Modeling Unvoiced Sounds In Statistical Parametric Speech Synthesis with a Continuous Vocoder

Tamás Gábor Csapó, Géza Németh
Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics
Budapest, Hungary
Email: {csapot,nemeth}@tmit.bme.hu

Milos Cernak, Philip N. Garner
Idiap Research Institute
Martigny, Switzerland
Email: {Milos.Cernak,Phil.Garner}@idiap.ch

***Abstract*—In this paper, we introduce an improved excitation model for statistical parametric speech synthesis. Our earlier vocoder [1], which applies continuous F0 in combination with Maximum Voiced Frequency (MVF), is extended. The focus of this paper is on the modeling of unvoiced consonants, for which two alternative methods are proposed. The first method applies no postprocessing during MVF estimation to reduce the unwanted voiced component of unvoiced speech sounds. The second separates voiced and unvoiced excitation based on the phonetic labels of the text to be synthesized. In an objective experiment we found that the first method produces unvoiced sounds that are closer to natural speech in terms of Harmonics-to-Noise Ratio. A subjective listening test showed that both methods are more natural than our baseline system, and the second method is significantly preferred.**

I. Introduction

There are several main factors in statistical parametric speech synthesis that are needed to deal with in order to achieve as high quality synthesized speech as with the unit selection approach. These include the improved vocoder techniques, acoustic modeling accuracy and over-smoothing during parameter generation [2]. In this paper, we investigate the first of these: vocoding in hidden Markov-model (HMM) based text-to-speech (TTS) synthesis. A large number of such vocoders, also called as excitation models, have been proposed in the last few years, including mixed excitation [3], [4], glottal source parameter based [5]–[8], Harmonics-to-Noise model based [9], [10] and residual based [11]–[14] vocoders (for a comparison, see [15]). These all have the aim of reducing the "buzziness" caused by oversimplified vocoding methods in early versions of HMM-TTS. Although there are vocoding methods which yield in close to natural synthesized speech, they are typically computationally expensive, and are thus not suitable for real-time implementation. Therefore, we are seeking a computationally feasible solution in this paper.

Traditionally, using standard pitch tracking methods in excitation models, the F0 contour is discontinuous at voiced-unvoiced (V-UV) and unvoiced-voiced (UV-V) boundaries. For handling discontinuous F0, Multi-Space Distribution (MSD) was proposed for use with HMMs, which involves building separate models for voiced and unvoiced frames [16]. However, it has been recently shown that excitation models using continuous F0 have several advantages in statistical parametric speech synthesis [17]. First of all, the inaccurate MSD-HMM modeling around V-UV and UV-V transitions can be omitted. Second, it was found that more expressive F0 contours can be generated using a continuous F0 than using standard F0 models [18], [19]. In such continuous systems, voicing strength or voicing label is often used for modeling the voicing feature separately [20], [21]. Another important observation is that the voiced/unvoiced decision can be left up to the aperiodicity features in a mixed excitation vocoder [22]. This decision can also be modeled using a dynamic voiced frequency [14], [23]. Furthermore, continuous F0 models can be effectively used with noisy speech [24].

In our earlier work, we proposed a residual codebook based excitation model [25], which was integrated into HMM-TTS [26]. In [1], we extended the model with 1) a Principal Component Analysis based residual to reduce buzziness, 2) a continuous F0 model [17] to decrease the disturbing effect of creaky voice and 3) Maximum Voiced Frequency (MVF) [23] to model the voiced/unvoiced characteristics of sounds. Here, MVF is a continuous measure of voicing and the excitation is composed as the sum of a lower frequency voiced component and a higher frequency unvoiced component, separated by the MVF parameter stream. Although the above combination was successful in diminishing the artifacts caused by creaky voice, in a subjective listening test we found that unvoiced sounds have sometimes a too strong voiced component, as a result of using continuous F0 in combination with MVF.

In this paper, we extend our earlier vocoder [1] by two alternative methods for modeling unvoiced sounds. In Section II, the novel methods are proposed, followed by a discussion in Section III. Section IV shows an objective evaluation and a subjective evaluation of these methods. Finally, in Section V we conclude the paper.

