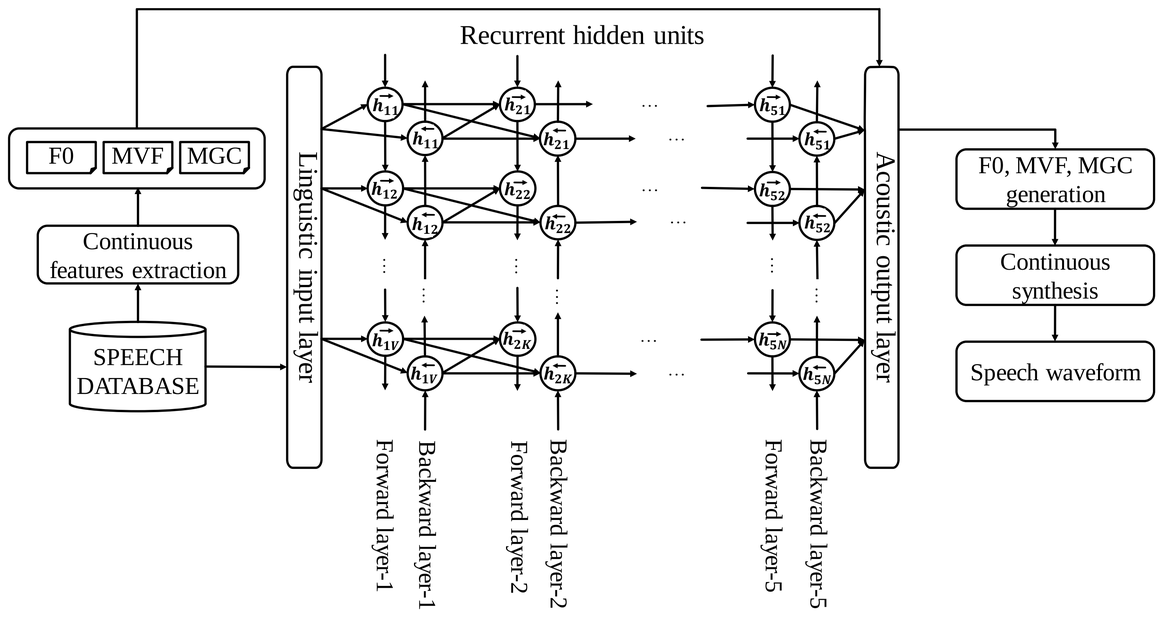

Figure 2. *Workflow of the DNN/RNN based TTS system.*

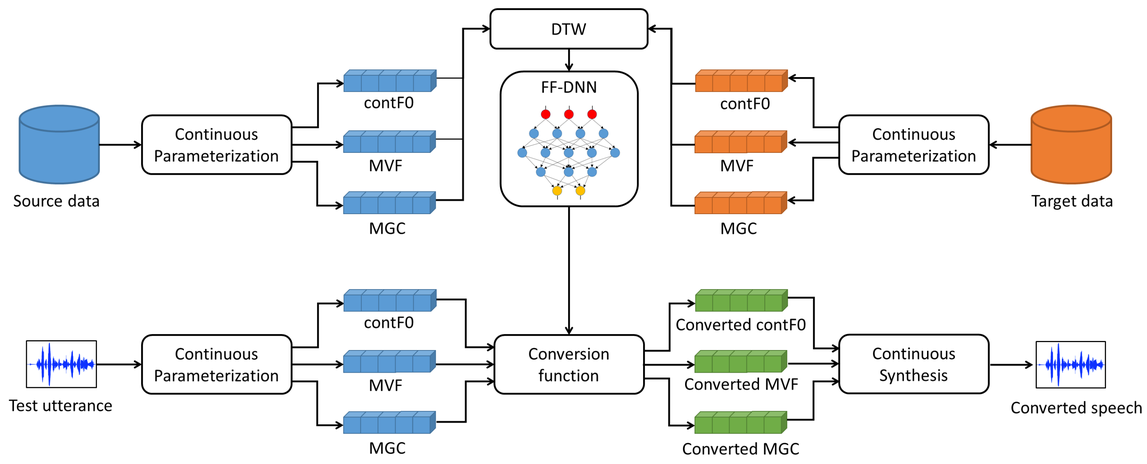

Figure 3. *Voice conversion process with Continuous based waveform generation.*

3. Objective evaluation

- Data: from CMU-ARCTIC
  - AWB (Scottish English, male), JMK (Canadian English, male), SLT (American English, female), and BDL (American English, male).
    - each one consisting of 1132 sentences.
    - 90% of these sentences were used in the training experiment, while the rest were used for testing and evaluating.
  - Training procedures were conducted on an NVidia Titan X GPU.
- Empirical measures
  - Mel-Cepstral Distortion
  - Root mean squared error
  - The correlation measures
  - frequency-weighted segmental SNR
  - Normalized Covariance Metric
  - Log Spectral Distortion
  - Weighted spectral slope

