Deep Recurrent Neural Networks in Speech Synthesis Using a Continuous Vocoder

Budapest University
of Technology and Economics,
Department of Telecommunications and Media Informatics (BME TMIT),
Budapest, Hungary
Mohammed Salah Al-Radhi,
Tamás Gábor Csapó, Géza Németh,
{malradhi,csapot,nemeth}@tmit.bme.hu

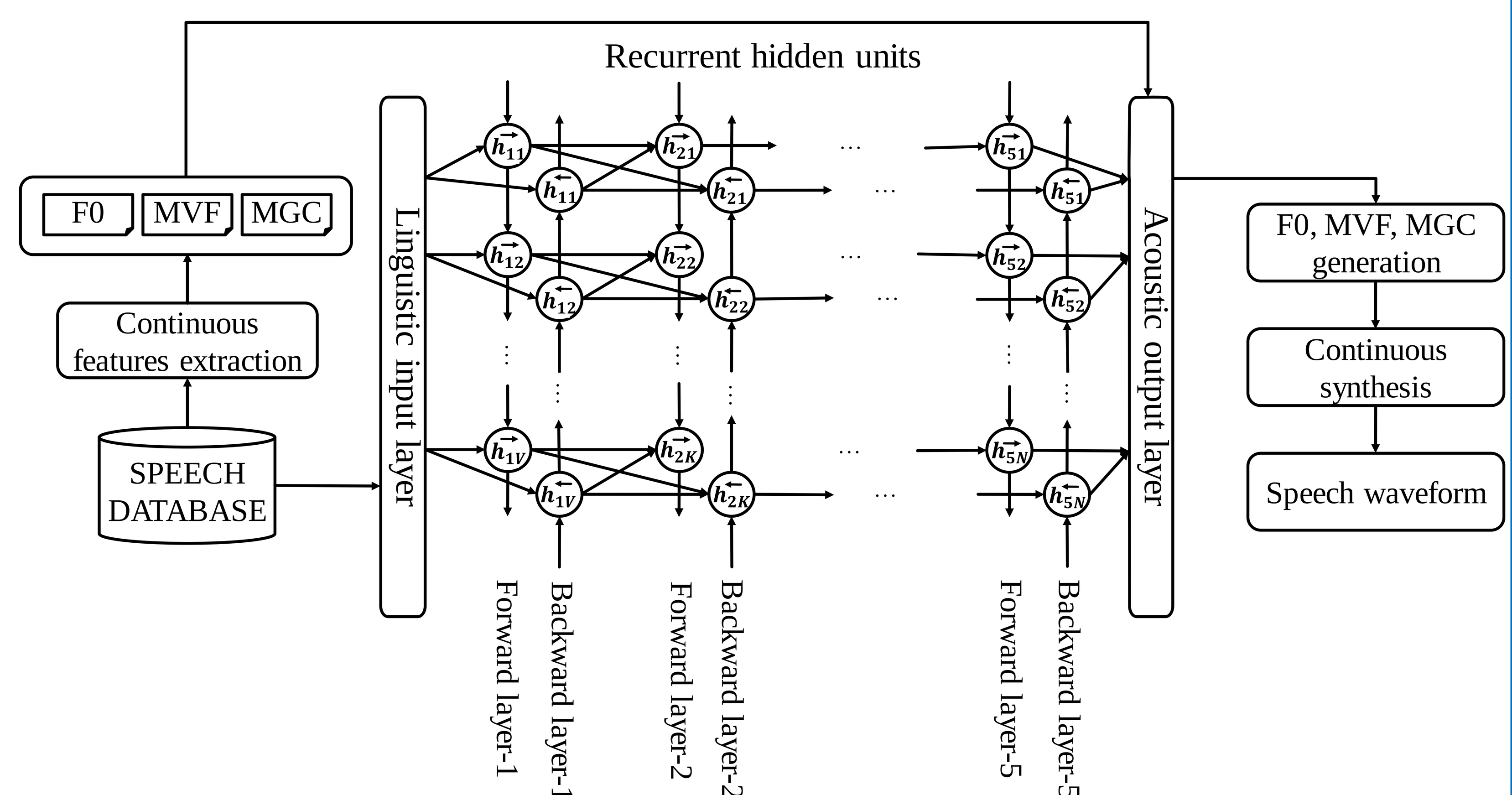
1. Introduction

- **vocoder problems**
  - buzziness
  - real-time processing
- **fundamental frequency (F0)**
  - continuous in voiced regions
  - discontinuous in unvoiced regions
  - hard to model boundaries between voiced and unvoiced segments
- **maximum voiced frequency (MVF)**
  - excitation parameter
  - separate the voiced and unvoiced components
- **standard Mel-Generalized Cepstral analysis (MGC)**
- **Feed-forward deep neural network**
  - in [1], we proposed a vocoder using continuous F0 in combination with MVF, which was successfully used with a feed-forward DNN based text-to-speech (TTS).
  - according to [2], DNNs have a lack of sequence modeling and ability to predict variances which might degrade the quality of synthesized speech
- **goal of this paper**
  - **Spectral envelope refinement**
  - **propose the use of sequence-to-sequence acoustic modeling with recurrent neural networks (RNNs).**
  - four RNN architectures are investigated and applied using this continuous vocoder to model F0, MVF, and proposed MGC