II. Methods

The baseline and both proposed systems are composed of analysis, statistical modeling and synthesis phases. First, the baseline system is introduced with all three phases in Section II-A. After that, the novelties of the proposed data-driven (Section II-B) and rule-based (Section II-C) extensions are shown for modeling unvoiced sounds. The general framework of the proposed methods is shown in Fig. 1.

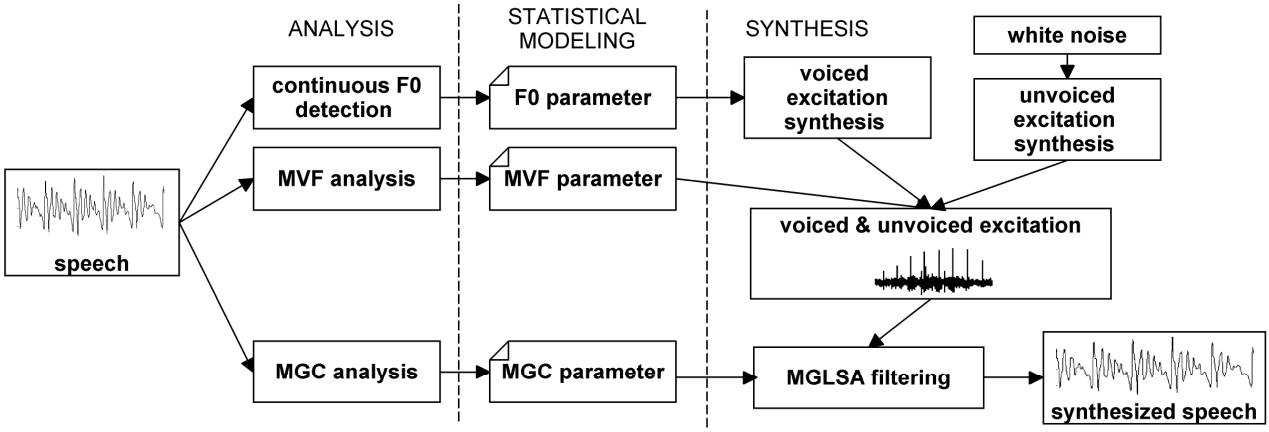

Fig. 1. General framework of the proposed methods.

A. Baseline

1) Analysis: In the baseline system, first the fundamental frequency (F0) parameter is calculated on the input waveforms sampled at 16 kHz by the open-source implementation [27] of a simple continuous pitch tracker [17], denoted as 'F0cont'. 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 [23], resulting in the MVF parameter. In the next step 24-order Mel-Generalized Cepstral analysis (MGC) [28] 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 on the pitch synchronous residuals in the baseline system. In the synthesis phase, the first principal component of this PCA residual is used (for the details of the calculation and samples, see Fig. 1 in [1]).

2) Statistical modeling: For training, the logarithmic values of the parameters are calculated from each frame to describe the excitation (F0cont and MVF) and the spectrum (MGC). As all parameter streams are continuous, they are modeled as simple HMMs, avoiding thus MSD-HMM modeling. The first and second derivatives of all the parameters are also stored in the parameter files and used in the training phase. Decision tree-based context clustering is used with context dependent labeling applied in the English version of HTS 2.3beta [29], [30]. Independent decision trees are built for all the parameters and duration using a maximum likelihood criterion.

3) Synthesis: The right part of Fig. 1 shows the steps applied in the synthesis part of the baseline system. First, PCA residuals are overlap-added resulting in a voiced excitation, and the density of the residual frames is dependent on the F0cont parameter. The unvoiced part of the excitation is based on white noise. As there is no strict voiced / unvoiced decision in this stream, the MVF parameter models the voicing information: for unvoiced sounds, the MVF is low (around 1 kHz), for voiced sounds, the MVF is high (typically above 4 kHz), whereas for mixed excitation sounds, the MVF is in between (e.g. for voiced fricatives, MVF is around 2–3 kHz).

