EXCITATION MODELING FOR HMM-BASED SPEECH SYNTHESIS: BREAKING DOWN THE IMPACT OF PERIODIC AND APERIODIC COMPONENTS

Thomas Drugman[1]*, Tuomo Raitio*[2]

[1]TCTS Lab – University of Mons, Belgium
[2]Aalto University, Department of Signal Processing and Acoustics, Espoo, Finland

ABSTRACT

HMM-based speech synthesis generally suffers from typical buzziness due to over-simplified excitation modeling of voiced speech. In order to alleviate this effect, several studies have proposed various new excitation models. No consensus has however been reached on what is the perceptual importance of the accurate modeling of the periodic and aperiodic components of voiced speech, and to what extent they separately contribute in improving naturalness. This paper considers a generalized mixed excitation modeling, common to various existing approaches, in which both periodic and aperiodic components coexist. At least three main factors may alter the quality of synthesis: periodic waveform, noise spectral weighting, and noise time envelope. Based on a large subjective evaluation, the goal of this paper is threefold: *i)* to evaluate the relative perceptual importance of each factor, *ii)* to investigate what is the most appropriate method to model the periodic and aperiodic components, and *iii)* to provide prospective clues for future work in excitation modeling.

Index Terms— HMM-based speech synthesis, excitation modeling, glottal flow, residual signal

1. INTRODUCTION

Statistical parametric speech synthesis based on hidden Markov models (HMMs) [1] emerged this last decade as a promising technique for the automatic generation of speech from text. This approach exhibits several advantages over concatenative speech synthesis approach [2]: flexibility to change the voice characteristics [3, 4, 5, 6], reduced memory footprint [7, 8], and enhanced robustness [9]. Nonetheless, although some progress has been achieved these last years, its main flaw is a degraded speech quality. This can be explained by two main factors: *i)* the synthesis relies on a parametric representation of the speech signal which results in a typical *buzziness*; *ii)* the synthesis relies on a statistical modeling of a given speech database, which results in a typical *muffledness* caused by oversmoothed generated trajectories.

This paper addresses the first issue and aims to enhance the naturalness of synthesized speech by improving the *excitation modeling*. In general, the modeling of speech is based on the source-filter approach. In this framework, two options are possible according to what is considered to be the source and the filter. In the first case, the source is the glottal (air) flow as physiologically produced by the vocal organs, and the filter refers to the vocal tract response. Beyond the physiological motivation, this approach has the advantage to be more flexible, as proper modifications of the glottal contribution are expected to reflect changes in voice quality. Nonetheless, this approach requires the reliable and accurate separation of these components from each other using glottal inverse filtering, which is a difficult inverse problem [10, 11]. In the second case, the filter corresponds to the overall spectral envelope of speech and the excitation is the residual signal obtained by feeding the speech signal through the inverse of the estimated filter. The residual signal has the advantage to be easily obtained, however its amplitude spectrum is by definition flat and the information about the glottal spectral contribution is inextricably mixed in the filter component. As a consequence, its flexibility for speech modifications is more limited.

In all cases, separating the source and filter contribution is important as it can lead to their distinct characterization and modeling. Methods parameterizing the filter, such as the well-known linear prediction (LP) or mel-cepstral features [12], are widely used. On the contrary, methods modeling the excitation signal are still not well established and the accurate and perceptually relevant modeling of the excitation would benefit many speech processing areas.

The basic excitation model makes use of either a quasi-periodic impulse train for voiced speech, or white noise for unvoiced speech. The simple representation of voiced speech makes the resulting synthesis sound buzzy due to zero-phase nature of the excitation. Various studies have focused on improving the excitation model by mixing periodic excitation with aperiodic noise, such as in the mixed excitation (ME) [13] approach. In ME, voiced excitation is composed of both periodic and aperiodic components of which relative magnitudes are controlled by band-pass voicing strengths. In a similar way, a ME consisting of a set of high-order state-dependent filters derived through a closed-loop procedure was proposed in [14]. In [15], a hybrid approach makes use of a codebook of pitch-synchronous residual frames which are selected at synthesis time according to the down-sampled version of the excitation. In [16, 17], the deterministic plus stochastic model (DSM) of the residual signal is proposed. DSM excitation consists of two components: the deterministic waveform called eigenresidual, which is obtained by principal component analysis (PCA) on a set of pitch-synchronous residual frames, and an aperiodic excitation delimited by maximum voiced frequency F_m and modulated in time according to a speaker-specific time envelope.

