Time-domain envelope modulating the noise component of excitation in a continuous residual-based vocoder for statistical parametric speech synthesis

Mohammed Salah Al-Radhi[1], *Tamás Gábor Csapó*[1,2], *Géza Németh*[1]

[1]Department of Telecommunication and Media Informatics
Budapest University of Technology and Economics, Budapest, Hungary
[2]MTA-ELTE Lendület Lingual Articulation Research Group, Budapest, Hungary
{malradhi,csapot,nemeth}@tmit.bme.hu

Abstract

In this paper, we present an extension of a novel continuous residual-based vocoder for statistical parametric speech synthesis. Previous work has shown the advantages of adding envelope modulated noise to the voiced excitation, but this has not been investigated yet in the context of continuous vocoders, i.e. of which all parameters are continuous. The noise component is often not accurately modeled in modern vocoders (e.g. STRAIGHT). For more natural sounding speech synthesis, four time-domain envelopes (Amplitude, Hilbert, Triangular and True) are investigated and enhanced, and then applied to the noise component of the excitation in our continuous vocoder. The performance evaluation is based on the study of time envelopes. In an objective experiment, we investigated the Phase Distortion Deviation of vocoded samples. A MUSHRA type subjective listening test was also conducted comparing natural and vocoded speech samples. Both experiments have shown that the proposed framework using Hilbert and True envelopes provides high-quality vocoding while outperforming the two other envelopes.

Index Terms: Speech synthesis, continuous vocoder, envelope, triangular, parametric.

1. Introduction

Statistical parametric speech synthesis has been an important research field during the last years due to the development of Hidden Markov Model (HMM) based [1] and deep neural network based approaches [2]. Such a statistical framework is guided by the encoding-decoding (Vocoder) concept which is based on computational models of speech. Since the design of a vocoder depends on speech characteristics, several approaches have been devised as potential solution to the problem of "buzziness" in vocoders. Hu et al. [3] present an experimental comparison of a wide range of important vocoder types which have been previously invented. Despite the fact that most of these vocoders have been successful in synthesizing speech, they are not always successful in synthesizing high quality speech. The reason for this is the inaccurate estimation of the speech model parameters which leads to a degradation in the speech signal. Consequently, we are trying to find a solution in this paper to remove the buzzy quality, while the vocoder remains still computationally efficient.

In HMM based speech synthesis, the accurate modelling of fundamental frequency (also referred to as pitch or F0) is an important aspect. Its modeling with traditional approaches is complicated because pitch values are continuous in voiced regions and discontinuous in unvoiced regions. For modeling discontinuous F0, [4] proposed MSD-HMM and it is generally accepted. However, because of the discontinuities at the boundary between voiced and unvoiced regions, the MSD-HMM is not optimal [5]. To solve this, among others, [6] proposed a continuous F0 model, showing that continuous F0 observations can similarly appear in unvoiced regions. It has also been shown recently that continuous modeling can be more effective in achieving natural synthesized speech [7].

Another excitation parameter is the Maximum Voiced Frequency (MVF) which was recently proposed and shown to result in major improvement in the quality of synthesized speech [8]. During the synthesis of various sounds, the MVF parameter can be used as a boundary frequency to separate the voiced and unvoiced components.

To reconstruct the time-domain characteristics of voiced frames, a time-domain envelope is often applied which was shown to be related to speech intelligibility [9]. There are various methods to obtain a more reliable representation of such envelopes. An early attempt [10], the time envelope is shaped by obtaining peaks of the signal in a window that runs in the data. In [11] a pitch-synchronous triangular envelope is proposed. In [12], Hilbert and energy envelops are introduced. In [13], an iterative technique used to estimate the true envelope. Frequency Domain Linear Prediction (FDLP) envelope is presented in [14]. In vocoding, such envelopes are often used to enhance the source model (e.g. [15], [16], and [17]).

The latest generation of vocoders focuses on uniformly modeling the various types of speech sounds within the same framework. [18] considers a generalized mixed excitation model, in which both periodic and aperiodic components coexist. [19] shows a synthesizer, which uses a uniform representation for voiced and unvoiced segments. This initial study emphasizes the importance of simple excitation models: it claims that the over-parametrization of previous models often leads to statistical learning inefficiencies and intractable tuning issues. As the noise component is not accurately modeled even in the widely used STRAIGHT vocoder [20], Degottex and his colleagues aim to improve this and present a novel noise model, which is of slightly worse quality than STRAIGHT, but it is much simpler [19].