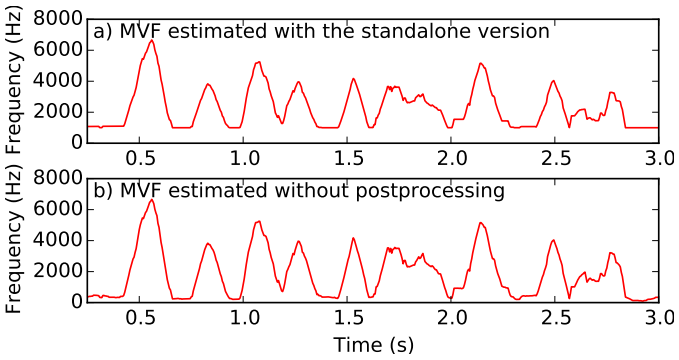

Fig. 3. MVF with the standalone version of MVF_Toolkit and without postprocessing on the '*He had fulfilled his duty and paid properly.*' sentence.

Voiced excitation is lowpass filtered, unvoiced excitation is highpass filtered depending on the MVF parameter stream, and they are added together frame by frame. Finally, an MGLSA filter is used to synthesize speech from the excitation and the MGC parameter stream [31].

4) Demonstration sample: A sample for the generated MVF parameter stream and for the spectrogram of a synthesized sentence can be seen in Fig. 2 a). At unvoiced segments (e.g. around 0.6 s, 1.6 s, 1.8 s), the MVF value is close to 1 kHz, and thus there is a relatively strong voiced component below this frequency even in case of the unvoiced sounds.

In [1], we conducted a listening test of English speech synthesis samples, where the ratings of the listeners showed that there is room for improvement in modeling the unvoiced sounds with this continuous F0 model. Although MVF-based mixed voiced and unvoiced excitation was found to be extremely useful for modeling the voiced fricatives and other voiced sounds, the voiced component in case of the unvoiced sounds resulted in a disturbing 'buzzy' effect.

B. Proposed #1: data-driven modeling of unvoiced sounds

The goal of the first proposed system is to improve the estimation of the Maximum Voiced Frequency, in order to obtain more realistic values for unvoiced sounds. The standalone version of MVF_Toolkit contains a post-processing step that smooths the estimated MVF [23]. In the context of HMM