In parallel, similar improvements using a glottal flow modeling have been introduced. The approach described in [18] incorporates the Liljencrants–Fant (LF) [19] model so as to reduce the buzziness and increase the flexibility. A natural glottal flow pulse estimated by glottal inverse filtering from a sustained vowel is modified according to voice source features and mixed with noise in the so-called GlottHMM approach presented in [20] and further refined in [21]. A synthesis approach using LF model was also introduced in [22]. In [23], a glottal source pulse library is extracted from natural speech and pulses are selected according to voice source features for synthesis.

T. Drugman is supported by FNRS. T. Raitio is supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 287678. The authors would like to thank Vasilis Karaiskos for running the listening tests.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

All these techniques (modeling either residual or glottal flow) have been shown to provide a higher naturalness in HMM-based speech synthesis, compared to the traditional pulse excitation.

Despite all the advances in excitation modeling, no consensus has been reached yet on the perceptual effect of each component in voice source modeling, and to what extent they separately contribute in improving naturalness. In the frame of HMM-based speech synthesis, this paper investigates the perceptual impact of the three main factors in excitation modeling: waveform used for periodic excitation, spectral weighting between the periodic and aperiodic components, and the envelope used for the time modulation of the noise. The goal of this paper is threefold: *i)* to evaluate the relative importance each component in modeling the excitation, *ii)* to investigate what is the most appropriate method to model these components, *iii)* to provide prospective clues for future work in excitation modeling.

The paper is structured as follows. Section 2 presents the general vocoding framework used in various existing approaches and describes the alternatives considered throughout this paper. Section 3 deals with the experimental protocol, providing details about the implementation of our HMM-based speech synthesizers and describing the subjective evaluation and its results. Section 4 finally discusses the implications of the study and concludes the paper.

2. GENERAL VOCODING FRAMEWORK

The great majority of excitation models rely on a similar mixed excitation model in which both periodic and aperiodic components coexist during the production of voiced sounds. The workflow of this generalized vocoder is displayed in Fig. 1. The periodic contribution of the excitation $e_p(t)$ is obtained from a specific waveform whose duration is adapted to the current F_0 value, and which is then filtered using some aperiodicity measurements. As for the aperiodic excitation component $e_a(t)$, it results from a white Gaussian noise that is spectrally modified using these same aperiodicity measurements and modulated in time using a given envelope. Note that all this process is achieved pitch-synchronously. The two components $e_p(t)$ and $e_a(t)$ are then summed up and the pitch-synchronous windowed frames are overlap-added. The resulting excitation contribution finally goes through the filter to give the speech signal. The three main factors impacting the performance of this generalized excitation model are now studied in the remainder of this paper: periodic waveform, noise spectral weighting and noise envelope.

2.1. Periodic waveform

In the simplest source-filter vocoder, Dirac pulses at fundamental period intervals are used to create the voiced excitation. Usually improvements in excitation modeling are compared with either this simple model or the mixed excitation [13], which is used e.g. in the most commonly used vocoder STRAIGHT [24, 25]. Improvements

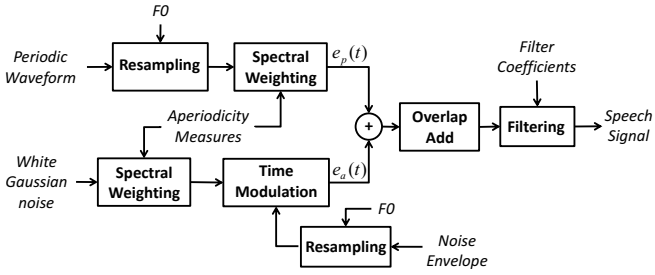

Fig. 1. Workflow of generalized vocoder using mixed excitation.

