COMPARING GLOTTAL-FLOW-EXCITED STATISTICAL PARAMETRIC SPEECH SYNTHESIS METHODS

Tuomo Raitio[1]*, Antti Suni*[2]*, Martti Vainio*[2]*, and Paavo Alku*[1]

[1]Aalto University, Department of Signal Processing and Acoustics, Espoo, Finland
[2]University of Helsinki, Institute of Behavioural Sciences, Helsinki, Finland

ABSTRACT

This paper studies the performance of glottal flow signal based excitation methods in statistical parametric speech synthesis. The current state of the art in excitation modeling is reviewed and three excitation methods are selected for experiments. Two of the methods are based on the principal component analysis (PCA) decomposition of estimated glottal flow pulses. While the first one uses only the mean of the pulses, the second method uses 12 principal components in addition to the mean signal for modeling the glottal flow waveform. The third method utilizes a glottal flow pulse library from which pulses are selected according to target and concatenation costs. Subjective listening tests are carried out to determine the quality and similarity of the synthetic speech of one male and one female speaker. The results show that the PCA-based methods are rated best both in quality and similarity, but adding more components does not yield any improvements.

Index Terms— Statistical parametric speech synthesis, excitation, glottal flow, principal component analysis, pulse library

1. INTRODUCTION

Statistical parametric speech synthesis, also known as HMM-based synthesis [1], is a flexible framework for creating synthetic speech. Despite its several attractive features, hidden Markov model (HMM) based synthesis is known to suffer from poor voice quality in comparison to the best unit selection systems. Recent advances in statistical modeling and vocoding techniques, however, have indicated that adequate quality can also be achieved in HMM-based synthesis [1]. One of the key factors for this progress has been the advances in the excitation modeling methods for the HMM-vocoders.

There are several different approaches for modeling the excitation of a speech signal. The earliest vocoders used only a periodic train of impulses [2] located at glottal closure instants to model the source of voiced speech. The quality of impulse-train-excited speech is poor with a buzzy sound due to unnaturally strong higher harmonics. In addition, excitation features other than the fundamental frequency (F0) and energy cannot be modeled. Many excitation generation methods have been proposed as alternatives to the use of simple impulse trains. Combining the periodic, voiced excitation with additive noise has been used in several more advanced methods, such as in mixed excitation [3] and two-band excitation [4] techniques. In mixed excitation, noise is added to different frequency bands according to weights that define the relative amplitudes of periodic excitation versus aperiodic noise excitation. Mixed excitation is used for example in STRAIGHT [5, 6], which is one of the most widely used vocoders in statistical parametric speech synthesis. In two-band excitation, a maximum voiced frequency is defined above which voiced excitation is composed only of an aperiodic component. Both mixed and two-band excitation have been shown to improve the synthesis quality compared to systems using the traditional impulse train excitation. In another approach, closed-loop training [7, 8], voiced periodic impulse excitation and unvoiced aperiodic noise excitation are fed through state-dependent filters, thus maximizing the likelihood of the excitation signal in comparison to the original one. The synthesis quality is greatly improved compared to a conventional impulse train excitation [7] and was comparable to the quality of a STRAIGHT based method [9]. Also parametric models of the glottal flow have been used in speech synthesis [10, 11] hence allowing some ability for modification of the voice source characteristics. Results obtained indicate that the problem of buzziness can be partly avoided.

