High Quality Continuous Residual-based Vocoder for Statistical Parametric Speech Synthesis

Mohammed Salah Al-Radhi

Department of Telecommunication and Media Informatics
Budapest University of Technology and Economics, Budapest, Hungary
malradhi@tmit.bme.hu

1- Introduction

1.1. Motivation

Nowadays Text-to-speech systems are intelligible, however, a limitation of current parametric techniques does not allow full naturalness yet and there is a room for improvement in being close to human speech. According to recent summaries, there are three main factors in statistical parametric speech synthesis that are needed to deal with in order to achieve as high quality synthesized speech as unit selection: improved vocoder techniques, acoustic modeling accuracy and over-smoothing during parameter generation.

Vocoders received renewed attention recently as basic components in speech synthesis applications such as statistical parametric speech synthesis, voice transformation, and voice conversion. Despite the new WaveNet type of TTS to synthesize samples of speech directly without using vocoders as an intermediate step [1], it is still worth to develop advanced vocoders. There are vocoding methods which yield in close to natural synthesized speech (e.g. the STRAIGHT), but they are typically computationally expensive, and are thus not suitable for real-time implementation, especially in embedded environments. Therefore, there is a need for simple and computationally feasible digital signal processing algorithms that can handle non-regular voice qualities as well.

The motivation behind these was to construct a vocoder that would be a very flexible system whose parameters can be controlled with respect to achieve high quality synthesized speech.

1.2. Objectives

Given the context presented before, the main objective of my PhD work is to provide new advances in a novel continuous residual-based vocoder for statistical parametric speech synthesis. The first part of this challenge requires to address three principle objectives: 1) New technique for modelling unvoiced sounds is proposed by adding time domain envelope to the voiced excitation, 2) Investigate the Phase Distortion Deviation of vocoded samples as an objective test, and 3) Build a deep learning model based speech synthesis system based on the Merlin toolkit which could synthesize very high quality speech, and ensure that all continuous parameters used by the vocoder were taken through training. A MUSHRA type subjective listening test was also conducted comparing natural and vocoded speech samples.

The main goal presented here is to provide a large potential to improve the naturalness, expressivity, and suitability of Continuous vocoder based speech synthesis.

Keywords: Statistical parametric speech synthesis, continuous vocoder, time envelope, unvoiced source modelling, deep learning.

2- Related work

In the last decade, a large number of vocoders have been proposed [2] [3] [4] [5]. The direct antecedent of the current work had been carried out by [6] which proposed a computationally feasible residual-based vocoder. During the analysis phase, fundamental frequency (F0) is calculated on the input waveforms of a simple continuous pitch tracker [7]. In regions of creaky voice and in case of unvoiced sounds or silences, this pitch tracker interpolates F0 based on a linear dynamic system and Kalman smoothing. After this step, Maximum Voiced Frequency (MVF) is calculated from the speech signal [8], resulting in the MVF parameter. In the next step, 24-order Mel-Generalized Cepstral analysis (MGC) [9] is performed on the speech signal with alpha=0.42 and gamma=-1/3. In all steps, 5 ms frame shift is used. The results are the F0cont, MVF and the MGC parameter streams. Finally, the baseline system performs Principal Component Analysis (PCA) on the pitch synchronous residuals.

During the synthesis phase of the baseline system, voiced excitation is composed of PCA residuals overlap-added pitch synchronously, depending on the continuous F0. After that, this voiced excitation is lowpass filtered frame by frame at the frequency given by the MVF parameter. In the frequencies higher than the actual value of MVF, white noise is used. Voiced and unvoiced excitation is added together. Finally, an MGLSA filter is used to synthesize speech from the excitation and the MGC parameter stream [10].

3- Proposed systems

1.1. Modulating the noise component of excitation

Degottex et al. argue that the noise component is not accurately modeled in modern vocoders (even in the widely used STRAIGHT vocoder) [11]. In our baseline system [6], there is a lack of voiced component in higher frequencies. However, it was shown that in natural speech, the high-frequency noise component is time-aligned with the pitch periods [12]. The novelty of the proposed model differs from the baseline in the synthesis phase: we test various time envelopes to shape the high-frequency component (above MVF) of the excitation by estimating the envelope of the PCA residual and modifying the noise component by this envelope to make it more similar to the residual of natural speech.