In our earlier work, we proposed a computationally feasible residual-based vocoder [21], using a continuous F0 model [7], and MVF [8]. In this method, the voiced excitation consisting of pitch synchronous PCA residual frames is low-pass filtered while the unvoiced part is high-pass filtered according to the MVF contour as a cutoff frequency. The approach was especially successful for modeling speech sounds with mixed

excitation. In [22], we removed the post-processing step in the estimation of the MVF parameter and thus improved the modelling of unvoiced sounds within our continuous vocoder. The goal of this paper is to further improve our earlier vocoder [22] by using various time domain envelopes for advanced modeling of the noise excitation. We expect that adding such an envelope-modulated noise to the excitation component, the quality of synthesized speech will be closer to natural speech. The novelty of our paper is the unique combination of 1) PCA residual based excitation, 2) continuous F0 modeling, 3) MVF-based mixed voiced and unvoiced excitation, and 4) time-domain envelopes for shaping the high-frequency component of the excitation.

This paper is organized as follows: In Section 2, several time domain envelopes are described. Then, discussion is presented in Section 3. Objective and subjective evaluation is discussed in Section 4. Finally, Section 5 concludes the paper.

2. Methods

2.1. Baseline

The baseline system is our earlier continuous vocoder [22]. During the analysis phase, F0 is calculated on the input waveforms by the open-source implementation (https://github.com/idiap/ssp) 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, MVF is calculated from the speech signal using the MVF_Toolkit (http://tcts.fpms.ac.be/~drugman/files/MVF.zip), resulting in the MVF parameter [8]. In the next step 24-order Mel-Generalized Cepstral analysis (MGC) [23] 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, we perform Principal Component Analysis (PCA) on the pitch synchronous residuals in the baseline system [22].

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 [24].

In the baseline system, 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 [11]. The proposed systems of the current paper differ from the baseline only in the synthesis phase: we test various time envelopes to shape the excitation component and aim to make it more similar to the residual of natural speech. In the next part, we show the envelopes that were used in the study. These will be used 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.

2.2. Time-domain envelope estimation

This work evaluates four different methods for estimating the time envelope of the speech residual signal. The aim of this section is to present a more reliable envelope for our continuous vocoder by assessing and improving the most widely used techniques. The general workflow of the proposed method is presented in Figure 1. The steps involved in this framework are composed of three parts: analysis, statistical modeling, and synthesis. In this paper, we only deal with the analysis and synthesis phases; the statistical modeling is investigated in [25].

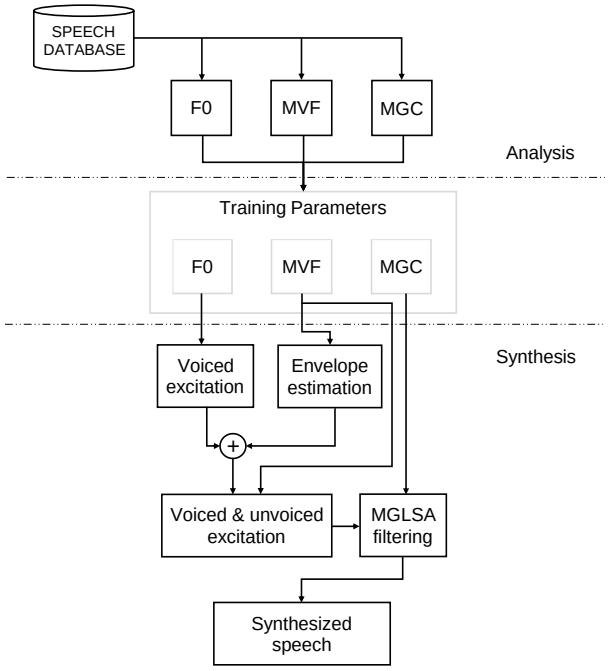

Figure 1. *Workflow of the proposed method.*

2.2.1. Amplitude Envelope

The amplitude envelope is usually defined as filtering the absolute value of the voiced frame $f(n)$ by moving the average filter to the order of $2N+1$, where N is chosen to be 10 [12]. The amplitude envelope is given by

$$A(n) = \frac{1}{2N+1} \sum_{k=-N}^{N} |f(n-k)| \quad (1)$$

Previous work showed that by down-sampling the amplitude envelope to a different number of samples will reduce the relative time square error (RTSE) [26] during parameterizing the noise components.