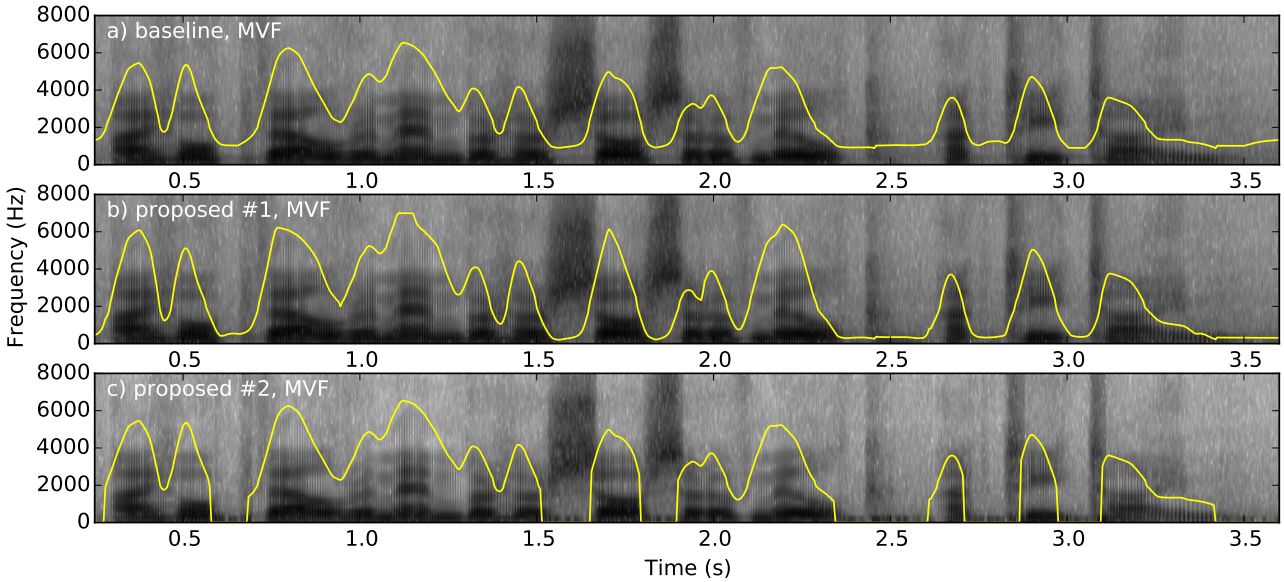

Fig. 2. Sample from the baseline system and the two proposed methods. The synthesized sentence is '*I hope they'll remember her saucer of milk at tea-time.*'

synthesis, we find that this can lead to voicing errors and too high minimal values for the MVF stream. However, we would also expect that the HMM would also smooth the MVF when MVF is used as a feature. This implies that the post-processing is not necessary. Fig. 3 compares the result of the MVF estimation, showing that the standalone version has a 1000 Hz floor for the MVF contour, whereas in the version without post-processing lower values appear as well.

1) Analysis, statistical modeling and synthesis: In system Proposed #1, the analysis is similar to the baseline system, but the MVF estimation does not include post-processing. The phases of statistical modeling and synthesis are the same as in the baseline system.

2) Demonstration sample: Fig. 2 b) shows a sample for the generated MVF parameter stream and for the spectrogram of a synthesized sentence using system Proposed #1. At unvoiced segments, the generated MVF value is now significantly lower (around 200–500 Hz) than with the baseline system. This means that although there is still a voiced component, it occurs only in a very short spectral band at the lowest frequencies. It is also worth noting that because no post-processing was used in the MVF estimation, the HMMs learnt better this parameter stream, and the MVF values for voiced sounds are typically higher than with the baseline system (e.g. around 1.2 s and 2.2 s).

C. Proposed #2: rule-based modeling of unvoiced sounds

In system Proposed #2, we introduce a rule-based extension to generate excitation for the unvoiced sounds.

1) Analysis, statistical modeling and synthesis: The analysis and the statistical modeling phases are the same as in the baseline system. During synthesis, unvoiced sounds are found based on the phonetic labels from the text to be synthesized, and the excitation is generated fully based on white noise. This is equal to setting the MVF value to 0 during the addition of the voiced and unvoiced components in the excitation generation of unvoiced sounds.

2) Demonstration sample: A synthesized sentence is shown in Fig. 2 c) for the Proposed #2 system. Here, the voiced component is fully omitted from the unvoiced sounds, which can be an advantage. However, a potential disadvantage of this system might be that abrupt changes occur at voiced-unvoiced and unvoiced-voiced transitions in the excitation, because there is no smooth transition as in the baseline and Proposed #1 systems. This contradicts to natural speech where V-UV and UV-V changes are smoother.

D. Benchmark

For the subjective evaluation in Section IV-C, we used a benchmark system from HTS-demo [29], [30]. This system uses pulse-noise excitation and MSD-HMMs for statistical modeling of the F0 stream.

III. Discussion

The MVF estimation is dependent on the estimated pitch track [23]. Although the standalone MVF estimation was originally proposed to use with standard discontinuous F0 (e.g. [14]), our previous research has shown that the MVF_Toolkit is suitable to estimate a continuous MVF trajectory when using a continuous F0 as input [1]. However we found that a vocoder using continuous F0 in combination with continuous MVF cannot accurately model the unvoiced components of speech. We hypothesize that both the data-driven and rule-based MVF modeling strategies of this paper will be superior to the baseline system because of the novel modeling of the unvoiced sounds. We expect that the lower MVF values of the Proposed #1 system will lead to less buzziness while keeping the advantages of using only continuous parameters (e.g. having a smoother transition at voiced-unvoiced and unvoiced-voiced boundaries). The Proposed #2 system is expected to be more natural than a vocoder with discontinuous F0,

TABLE I
MEAN HNR VALUES GROUPED BY SOUND AND SENTENCE TYPE.