Table 2. *Objective measures for all training based VC systems.*

Error metrics	Model	SLT-to-BDL	BDL-to-SLT	SLT-to-JMK	JMK-to-SLT
MCD	Reference	5.624	5.355	5.856	5.765
	Proposed	**5.609**	**5.341**	**5.846**	**5.754**
fwSNRseg	Reference	1.660	1.119	**2.162**	0.558
	Proposed	**3.072**	**1.873**	1.970	**1.312**
LSD	Reference	2.423	2.208	2.506	2.557
	Proposed	**2.214**	**2.107**	**2.368**	**2.401**
WSS	Reference	8.842	16.299	8.068	17.310
	Proposed	**7.723**	**13.683**	**7.783**	**14.046**
NCM	Reference	0.103	0.102	0.024	0.030
	Proposed	**0.115**	**0.124**	**0.028**	**0.035**

Table 1. *Objective measures for all training based TTS systems.*

Systems	MCD (dB)		MVF (Hz)		F0 (Hz)		CORR	
	SLT	AWB	SLT	AWB	SLT	AWB	SLT	AWB
DNN(baseline)	4.923	4.592	0.044	0.046	17.569	22.792	0.727	0.803
LSTM	4.825	4.589	0.046	0.047	17.377	23.226	0.732	0.793
GRU	4.879	4.649	0.046	0.047	17.458	23.337	0.731	0.791
Bi-LSTM	**4.717**	**4.503**	**0.042**	**0.044**	**17.109**	**22.191**	**0.746**	**0.809**

Acknowledgements

The research was partly supported by the AI4EU project and by the National Research, Development and Innovation Office of Hungary (FK 124584). We would like to thank the listeners participating in the test.

4. Perceptual evaluation

- Multi-Stimulus test with Hidden Reference and Anchor (MUSHRA).
- 20 participants (mean age: 38 years) with engineering background.
- rate from 0 (highly unnatural) to 100 (highly natural).
- samples:
  - http://smartlab.tmit.bme.hu/vc2019
  - http://smartlab.tmit.bme.hu/vocoder2019

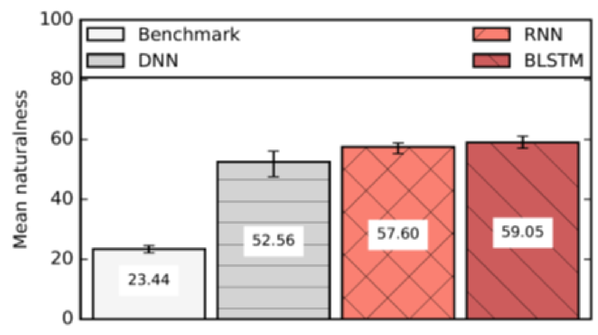

Figure 4. *Results of the subjective evaluation for the naturalness question. Higher value means larger naturalness.*

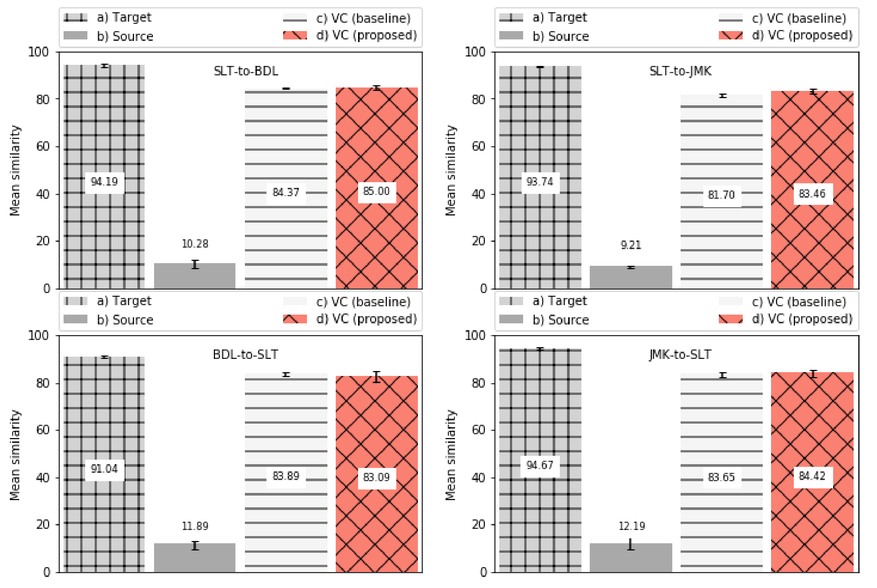

Figure 5. *MUSHRA scores for the similarity question. Higher value means larger similarity to the target speaker.*

Key references

[1] M.S. Al-Radhi, T.G. Csapó, G. Németh, "Time-Domain Envelope Modulating the Noise Component of Excitation in a Continuous Residual-Based Vocoder for Statistical Parametric Speech Synthesis", INTERSPEECH, pp. 434-438, 2017.
[2] P.N. Garner, M. Cernak, P. Motlicek, "A simple continuous pitch estimation algorithm," IEEE Signal Processing Letters, vol. 20, no. 1, pp. 102-105, 2013.
[3] T. Drugman, Y. Stylianou, "Maximum Voiced Frequency Estimation : Exploiting Amplitude and Phase Spectra," IEEE Signal Processing Letters, vol. 21, no. 10, pp. 1230–1234, 2014.
[4] M.S. Al-Radhi, T.G. Csapó, G. Németh, "Deep Recurrent Neural Networks in Speech Synthesis Using a Continuous Vocoder", SPECOM, pp. 282-291, 2017.