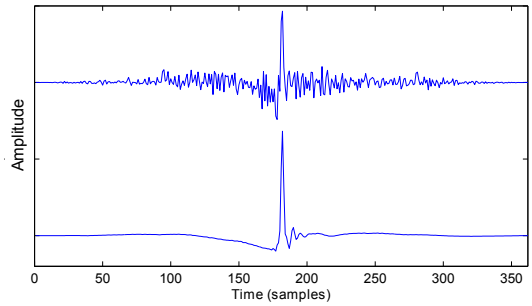

Fig. 2. Natural residual excitation frame (*upper signal*) and eigenresidual (*lower signal*) for speaker AWB.

over the simple excitation are rather easy to achieve either by using more natural periodic waveform or by mixing the periodic component with noise. However, the comparison between more complex methods (e.g. STRAIGHT) may be ambiguous, since evaluations are usually made between whole vocoder architectures using different parameterization methods, parameters representations, and HMM training. Also, the contributions of the periodic and aperiodic components are usually left undetermined.

Only few studies have addressed the perceptual differences between deterministic waveforms other than impulse. Experiments in [26] have shown that a mean glottal flow pulse (similar to eigenresidual in [17]) was rated better in quality than excitation using selection of natural pulses and equal to a pulse reconstructed from 12 PCA components. The latter comparison was also informally done using residual waveform in [17] with the same conclusion that using more components for modeling does not improve quality. In creaky voice synthesis [27], the type of deterministic waveform has also been shown to have relevant perceptual effect. However, it is still not clear what is the perceptual effect of the type of deterministic waveform on speech quality in general.

In this paper, we consider the reconstruction of the residual signal with three possible periodic waveforms: *i)* the Dirac impulse as used in the simplest vocoder; *ii)* a *natural* excitation residual frame; *iii)* speaker-dependent eigenresidual as proposed in [17]. Note that the choice of the natural residual frame was not arbitrary and resulted from the consideration of several criteria: *a)* having a low pitch to avoid as much as possible up-sampling to the target F_0 (as this will cause energy holes in high frequencies); *b)* its amplitude spectrum must be as flat as possible to avoid artefacts due to residual resonances; *c)* having a clear discontinuity at the glottal closure instant (GCI). The natural residual and eigenresidual for the male speaker considered in this paper are illustrated in Fig. 2.

2.2. Noise spectral weighting

In order to reduce buzziness caused by the zero-phase excitation, it has been shown to be beneficial to adopt an approach in which both periodic and aperiodic components may coexist [13]. Two main techniques were proposed in the literature for this purpose. The first one relies on a multiband approach where, for each spectral band, the energy ratio between the periodic and aperiodic contributions is controlled by *aperiodicity measurements*. These measurements can be computed in various ways. In [13], they consist of correlation coefficients calculated in each band, while in [25, 23] they are determined based on the ratio between the upper and lower smoothed spectral envelopes. The second technique for spectral weighting makes use of a maximum voiced frequency (F_m) which demarcates the boundary between the periodic component (which holds only in the low

frequencies) and the aperiodic component (which holds only in the high frequencies). This idea originates from the multiband excitation vocoder [28], which was was later integrated into several methods for excitation modeling in HMM-based speech synthesis [16, 22].

The perceptual effect of these methods has not been studied in the context of HMM-based speech synthesis, Thus, four options for spectral weighting are investigated in this paper: *i)* the aperiodic component is discarded and the excitation consists only of the periodic contribution; *ii)* use of a static maximum voiced frequency F_m fixed to 4 kHz as is done in [29] and [17]; *iii)* use of dynamic F_m value estimated using the algorithm described in [30]; *iv)* use of the harmonic-to-noise ratio (HNR) measurements proposed in [23].

2.3. Envelope for noise modulation

In addition to modeling the spectral characteristics of the noise, some studies have addressed its time properties. The motivation for this arises from the observation that the time distribution of the noise is not uniform and exhibits a synchronization with the glottal cycle. In [30], a pitch-synchronous triangular envelope is proposed. In [31], authors compare the triangular and Hilbert energy envelopes in the frame of HNM and report a slight improvement. In [32], an alternative parametric representation of a triangular envelope is proposed. It is however worth mentioning that none of these works have been tested in HMM-based speech synthesis, which requires slowly-varying parameter trajectories for a proper statistical modeling. Finally, a speaker-dependent noise waveform envelope was proposed in [17], which is extracted by averaging GCI-synchronous Hilbert envelopes of the stochastic part of the excitation.