	natural		baseline		proposed1		proposed2	
	AWB	SLT	AWB	SLT	AWB	SLT	AWB	SLT
ch	0.79	1.11	0.85	1.63	0.81	1.54	0.63	1.01
f	1.13	1.21	1.36	2.23	1.27	2.09	1.13	1.55
k	1.18	1.78	1.55	2.30	1.34	2.08	0.83	1.24
p	0.66	1.53	3.08	2.61	2.43	2.35	0.53	1.53
s	0.73	1.25	0.82	1.34	0.78	1.19	0.62	0.88
sh	0.69	1.25	0.94	1.70	0.89	1.57	0.65	1.26
t	1.02	2.92	1.73	4.69	1.50	4.38	0.90	1.66
th	1.05	1.76	1.81	2.89	1.64	2.72	0.99	2.03

because the result of continuous modeling and interpolation in F0 estimation is that there are no voicing errors caused by the statistical modeling. We test these hypotheses with both objective and subjective evaluation experiments.

IV. EXPERIMENTAL RESULTS

A. Data

Two English speakers were chosen from the CMU-ARCTIC database [32], denoted AWB (Scottish English, male) and SLT (American English, female). 90% of the sentences (1024 and 1018, respectively) were used for single speaker training with the HMMs and the remaining sentences were used for testing.

B. Objective evaluation

First, we selected the last 10% of the natural sentences from the CMU-ARCTIC database (114 sentences for both speakers), and synthesized them based on their labels for both speakers and using all three systems. After that, we compared the natural and synthesized sentences by investigating the unvoiced sounds. From the natural and synthesized speech data, we measured a Harmonics-to-Noise (HNR) ratio at a 5 ms frame shift using SSP [17], [27]. After that, the frame by frame HNR values corresponding to the unvoiced sounds were collected and compared with each other, grouped by the sound and by the sentence type. Table I shows the results of this comparison for each unvoiced sound. In all cases, $HNR_{baseline} > HNR_{proposed1} > HNR_{proposed2}$, showing that the unvoiced components were decreased in both proposed systems compared to the baseline. The $HNR_{natural}$ values are usually between those of the two proposed systems. Fig. 4 shows the same trend for both speakers. According to a statistical analysis, all differences are statistically significant at $p < 0.05$ level. From these results, we can see that the synthesized sentences by Proposed #1 and Proposed #2 systems are closer to the natural sentences in terms of the Harmonics-to-Noise ratio of the unvoiced sounds than the baseline system. However, the $HNR_{proposed2}$ values are somewhat lower than $HNR_{natural}$, indicating that the ratio of the unvoiced components might be too high in this system.

C. Subjective evaluation

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 [33]. The advantage of MUSHRA is that it enables evaluation

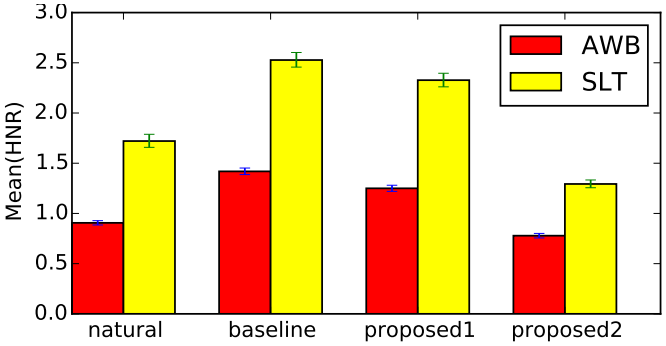

Fig. 4. Mean HNR values by sentence type. Errorbars show the bootstrapped 95% confidence intervals.