Amplitude, Hilbert, Triangular and True envelopes are investigated, enhanced, and then applied to the noise component of the excitation in our continuous vocoder to present a more reliable envelope. I have also proposed in [13] [14] that the True envelope with weighting factor will bring a unique time envelope which makes the convergence more closely to the natural speech. In practice, the weight factor which was found to be the most successful is 10.

The natural and vocoded sentences were compared by measuring the Phase Distortion Deviation at a 5 ms frame shift using covarep/HMPD (http://covarep.github.io/covarep/). As can be seen from the Figure 3 of [13], the baseline vocoding sample has too much noise component compared to the natural sample (e.g. see the colors between 0.5-0.8s). On the other hand, the proposed systems with envelopes have PDD values (i.e., colors in the figure) closer to the natural speech. We also quantified the distribution of the PDD measure across all of the natural and vocoded variants of several sentences [13].

1.2. Acoustic modeling

In [6] and [13], a simple spectral model represented by 24-order mel-generalized cepstral coefficients was used [9]. However, more advanced spectral estimation methods might increase the quality of synthesized speech. A new spectral estimation method described in [15] based on the 60-order mel-generalized cepstral representation has been used in the following subsections, which can be regarded as a unified approach to speech spectral estimation.

1.2.1. Feed-Forward Deep Neural Network

The earlier vocoder [6] was successfully applied in hidden Markov models (HMM) based TTS. However, recent work in speech synthesis has pointed out the benefit of using deep neural network (DNN) models over HMM. With help of the Merlin toolkit [16], we used a DNN with 6 hidden layers, each consisting of 1024 units to train parameters of continuous vocoder (F0, MVF, and MGC) [17]. We applied a hyperbolic tangent activation function whose outputs lie in the range (-1 to 1) which can yield lower error rates and faster convergence than a logistic sigmoid function (0 to 1). In a subjective listening test, we found that the DNN-TTS using the Continuous vocoder was rated better than an earlier HMM-TTS system (Figure 1).

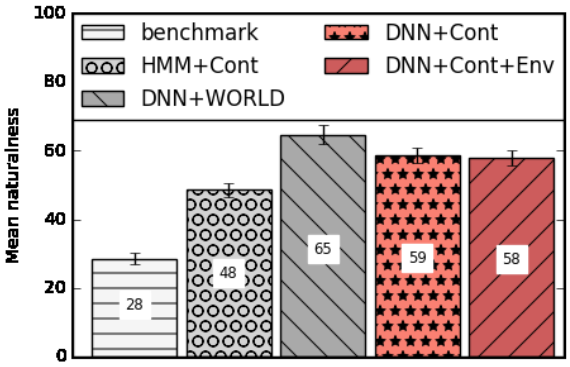

Figure 1. *Results of the MUSHRA listening test for the naturalness question. Error bars show the bootstrapped 95% confidence intervals. The score for the reference (natural speech) is not included.*

1.2.2. Recurrent neural network

In the previous section, we proposed a vocoder which was successfully used with a feed-forward deep neural network [17]. However, Zen, et al. [18] comprehensively listed the limitations of the conventional DNN-based acoustic modeling for speech synthesis, e.g. its lack of ability to predict variances, unimodal nature of its objective function, and the sequential nature of speech is ignored. In order to avoid these problems, we propose the use of sequence-to-sequence modeling with recurrent neural networks (RNNs). In [14], four neural network architectures (long short-term memory (LSTM), bidirectional LSTM (BLSTM), gated recurrent network (GRU), and standard RNN) are investigated and applied using this continuous vocoder to model F0, MVF, and MGC for more natural sounding speech synthesis.

From both objective and subjective evaluation metrics, experimental results demonstrated that the proposed RNN models can improve the naturalness of the speech synthesized significantly over the DNN baseline (Figure 2). In particular, the BLSTM network achieves better performance than others [14].