Three possible noise envelopes are studied in this paper: *i)* uniform distribution; *ii)* the triangular window proposed in [30]; *iii)* the speaker-dependent Hilbert envelope proposed in [17]. An illustration of this latter waveform is shown in Fig. 3 for the female speaker considered in this study.

3. EXPERIMENTS

3.1. HMM-based voice building

In order to find out the perceptual effect of each of the studied excitation component, HMM-based voices were built and used in subjective listening tests. To prevent perceptual effects due to other factors than the ones in study, a single system architecture was used, capable of producing all the different component combinations. The speech features used to train the HMMs are depicted in Table 1. In feature extraction, fundamental frequency (F_0) and HNR were extracted using the GlottHMM vocoder [21, 23] while SPTK 3.6 [33] was used to extract speech spectrum. The spectrum was estimated using a 30th order mel-generalised cepstral (MGC) analysis [34] with $\alpha = 0.42$ and $\gamma = -1/3$. MGCs were then converted to line spectral frequencies (LSF) for better parameter representation for HMM training. F_m was estimated by the algorithm described in [30]. All other data such as the periodic waveform or the noise envelope have been extracted as explained in Section 2 by a GCI-synchronous analysis, where GCIs are detected using the SEDREAMS algorithm [35].

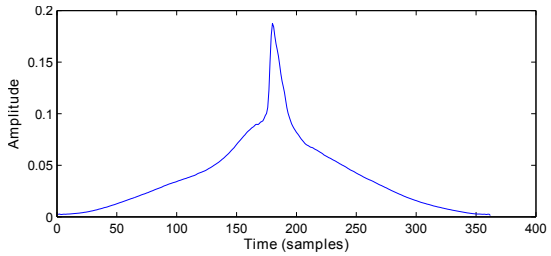

Fig. 3. Speaker-dependent Hilbert envelope for speaker SLT.

The HTS 2.1 HMM architecture [36] was used for training the features using F_0 and spectrum for the alignment. In synthesis, parameters were generated considering global variance [37] except for the spectrum (due to unstability issues with MGCs). Excitation was generated using the vocoder described in Section 2, where the excitation waveform and noise modeling were varied according to the desired setup. Finally, the excitation was filtered with the mel-generalised log spectral approximation (MGLSA) filter [38].

Two databases recorded for the purpose of developing text-to-speech (TTS) synthesis were used to build voices for the experiments. These voices are Scottish English male AWB and US English female SLT from the ARCTIC database [39], which consist of 1,138 and 1,132 sentences, respectively. 1,000 sentences were used for training both voices and the rest was used for testing.

3.2. Subjective evaluations

Subjective evaluation was performed in three separate steps in order to find out the effect of each component and also their possible interactions. The idea was to first select the best noise spectral weighting according to a subjective evaluation. Then, the best spectral weighting method according to the first evaluation is used to study the effect of the noise time envelope. Finally, in the third test, both the best noise spectral weighting and the best time envelope are used in the study of the effect of the periodic waveform.

Comparison category rating (CCR) test was used in order to determine the quality difference between the systems. In CCR test, listeners are presented with speech sample pairs from which listeners rate the difference of the two samples on the comparison mean opinion score (CMOS) scale, which is a seven-point scale ranging from *"much worse"* (-3) to *"much better"* (3). All possible system combinations were evaluated (e.g. for three systems: 1–2, 1–3, 2–3) in both directions (e.g. 1–2 and 2–1). Thus, there were 6 comparisons per sample for the 3-system test and 12 for the 4-system test. The CCR test responses were summarized by calculating the mean scores and 95% confidence intervals for each method. The mean yields the order of preference and distances between all the methods (i.e., the amount of preference relative to each other). A Wilcoxon signed-rank test was finally used for further testing the significance between the means of each method pair. The systems used across the 3 CCR tests are summarized in Table 2 in concordance with the methods explained in Section 2.