2. Methods

- **Continuous vocoder (baseline [1])**
  - continuous F0 model [3] to decrease the disturbing effect of creaky voice
    - standard autocorrelation
    - no voiced/unvoiced decision
    - Kalman smoothing-based interpolation
  - MVF to model the voiced/unvoiced characteristics of sounds [4]
- **Spectral envelope estimator**
  - CheapTrick algorithm [5]: accurate and temporally stable spectral envelope
    - F0-adaptive Hanning window
    - smoothing of the power spectrum
    - spectral recovery in the quefrency domain
- **Noise component**
  - shaping the high-frequency component by adding envelope modulated noise to the voiced excitation
  - True envelope [6]
    - the original spectrum signal and the current cepstral representation are maximized (see Fig. 2).
    - weighting factor makes the convergence more closely to the natural speech. In practice, the most successful weighting factor is 10 (see Fig. 3).
- **Acoustic modeling using RNN (see Fig. 1)**
  - applied a hyperbolic tangent activation function
    - lower error rates and faster convergence
  - 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)
    - Hybrid RNN

Figure 1. *Workflow of the proposed method.*



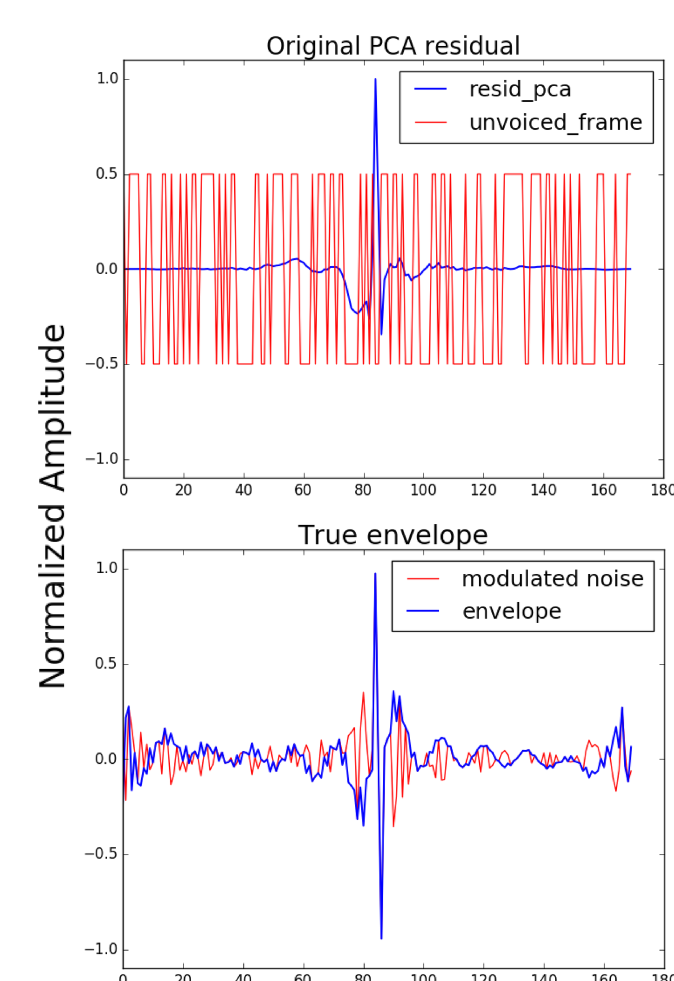


Figure 2. *Illustration the effect of applying the time envelope.*

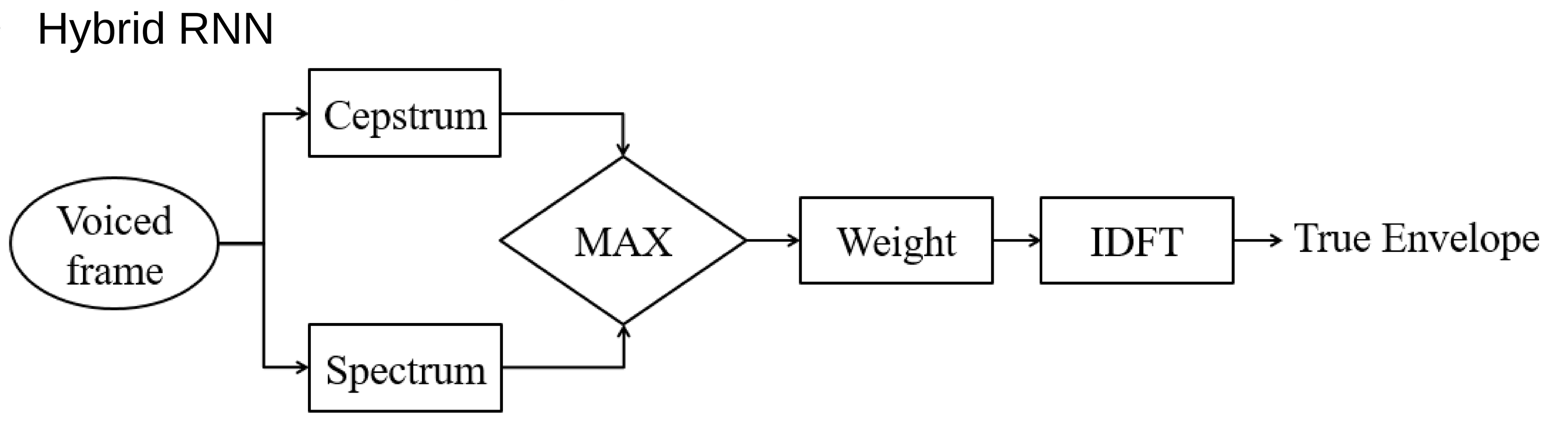


Figure 3. *Procedures for estimating the True envelope.*

3. Objective evaluation

- Data: from CMU-ARCTIC
  - AWB (Scottish English, male) and SLT (American English, female)
  - 90% of the sentences were used for training and the rest was used for testing
- RMS - Log Spectral Distance
  - root mean square (RMS) log spectral distance (LSD) evaluation was carried out
  - LSD is getting lower by using CheapTrick spectral algorithm than the simple spectral algorithm used in the baseline vocoder (see Fig. 4).
- Empirical measures (see Table 1)
  - Mel-Cepstral Distortion
  - Root mean squared error
  - Overall validation error
  - The correlation measures

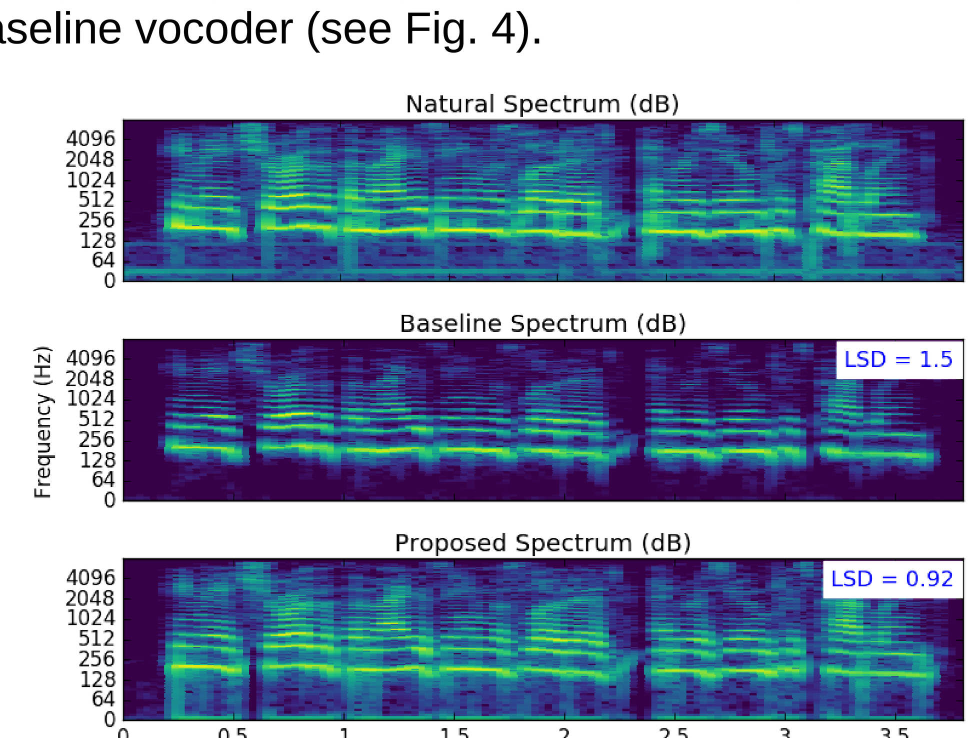


Figure 4. *Comparison of the speech spectrums synthesized by proposed continuous vocoder. The sentence is "He made sure that the magazine was loaded, and resumed his paddling." from speaker SLT.*