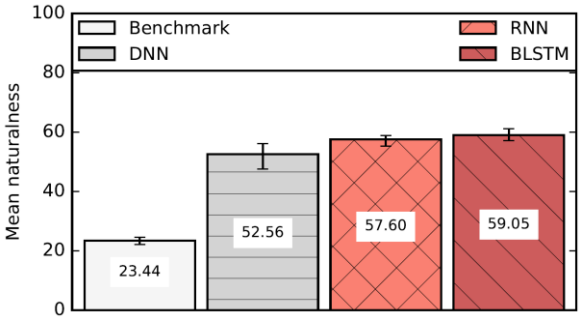

Figure *2. Results of the MUSHRA listening test for the naturalness question. Error bars show the bootstrapped 95% confidence intervals. The score for the reference (natural speech) is not included.*

4- Planned future work

Plans of future research involve some extensions to the current version of the Continuous vocoder. The author plans to add a Harmonics-to-Noise Ratio parameter to the analysis, statistical learning and synthesis steps in order to further reduce the buzziness caused by vocoding. In follow-up work, he also tries to investigate a sinusoidal-continuous vocoder based same continuous parameters and then implement an experimental comparison between them as better results can be expected. Besides, a mixture density recurrent network by combining with BLSTM-RNN based TTS can be also used as a basis of future developments to further improve and refine continuous parameters.

Acknowledgements

This author would like to thank his supervisors Prof. Géza Németh and Dr. Tamás Gábor Csapó for their supports and encouragements. He would also like to thank the Stipendium Hungaricum for providing him a fully funded doctoral program.

References

[1] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," in *arXiv preprint, https://arxiv.org/abs/1609.03499*, 2016.

[2] T. Drugman and T. Dutoit, "The deterministic plus stochastic model of the residual signal and its applications," *IEEE Trans. Audio, Speech and Language Processing*, vol. 20, no. 3, pp. 968-981, 2012.

[3] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Trans. on Audio, Speech, and Language*, vol. 19, no. 1, pp. 153-165, 2011.

[4] G. Degottex and Y. Stylianou, "A Full-Band Adaptive Harmonic Representation of Speech," in *Proc. Interspeech*, Portland, USA, 2012.

[5] D. Erro, I. Sainz, E. Navas, I. Hernaez, "Improved HNM-based Vocoder for Statistical Synthesizers," in *Proc. Interspeech*, Florence, 2011.

[6] Tamás Gábor Csapó, Géza Németh, Milos Cernak, and Philip N. Garner, "Modeling Unvoiced Sounds In Statistical Parametric Speech Synthesis with a Continuous Vocoder," in *EUSIPCO*, Budapest, 2016, pp. 1338-1342.

[7] 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.

[8] T. Drugman and Y. Stylianou, "Maximum Voiced Frequency Estimation : Exploiting Amplitude and Phase Spectra," *IEEE Signal Processing Letters*, vol. 21, no. 10, p. pp. 1230–1234, 2014.

[9] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis - a unified approach to speech spectral estimation," *in Proc. ICSLP*, p. 1043–1046, 1994.

[10] S. Imai, K. Sumita, and C. Furuichi, "Mel Log Spectrum Approximation (MLSA) filter for speech synthesis," *Electronics and Communications in Japan (Part I: Communications)* , vol. 66, no. 2, pp. 10-18, 1983.

[11] G. Degottex, P. Lanchantin, and M. Gales, "A Pulse Model in Log-domain for a Uniform Synthesizer," *in Proc. ISCA SSW9*, p. 230–236, 2016.

[12] Y. Stylianou, "Applying the harmonic plus noise model in concatenative Speech Synthesis," *IEEE Trans. Audio, Speech, Lang. Proc*, vol. 9, no. 1, pp. 21-29, 2001.

[13] Mohammed Salah Al-Radhi, Tamás Gábor Csapó, and Géza Németh, "Time-domain envelope modulating the noise component of excitation in a continuous residual-based vocoder for statistical parametric speech synthesis," in *Interspeech (accepted)*, Stockholm, 2017.

[14] Mohammed Salah Al-Radhi, Tamás Gábor Csapó, and Géza Németh, "Deep Recurrent Neural Networks in Speech Synthesis Using a Continuous Vocoder," in *SPECOM*, England, UK, 2017.

[15] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE transactions on information and systems*, vol. 7, no. E99-D, pp. 1877-1884, 2016.

[16] Zhizheng Wu, Oliver Watts, and Simon King, "Merlin: An Open Source Neural Network Speech Synthesis System," in *Proceedings of the 9th ISCA Speech Synthesis Workshop*, Sunnyvale, USA, 2016.

[17] Mohammed Salah Al-Radhi, Tamás Gábor Csapó, and Géza Németh, "Continuous vocoder in deep neural network based speech synthesis," *in preparation*, 2017.

[18] Zen H., and Senior A., "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," *ICASSP*, pp. 3844-3848, 2014.