2.2.2. Hilbert Envelope

Another method of calculating an envelope is based on the Hilbert transform technique [27] [28], which has been used to obtain an analytical signal (complex signal). Then, the Hilbert envelope can be estimated by taking the magnitude of the analytical signal.

2.2.3. Triangular Envelope

A further time domain parametric envelope that can be easily applied to each frame signal is the triangular envelope. It was proposed in [12] by using three parameters as it assumes the triangular is symmetric. In [17], it used a polynomial curve to detect these three parameters (two inflection points and the maximum point on that curve). In the current experiments, we applied similar parameters as [12].

2.2.4. *True Envelope*

Another new approach, which is widely used for estimating the time domain envelope, is called the true envelope (TE). In an iterative procedure, the TE algorithm starts with estimating the cepstrum and updating it in such a way that the original spectrum signal and the current cepstral representation is maximized [29] [30]. To have an efficient real time implementation, [31] proposed a concept of a discrete cepstrum which consists of a least mean square approximation, and [32] added a regularization technique that aims to improve the smoothness of the envelope. Here, the procedures for estimating the TE is shown in Figure 2 in which the cepstrum can be calculated as the inverse Fourier transform of the log magnitude spectrum of the voiced frame. Moreover, TE with weighting factor will bring us 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.

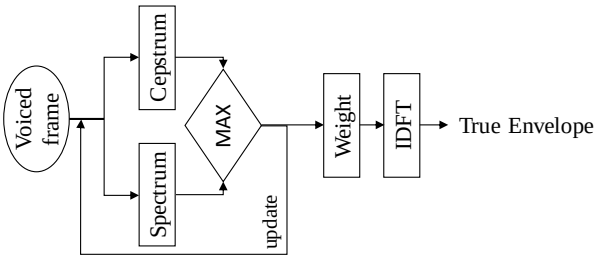

Figure 2. *Procedures for estimating the true envelope.*

3. Discussion

In [22] we found that although the modelling of voiced, unvoiced and mixed excitation sounds was acceptable with a continuous vocoder in HMM-based modelling, there is room for improvement in the noise component of the excitation. Degottex and Erro also argue that the noise component is not accurately modeled in modern vocoders (even in the widely used STRAIGHT vocoder) [33]. They propose a new uniform synthesizer and show that the noise component is improved compared to earlier solutions. It is claimed that the Phase Distortion Deviation (PDD) is a good measure of noisiness [33], therefore we will use this as an objective measure to compare the various versions of our vocoder. The definition of PDD can be found in [33]. Phase Distortion is claimed to be a strong correlate of the maximum-phase component of the voice source. We expect that the vocoded speech samples by the proposed systems are more close to natural speech than those of the baseline. We test this hypothesis with objective and perceptual evaluations.

4. Experimental Results

4.1. Data

Two English speakers were chosen from the CMU-ARCTIC database [34], denoted AWB (Scottish English, male) and SLT (American English, female). 100 sentences (sampled at 16 kHz) from each speaker were analyzed and synthesized with the baseline and proposed vocoders.

4.2. Objective evaluation

We compared the natural and vocoded sentences by measuring the Phase Distortion Deviation at a 5 ms frame shift using covarep/HMPD (http://covarep.github.io/covarep/). As we wanted to quantify the noisiness in the higher frequency bands only, we zeroed out the PDD values below the MVF contour.