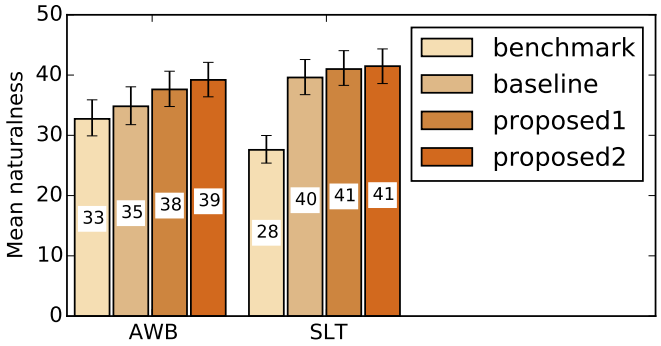

Fig. 5. Results of the subjective evaluation for the naturalness question. Errorbars show the bootstrapped 95% confidence intervals. The score for the natural speech is not included, because it is always 100.

of multiple samples in a single trial without breaking the task into many pairwise comparisons. Our aim was to measure the perceived correlate of the ratio of the voiced and unvoiced components, therefore we compared natural sentences with the synthesized sentences from the baseline, Proposed #1, Proposed #2 systems and the benchmark system. From the 114 sentences used in the objective evaluation, the 10 sentences having the highest ratio of unvoiced sounds were selected. Altogether, 100 sentences were included in the test (2 speakers · 5 systems · 10 sentences). 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).

Altogether 10 listeners participated in the test (1 female, 9 males). They were all speech experts, between 25-57 years (mean: 40 years). One of them was a native speaker of English. On average the whole test took 17 minutes to complete. The MUSHRA scores of the listening test are presented in Fig. 5 for the two speakers and five types. The figure shows that the two proposed systems outperform the baseline system for both speakers. The ratings of the listeners were compared by Mann-Whitney-Wilcoxon ranksum tests as well, with a 95% confidence level, showing that the Proposed #2 system was significantly preferred over the baseline in case of speaker AWB. The other differences are not statistically significant, but Fig. 5 indicates improvements even in case of speaker SLT and the Proposed #1 system. These tendencies show that

the result of the improved unvoiced sound modeling methods was perceivable for the subjects of the listening test.

V. Conclusions

In this paper, we introduced two alternative methods (data-driven and rule-based) for improved modeling of unvoiced sounds in statistical parametric speech synthesis. In an objective experiment and a subjective listening test both methods were found to be more natural than our baseline system, therefore the hypotheses of Section III can be accepted. Listeners slightly preferred the Proposed #2 system over the Proposed #1 system, and these differences were larger for the male speaker, whose original recordings contained high background noise.

The advantage of this continuous vocoder is that it is relatively simple: it has only two 1-dimensional parameters for modeling excitation (F0cont and MVF) and the synthesis part is a computationally feasible solution, therefore speech generation can be performed in real-time. In the future, we plan 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.

Acknowledgment

The authors would like to thank the listeners for participating in the subjective test. This research is partially supported by the Swiss National Science Foundation via the joint research project (SCOPES scheme) SP2: SCOPES project on speech prosody (SNSF no IZ73Z0_152495-1).

References

[1] T. G. Csapó, G. Németh, and M. Cernak, "Residual-Based Excitation with Continuous F0 Modeling in HMM-Based Speech Synthesis," in *Lecture Notes in Artificial Intelligence*, A.-H. Dediu, C. Martín-Vide, and K. Vicsi, Eds. Budapest, Hungary: Springer International Publishing, 2015, vol. 9449, pp. 27–38.

[2] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, nov 2009.

[3] T. Yoshimura and K. Tokuda, "Mixed excitation for HMM-based speech synthesis," in *Proc. Eurospeech*, Aalborg, Denmark, 2001, pp. 2263–2266.

[4] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE Transactions on Information and Systems*, vol. E90-D, no. 1, pp. 325–333, 2007.

[5] J. P. Cabral, S. Renals, J. Yamagishi, and K. Richmond, "HMM-based speech synthesiser using the LF-model of the glottal source," in *Proc. ICASSP*, Prague, Czech Republic, 2011, pp. 4704–4707.

[6] T. Raitio, A. Suni, M. Vainio, and P. Alku, "Comparing glottal-flow-excited statistical parametric speech synthesis methods," in *Proc. ICASSP*, Vancouver, Canada, 2013, pp. 7830–7834.