Systems	MCD (dB)		MVF (dB)		F0 (Hz)		CORR		Validation error	
	SLT	AWB	SLT	AWB	SLT	AWB	SLT	AWB	SLT	AWB
DNN (baseline)	4.923	4.592	0.027	0.028	17.569	22.792	0.727	0.803	1.543	1.652
LSTM	4.825	4.589	0.028	0.029	17.377	23.226	0.732	0.793	1.526	1.638
GRU	4.879	4.649	0.028	0.029	17.458	23.337	0.731	0.791	1.529	1.643
B-LSTM	**4.717**	**4.503**	**0.026**	**0.027**	**17.109**	**22.191**	**0.746**	**0.809**	**1.517**	1.632
Hybrid-RNN	5.064	4.516	0.028	0.027	18.232	22.522	0.704	0.805	1.547	**1.627**

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

4. Perceptual evaluation

- Multi-Stimulus test with Hidden Reference and Anchor (MUSHRA)
- 11 participants (mean age: 35 years) with engineering background
- rate from 0 (highly unnatural) to 100 (highly natural)
- both recurrent networks outperformed the DNN system (see Fig. 5)
- the BLSTM system reached the best naturalness scores

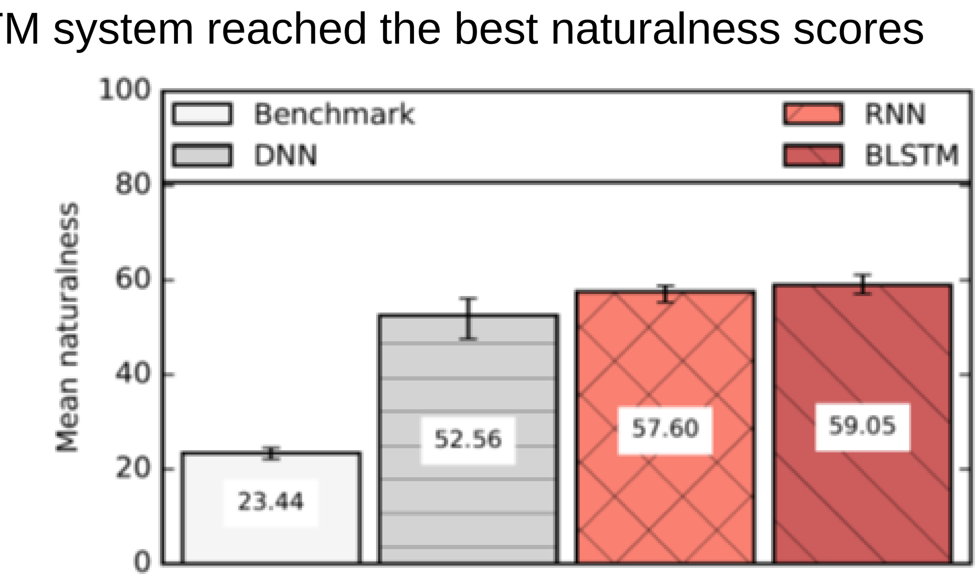


Figure 5. *Results of the MUSHRA listening test for the naturalness question. Error bars show the boot-strapped 95% confidence intervals.*

5. Discussion and Conclusion

- this work aims to apply a Continuous vocoder in recurrent neural network for more natural sounding speech synthesis
- it can be concluded that the BLSTM network converges faster and achieves better performance than others.
- plans of future research involve adding a Harmonics-to-Noise Ratio parameter to the analysis, statistical learning and synthesis steps in order to further reduce the buzziness caused by vocoding

Acknowledgements

The research was partly supported by the VUK (AAL-2014-1-183) and the EUREKA / DANSPLAT projects. We would like to thank the listeners participating in the test.

Key references

[1] T. G. Csapó, G. Németh, M. Cernak, and P. N. Garner, "Modeling Unvoiced Sounds In Statistical Parametric Speech Synthesis with a Continuous Vocoder," in EUSIPCO, Budapest, pp. 1338-1342, 2016.
[2] Zen H., and Senior A., "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," ICASSP, pp. 3844-3848, 2014.
[3] P. N. Garner, M. Cernak, and P. Motlicek, "A simple continuous pitch estimation algorithm," IEEE Signal Processing Letters, vol. 20, no. 1, pp. 102-105, 2013.
[4] T. Drugman and Y. Stylianou, "Maximum Voiced Frequency Estimation : Exploiting Amplitude and Phase Spectra," IEEE Signal Processing Letters, vol. 21, no. 10, pp. 1230–1234, 2014.
[5] M. Morise, "CheapTrick, a spectral envelope estimator for high-quality speech synthesis," Speech Communication, vol. 67, pp. 1-7, 2015.
[6] A. Röbel and X. Rodet, "Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation," in International Conference on Digital Audio Effects, Madrid, pp. 30-35, 2005.