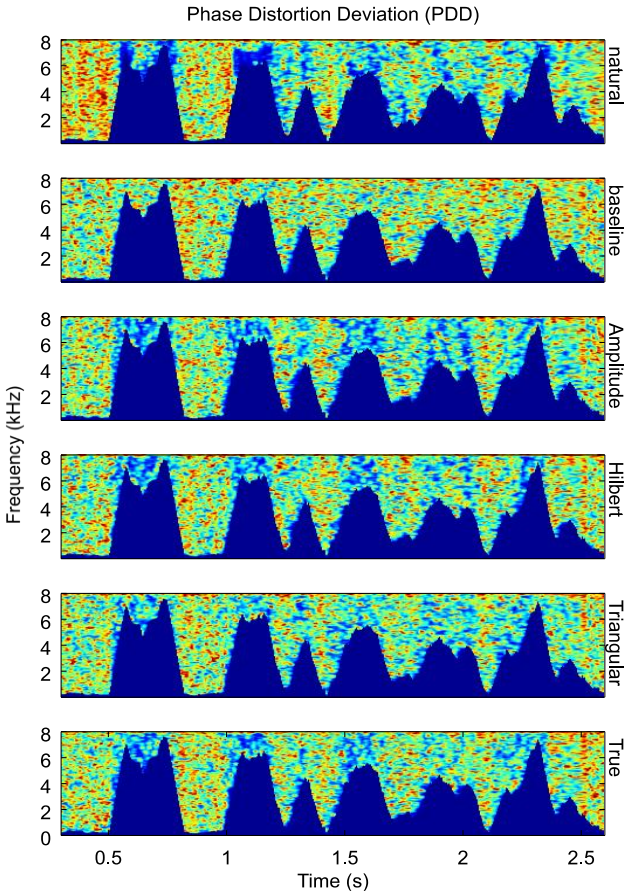

Figure 3. *Phase Distortion Deviation of a natural and vocoded speech samples above the Maximum Voiced Frequency region (sentence: "*He made sure that the magazine was loaded, and resumed his paddling*.", from speaker AWB). The warmer the color, the bigger the PDD value and the noisier the corresponding time-frequency region.*

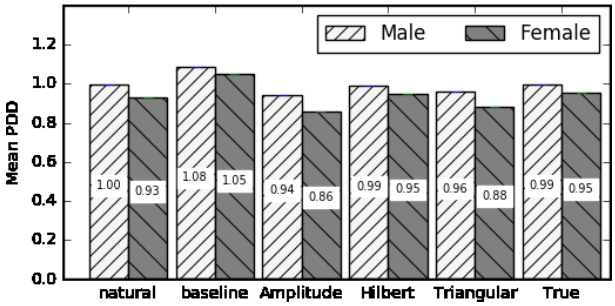

Figure 4. *Mean PDD values by sentence type.*

Samples for the PDD of the natural and vocoded utterances are shown in Figure 3. As can be seen, 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. Figure 4 shows the means of the PDD values of the two speakers grouped by the 6 variants. As can be seen, the PDD values of

the baseline system are significantly higher compared to natural speech. The various envelopes result in different PDD values, but in general they are closer to the natural speech than the baseline. We also conducted Mann-Whitney-Wilcoxon ranksum tests to check the statistical significance. For the male speaker, the systems with 'Hilbert' and 'True' envelopes are not significantly different from natural speech, whereas for the female speaker, only the 'Hilbert' is not significantly different from the natural samples. In general, the 'Amplitude' envelope system results in too low PDD values, meaning that the noisiness is too low compared to natural speech. Otherwise, in general the 'Hilbert' and the 'True' envelopes are closer to natural speech.

4.3. Perceptual evaluation

In this section, we discuss our results obtained with the proposed method by applying four different envelopes and assessing their quality compared to the original speech. For this purpose, the experimental setup is shown first, and then our results are evaluated.

4.3.1. Experimental protocol

As a subjective evaluation, the idea was to select the closeness between the synthesized and original speech signal that fits our goal. In order to evaluate which proposed system is closer to the natural speech, we conducted a web-based MUSHRA (MUlti-Stimulus test with Hidden Reference and Anchor) listening test [35]. The advantage of MUSHRA is that it enables evaluation of multiple samples in a single trial without breaking the task into many pairwise comparisons. Our aim was to measure the perceived correlation of the ratio of the voiced and unvoiced components, therefore we compared natural sentences with the synthesized sentences from the baseline, proposed and a hidden anchor system (the latter being a vocoder with simple pulse-noise excitation). From the 100 sentences used in the objective evaluation, 10 sentences were selected. Altogether, 70 utterances were included in the test (2 speakers x 7 types x 5 sentences). Before the test, listeners were asked to listen to an example from the male speaker to adjust the volume. In the test, the listeners had to rate the naturalness of each stimulus relative to the reference (which was the natural sentence), from 0 (highly unnatural) to 100 (highly natural). The utterances were presented in a randomized order (different for each participant). The listening test samples can be found online: http://smartlab.tmit.bme.hu/interspeech2017_vocoder_envelope.