[7] T. Raitio, H. Lu, J. Kane, A. Suni, M. Vainio, S. King, and P. Alku, "Voice source modelling using deep neural networks for statistical parametric speech synthesis," in *Proc. EUSIPCO*, Lisbon, Portugal, 2014, pp. 2290–2294.

[8] G. Degottex, P. Lanchantin, A. Roebel, and X. Rodet, "Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis," *Speech Communication*, vol. 55, no. 2, pp. 278–294, feb 2013.

[9] D. Erro, I. Sainz, E. Navas, and I. Hernáez, "Improved HNM-based Vocoder for Statistical Synthesizers," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 1809–1812.

[10] Z. Wen and J. Tao, "Inverse Filtering Based Harmonic plus Noise Excitation Model for HMM-based Speech Synthesis," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 1805–1808.

[11] ——, "Amplitude spectrum based Excitation model for HMM-based Speech Synthesis," in *Proc. Interspeech*, Portland, Oregon, USA, 2012, pp. 1428–1431.

[12] C.-s. Jung, Y.-s. Joo, and H.-g. Kang, "Waveform Interpolation-Based Speech Analysis/Synthesis for HMM-Based TTS Systems," *IEEE Signal Processing Letters*, vol. 19, no. 12, pp. 809–812, dec 2012.

[13] T. Drugman and T. Dutoit, "The Deterministic Plus Stochastic Model of the Residual Signal and its Applications," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 3, pp. 968–981, mar 2012.

[14] T. Drugman and T. Raitio, "Excitation Modeling for HMM-based Speech Synthesis: Breaking Down the Impact of Periodic and Aperiodic Components," in *Proc. ICASSP*, Florence, Italy, 2014, pp. 260–264.

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

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

[17] 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.

[18] 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, Japan, 2010, pp. 94–99.

[19] K. Yu and S. 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.

[20] Q. Zhang, F. K. Soong, Y. Qian, Z. Yan, J. Pan, and Y. Yan, "Improved modeling for F0 generation and V/U decision in HMM-based TTS," in *Proc. ICASSP*, Dallas, Texas, USA, 2010, pp. 4606–4609.

[21] K. Yu and S. Young, "Joint modelling of voicing label and continuous F0 for HMM based speech synthesis," in *Proc. ICASSP*, Prague, Czech Republic, 2011, pp. 4572–4575.

[22] J. Latorre, M. J. F. Gales, S. Buchholz, K. Knil, M. Tamura, Y. Ohtani, and M. Akamine, "Continuous F0 in the source-excitation generation for HMM-based TTS: Do we need voiced/unvoiced classification?" in *Proc. ICASSP*, Prague, Czech Republic, 2011, pp. 4724–4727.

[23] 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.

[24] K. U. Ogbureke, J. P. Cabral, and J. Carson-Berndsen, "Using Noisy Speech to Study the Robustness of a Continuous F0 Modelling Method in HMM-based Speech Synthesis," in *Proc. Speech Prosody*, Shanghai, China, 2012, pp. 67–70.

[25] T. G. Csapó and G. Németh, "A novel codebook-based excitation model for use in speech synthesis," in *IEEE CogInfoCom*, Kosice, Slovakia, dec 2012, pp. 661–665.

[26] ——, "Statistical parametric speech synthesis with a novel codebook-based excitation model," *Intelligent Decision Technologies*, vol. 8, no. 4, pp. 289–299, 2014.

[27] "Speech Signal Processing - a small collection of routines in Python to do signal processing [Computer program]," 2015. [Online]. Available: https://github.com/idiap/ssp

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

[29] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, and A. Black, "The HMM-based speech synthesis system version 2.0," in *Proc. ISCA SSW6*, Bonn, Germany, 2007, pp. 294–299.

[30] "HMM-based Speech Synthesis System (HTS) [Computer Program], Version 2.3beta." [Online]. Available: http://hts.sp.nitech.ac.jp/archives/2.3beta/HTS-2.3beta_for_HTK-3.4.1.tar.bz2

[31] 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.

[32] J. Kominek and A. W. Black, "CMU ARCTIC databases for speech synthesis," Language Technologies Institute, Tech. Rep., 2003.

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