All test samples (137 for AWB and 132 for SLT) were synthesized for the three tests using each system (4 + 3 + 3 = 10 systems). Thus, a total of 2,690 (10 $\times$ (137 + 132)) samples were synthesized. In order to reduce the workload on participants, 5 sentences from each speaker were randomly selected for each participant and presented to them in each test. Also ten null pairs (identical sample pair) were included in order to test the consistency of the listeners. Thus each participant rated a total of 130, 70, and 70 stimuli pairs in the first, second, and third test, respectively.

Table 1. Speech features used for training the HMM system.

Feature	Number of parameters
Fundamental frequency	1
Maximum voiced frequency (F_m)	1
Harmonic-to-noise ratio	5
Mel-generalized cepstrum	30

Listening tests were performed in sound proof booths with high-quality headphones. All participants were university students and native speakers of English, and they were paid for the participation. 24, 21, and 24 listeners participated in the three tests, respectively. However, after inspection of the results, some participants were removed due to inconsistent results for the null pairs. Thus, results from 20 listeners in each test were finally used. Note that since all participants are naive, they are known to use the CMOS scale in a smaller range than speech experts would for such a study [40].

3.3. Results

In the first test (CCR1), the perceptual effect of the noise spectral weighting was studied by evaluating the 4 approaches presented in Table 2. The Dirac pulse was used as the periodic waveform in synthesis to emphasize the perceptual effect of the noise models. Constant time envelope was also used. The results are shown in Figure 4 (uppermost graph). Discrepancies are observed across male and female speakers. For male, HNR and DynFm are rated the best, but for the female voice, DynFm is rated better than HNR. FixFm is rated always worse than HNR and DynFm except for the female speaker. The system without any noise (Imp) is always rated the worst. The results indicate that incorporation a noise model in voiced speech has a profound effect on speech quality, and the modeling of the time-varying characteristics of the noise spectrum is beneficial, as is done in DynFm and HNR. Since DynFm was rated better or equal than the rest of the systems, it is used in the rest of the experiments.

In the second test (CCR2), the effect of noise time envelope was studied. The 3 systems considered in CCR2 are depicted in Table 2 (middle part) and the corresponding results are displayed in Fig. 4 (middle graph). The results show no statistically significant differences between the methods. Thus, the results indicate that the noise time envelope has no perceptual relevance, and the simplest one, constant time envelope, is used in the third experiment.

In the third test (CCR3), the effect of periodic waveform was studied by including the 3 systems in Table 2 (bottom part). The corresponding results are shown in Figure 4 (bottom graph). The results diverge across male and female speakers. For male, the natural residual frame and the eigenresidual are rated best while the impulse excitation is rated worse than the natural residual. For the female speaker, impulse excitation and eigenresidual are rated equal while natural residual is rated worse than the two others. These results indicate that the perception of the deterministic waveform depends on the F_0 of the speaker. It is well known that phase information in periodic signals is less important for high-pitched signals than for low-pitched signals (see e.g. [41]). Thus, the natural phase characteristics preserved in the eigenresidual, and even more in the natural residual, are perceived as more natural in the low-pitched male speech. For the high-pitched female speaker, excitation phase characteristics have negligible perceptual effect and thus impulse excitation and eigenresidual are rated equal. It however turns out that the natural frame used for the female speaker exhibited a strong stochastic component, which might explain its degraded quality due to the repetitive structure of its inherent noisy phase.

Table 2. Systems in the three subjective evaluations (CCR1/2/3).

CCR1	Effect of noise spectral weighting
Imp	Impulse excitation without noise
FixFm	Impulse excitation + noise according to fixed F_m
DynFm	Impulse excitation + noise according to dynamic F_m
HNR	Impulse excitation + noise according to HNR
CCR2	Effect of noise time envelope
Con	Imp. exc. + dyn. F_m noise + constant time envelope
Tri	Imp. exc. + dyn. F_m noise + triangular time envelope
DSM	Imp. exc. + dyn. F_m noise + DSM time envelope
CCR3	Effect of deterministic waveform
Imp	Impulse excitation + dyn. F_m noise + const. time env.
Nat	Natural residual + dyn. F_m noise + const. time env.
Eig	Eigenresidual + dyn. F_m noise + const. time env.