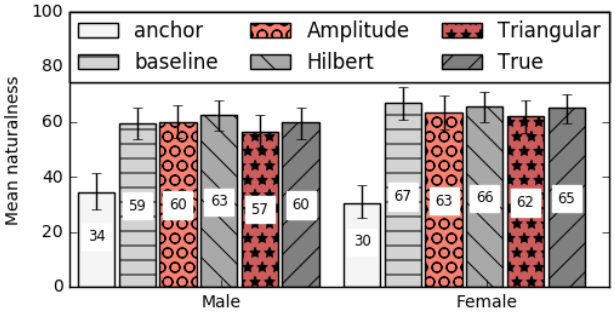

Figure 5. *Results of the subjective evaluation for the naturalness question. Higher value means larger naturalness. Errorbars show the bootstrapped 95% confidence intervals. The score for the reference (natural speech) is not included.*

4.3.2. Listening test results

Twelve participants between the age of $28-74$ (mean age: 38 years) mostly with engineering background were asked to conduct the online listening test. Seven of them were males and five were females. Two of them were native English speakers and none of them reported any hearing loss. On average, the test took 18 minutes to fill.

The MUSHRA scores of the listening test are presented in Figure 5. The results show that all of the systems are clearly better than the hidden anchor. However, there is no significant difference between the proposed systems using the envelopes and the baseline system. From the four envelopes in the study, the Hilbert and True reached the highest naturalness scores in the listening test. For the male speaker, the vocoder using the Hilbert envelope is slightly better than the baseline system, while this is not true for the female speaker.

5. Conclusions

This paper proposed a new approach with the aim of improving our continuous vocoder. The main idea was to further control the time structure of the high-frequency noise component by estimating a suitable time envelope. Four different envelopes (Amplitude, Hilbert, Triangular, and True) were tested from the literature [11], [12], [30].

Both objective (PDD) and subjective (MUSHRA) tests were conducted to evaluate the quality of the synthesized speech signal. Our aim was to find out which time envelope technique is the most effective in our continuous vocoder for high quality speech synthesis. From the objective experiments, which were measuring the Phase Distortion Deviation [34], it can be concluded that the Hilbert and True envelopes are the best when combined with the continuous vocoder (i.e. they are close to the natural sentences in terms of PDD). Despite the fact that these two of the proposed vocoders have a better capability for modeling the time structure of the noise component of the excitation (see e.g. the PDD measurements in Figure 3), in the listening test none of the proposed vocoders were found to be better than the baseline system. There might be two potential reasons for this: a) the weight factor for the high-frequency time envelope modulated noise excitation was not strong enough and listeners were not able to perceive this effect, b) the envelope modulated noise was strong enough, but in certain cases it made the synthesized samples too buzzy (i.e. less natural than the baseline), while other times it outperformed the baseline. Still, we feel that the unique combination of 1) PCA residual based excitation, 2) continuous F0 modeling, 3) MVF-based mixed voiced and unvoiced excitation, and 4) time-domain envelopes for shaping the high-frequency component of the excitation is an important contribution to the field of vocoding for statistical parametric speech synthesis.

As future work, we believe the results obtained in this paper will allow us to evaluate the performance of other types of vocoders. Finally, estimating the time envelope for unvoiced frames can be used as a basis of future developments for our continuous vocoder.

6. 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.

7. References

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, pp. 1039-1064, 2009.

[2] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural network," *in Proc. ICASSP*, p. 7962–7966, 2013.

[3] Q. Hu, K. Richmond, J. Yamagishi, and J. Latorre, "An experimental comparison of multiple vocoder types," in *Proc. ISCA SSW8*, pp.155–160, 2013.

[4] K. Tokuda, T. Mausko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Transactions on Information and Systems*, Vols. E85-D, no. 3, p. 455–464, 2002.

[5] Kai Yu and Steve Young, "Continuous F0 Modeling for HMM Based Statistical Parametric Speech Synthesis," *IEEE Transactions on Audio, Speech, and Language processing*, vol. 19, no. 5, pp. 1071-1079, 2011.