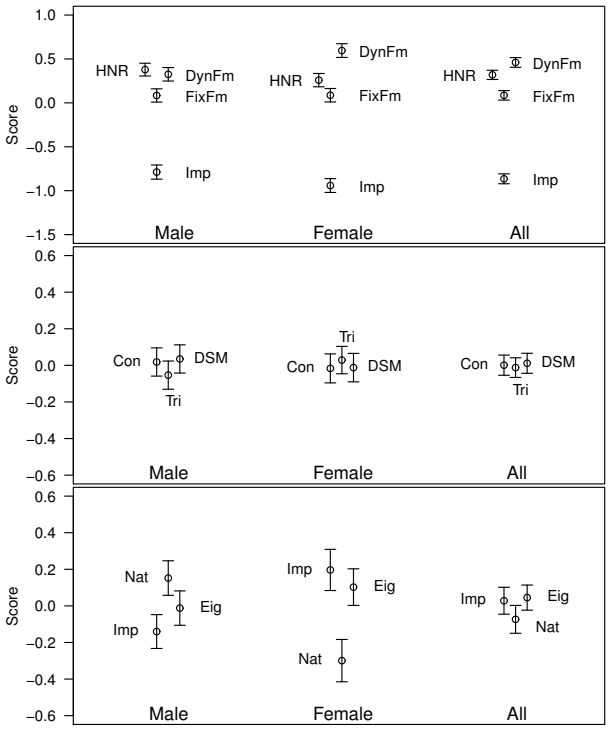

Fig. 4. Results (mean and 95% confidence intervals) of the subjective evaluation comparing noise spectral weighting (uppermost), noise time envelope (middle), and periodic waveform (bottom).

4. CONCLUSIONS

This paper addressed the problem of excitation modeling in order to improve the naturalness in HMM-based speech synthesis. Based on a generalized vocoder, three main factors influencing the quality of synthesis were studied: periodic waveform, noise spectral weighting, and noise time envelope. A subjective evaluation was performed in order to determine the perceptual importance of each factor. Our results clearly indicate that: *i)* incorporating a noise model during the production of voiced sound is crucial; *ii)* the modeling of the time-varying characteristics of the noise spectrum is beneficial. This can be efficiently achieved based on HNR measures or using a dynamic maximum voiced frequency; *iii)* the perceptual impact of the noise envelope seems to be negligible; *iv)* it is necessary to adapt the periodic waveform according the speaker's F_0 range as it will affect the perception of the excitation phase properties. These conclusions should be carefully considered when designing new excitation models. As a result, we believe that future research efforts should focus on new strategies to weight the energy of both periodic and aperiodic components in several spectral bands, as well as on a better understanding of the phase information in the periodic waveform.

5. REFERENCES

[1] Heiga Zen, Keiichi Tokuda, and Alan W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *ICASSP*, 1996, pp. 373–376.

[3] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Speaker interpolation in HMM-based speech synthesis system," in *Eurospeech*, 1997, pp. 2523–2526.

[4] K. Shichiri, A. Sawabe, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Eigenvoices for HMM-based speech synthesis," in *ICSLP*, 2002, pp. 1269–1272.

[5] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 9, pp. 1406–1413, 2007.

[6] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 17, no. 1, pp. 66–83, 2009.

[7] S.-J. Kim, J.-J. Kim, and M.-S. Hahn, "HMM-based Korean speech synthesis system for hand-held devices," *IEEE Trans. Consum. Electron.*, vol. 52, no. 4, pp. 1384–1390, 2006.

[8] A. Gutkin, X. Gonzalvo, S. Breuer, and P. Taylor, "Quantized HMMs for low footprint text-to-speech synthesis," in *Interspeech*, 2010, pp. 837–840.

[9] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "Robust speaker-adaptive HMM-based text-to-speech synthesis," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 17, no. 6, pp. 1208–1230, 2009.

[10] T. Drugman, B. Bozkurt, and T. Dutoit, "Causal-anticausal decomposition of speech using complex cepstrum for glottal source estimation," *Speech Communication*, vol. 53, no. 6, pp. 855–866, 2011.

[11] T. Drugman, B. Bozkurt, and T. Dutoit, "A comparative study of glottal source estimation techniques," *Computer Speech & Language*, vol. 26, no. 1, pp. 20–34, 2012.

[12] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis - A unified approach to speech spectral estimation," *ICSLP*, 1994.

[13] T. Yoshimura, K. Tokuda, T. Masuko, and T. Kitamura, "Mixed-excitation for HMM-based speech synthesis," *Eurospeech*, pp. 2259–2262, 2001.

[14] R. Maia, T. Toda, H. Zen, Y. Nankaku, and K. Tokuda, "An excitation model for HMM-based speech synthesis based on residual modeling," *ISCA SSW6*, 2007.

[15] T. Drugman, G. Wilfart, A. Moinet, and T. Dutoit, "Using a pitch-synchronous residual codebook for hybrid HMM/frame selection speech synthesis," *ICASSP*, pp. 3793–3796, 2009.

[16] T. Drugman, G. Wilfart, and T. Dutoit, "A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis," *Interspeech*, 2009.

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

[18] J. Cabral, S. Renalds, K. Richmond, and J. Yamagishi, "Towards an improved modeling of the glottal source in statistical parametric speech synthesis," in *Sixth ISCA Workshop on Speech Synthesis*, 2007, pp. 113–118.

[19] G. Fant, "The LF-model revisited. Transformations and frequency domain analysis," *STL-QPSR*, vol. 36, no. 2–3, pp. 119–156, 1995.

[20] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "HMM-based Finnish text-to-speech system utilizing glottal inverse filtering," in *Interspeech*, 2008, pp. 1881–1884.

[21] 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. Audio, Speech, Lang. Proc.*, vol. 19, no. 1, pp. 153–165, 2011.

[22] P. Lanchantin, G. Degottex, and X. Rodet, "A HMM-based speech synthesis system using a new glottal source and vocal-tract separation method," *ICASSP*, pp. 4630–4633, 2010.

[23] T. Raitio, A. Suni, H. Pulakka, and M. Vainioand P. Alku, "Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis," *ICASSP*, pp. 4564–4567, 2011.

[24] 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 Commun.*, vol. 27, no. 3–4, pp. 187–207, 1999.

[25] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," in *2nd MAVEBA*, 2001.

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

[27] T. Raitio, J. Kane, T. Drugman, and C. Gobl, "HMM-based synthesis of creaky voice," in *Interspeech*, 2013, pp. 2316–2320.

[28] D. Griffin and J. Lim, "Multiband excitation vocoder," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 36, no. 8, pp. 1223–1235, 1988.

[29] Y. Stylianou, "Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification," *PhD thesis, Ecole Nationale Superieure des Telecommunications*, 1996.

[30] 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.

[31] 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.

[32] J. Cabral and J. Carson-Berndsen, "Towards a better representation of the envelope modulation of aspiration noise," *NOLISP*, pp. 67–74, 2013.

[33] [Online], "Speech signal processing toolkit (SPTK) v. 3.6," 2013, sourceforge.net/projects/sp-tk/.

[34] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis – A unified approach to speech spectral estimation," in *ICSLP*, 1994, pp. 18–22.

[35] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: a quantitative review," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 20, no. 3, pp. 994–1006, 2012.

[36] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Sixth ISCA Workshop on Speech Synthesis*, 2007, pp. 294–299.

[37] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 816–824, 2007.

[38] T. Kobayashi, S. Imai, and T. Fukuda, "Mel generalized log spectrum approximation (MGLSA) filter," *Journal of IEICE*, vol. J68-A, no. 6, pp. 610–611, 1985.

[39] [Online], "CMU ARCTIC," 2013, festvox.org/cmu_arctic/.

[40] T. Drugman, "Advances in glottal analysis and its applications," *PhD thesis, University of Mons*, 2011.

[41] H. Pobloth and W.B. Kleijn, "On phase perception in speech," in *ICASSP*, 1999, vol. 1, pp. 29–32.