[6] K. Yu, B. Thomson, S. Young, and T. Street, "From Discontinuous To Continuous F0 Modelling In HMM-based Speech Synthesis," *in Proc. ISCA SSW7, Kyoto*, p. 94–99, 2010.

[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] R. Drullman, "Temporal Envelope and Fine Structure Cues for Speech Intelligibility," *Journal of the Acoustical Society of America*, vol. 97, no. 1, pp. 585-591, 1995.

[10] Schloss A., "On the Automatic Transcription of Percussive Music - From Acoustic Signal to High-Level Analysis," Ph.D. thesis, Stanford University, 1985.

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

[12] Y. Pantazis and Y. Stylianou, "Improving the modeling of the noise part in the harmonic plus noise model of speech," *ICASSP*, pp. 4609-4612, 2008.

[13] Axel Robel, Fernando Villavicencio, and Xavier Rodet, "On cepstral and all-pole based spectral envelope modeling with unknown model order," *Pattern Recognition Letters*, vol. 28, pp. 1343-1350, 2007.

[14] D. Ellis and M. Athineos, "Frequency Domain Linear Prediction for Temporal Features," in *Proc. IEEE ASRU Workshop*, 2003.

[15] R. Maia, H. Zen, K. Knill, M. Gales, and S. Buchholz, "Multipulse Sequences for Residual Signal Modeling," *in Proc. Interspeech*, p. 1833–1836, 2011.

[16] T. Drugman and T. Dutoit, "The Deterministic Plus Stochastic Model of the Residual Signal and its Applications," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 20, no. 3, p. 968–981, 2012.

[17] Joao P. Cabral and Julie C. Berndsen, "Towards a Better Representation of the Envelope Modulation of Aspiration Noise," in *Advances in Nonlinear Speech Processing, NOLISP 2013.*, Berlin Heidelberg, 2013.

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

[19] 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.

[20] H. Kawahara, I. Masuda-Katsuse and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3, pp. 187-207, 1999.

[21] Tamás Gábor Csapó, Géza Németh, and M. Cernak, "Residual-Based Excitation with Continuous F0 Modeling in HMM-Based Speech Synthesis," *3rd International Conference on Statistical Language and Speech Processing, SLSP 2015*, vol. 9449, pp. 27-38, 2015.

[22] 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.

[23] 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.

[24] 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.

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

[26] N.P. Narendra and K. Sreenivasa Rao, "Time-domain deterministic plus noise model based hybrid source modeling for statistical parametric speech synthesis," *Speech Communication*, vol. 77, pp. 65-83, 2015.

[27] Ruqiang Yan and Robert X. Gao, "Multi-scale enveloping spectrogram for vibration analysis in bearing defect diagnosis," *Elsevier, Tribology International*, vol. 42, no. 2, pp. 293-302, 2009.

[28] B. Moore, Auditory Processing of Temporal Fine Structure: Effects of Age and Hearing Loss, UK: World Scientific Co., 2014.

[29] A.Robel and X.Rodet, "Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation," in *International Conference on Digital Audio Effects*, Madrid, 2005.

[30] F. Villavicencio, A. RÖbel and X. Rodet, "Improving LPC Spectral Envelope Extraction of Voiced Speech by True-Envelope Estimation," *ICASSP*, vol. 6, pp. 869-872, 2006.

[31] T. Galas and X. Rodet, "An improved cepstral method for deconvolution of source-filter systems with discrete spectra," in *International Computer Music Conference*, Glasgow, 1990.

[32] O. Cappe and E. Moulines, "Regularization techniques for discrete cepstrum estimation," *IEEE Signal Processing*, vol. 3, no. 4, pp. 100-103, 1996.

[33] G. Degottex and D. Erro, "A uniform phase representation for the harmonic model in speech synthesis applications," *EURASIP J. Audio, Speech, Music Process.*, vol. 38, no. 1, pp. 1-16, 2014.

[34] J. Kominek and A. W. Black, "CMU ARCTIC databases for speech synthesis," Carnegie Mellon University, 2003.

[35] ITU-R Recommendation BS.1534, "Method for the subjective assessment of intermediate audio quality," 2001.