The real excitation of voiced speech, the glottal flow, is difficult to represent as a compressed parametric vector. Therefore, vocoding techniques have been proposed that utilize the excitation waveform *per se* rather than its pre-defined compressed representation, hence capturing the detailed characteristics of the signal. The excitation signal to be modeled can be either the glottal flow or the residual computed by linear predictive coding (LPC). The idea of using the natural excitation for improving the synthesis quality is not new (see e.g. [12, 13]), but the development of statistical parametric speech synthesis has given rise to novel excitation methods. In [14, 15], a glottal flow pulse estimated from natural speech with glottal inverse filtering is used for constructing the voiced excitation. The pulse is first interpolating according to F0, aperiodic noise component is added to five separate frequency bands in the frequency domain, and finally the modified pulses are concatenated in order to create a continuous excitation. The synthesis quality was shown in [15] to outperform STRAIGHT with a low-pitched male voice, and to be equal to or better than STRAIGHT in another experiment [16] with one male and one female voice. In [17, 18, 19], principal component analysis (PCA) is applied to pitch-synchronous residual signal in order to model the waveform with eigen-residuals. The method in [17] was shown to outperform a simple excitation, the method in [18] was rated better than a simple excitation and two-band excitation [4], and [19] was rated comparable to the quality of STRAIGHT. In [20], a pitch-synchronous residual codebook is constructed and residual frames are selected for synthesizing the excitation. The resulting quality was shown to outperform a simple impulse train excitation approach. In [21], a library of various estimated glottal flow pulses is constructed and selected for the synthesis of the excitation according to a target cost of voice source features and a concatenation cost between adjacent pulses. In [21], the pulse library method was shown to be equal in quality to the method in [15] but with slightly

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 287678, Academy of Finland (projects 135003 LASTU programme, 1128204, 1218259, 121252), and MIDE UI-ART.

better speaker similarity. In [22, 23], a pulse library technique was shown to perform comparably to STRAIGHT-based techniques.

The goal of this study is to compare the most potential recently proposed excitation generation techniques in statistical speech synthesis. The comparison involves methods that have been used either for the residual or the glottal flow. The common factor is the use of the natural excitation signal *per se*. All the methods involved have shown potential in producing high-quality synthetic speech but they have not been compared simultaneously until now.

2. VOCODERS IN COMPARISON

When comparing statistical parametric speech synthesis techniques, there is always the problem of how to make the comparison fair. The synthesis systems should be different only in terms of the technique to be tested. This is rarely possible and usually techniques are compared using more or less different system architectures. This may lead to various types of differences in the synthesized speech; prosodic differences are mixed with differences in segmental speech quality, and thus the rating becomes ambiguous. In this study, only a single configuration of a statistical speech synthesis system is trained per speaker in order to avoid the aforementioned problems. All parameters required to synthesize speech with the techniques to be tested are included in a single system.

Three different excitation techniques are experimented with. Two of the methods, related to [17, 18, 19], are based on the PCA decomposition of estimated glottal flow pulses. While the first one uses only the mean of the pulses, the second method uses 12 components in addition to the mean signal for modeling the glottal flow waveform. The third method [21] utilizes a library of various glottal flow pulses, which are selected from the library according to target and concatenation costs. The STRAIGHT vocoder is not involved in the present comparison for several reasons. First, it is not possible to integrate STRAIGHT into the other systems due to many differences in speech parametrization and synthesis techniques. Secondly, if STRAIGHT were used as a reference, the differences in prosody would certainly affect the results and thus the assessment would not measure purely the differences between the excitation generation techniques. Thirdly, the quality of STRAIGHT compared to the reference methods is already documented [15, 19]. The implementation of the single pulse GlottHMM method [15] is not included in the test since it is very similar to the first technique. All the three methods are described in more detail in Section 2.4.

Since two of the methods to be tested (similar to [19]) originally utilized the LPC residual as an excitation while the glottal flow is used in [21], it is not possible to integrate these methods as such into an individual system. Thus, all waveform modeling is performed in the glottal flow waveform domain, as is done in the GlottHMM vocoder [15, 21], which is used throughout this study. The implication of this choice is that the excitation signal to be modeled exhibits more spectral variation compared to whitened LPC residual. This also produces more natural variation to the excitation waveform that is to be modeled. Also the scheme of adding aperiodic noise is adopted from the GlottHMM vocoder. In the next section, GlottHMM is described in order to understand the methods used in this study and to depict the similarities and differences to the vocoder implementations used in other studies.

2.1. Speech Parametrization with GlottHMM

The parametrization of speech with the GlottHMM vocoder, used for all the methods, is illustrated in Figure 1. Speech signal is first

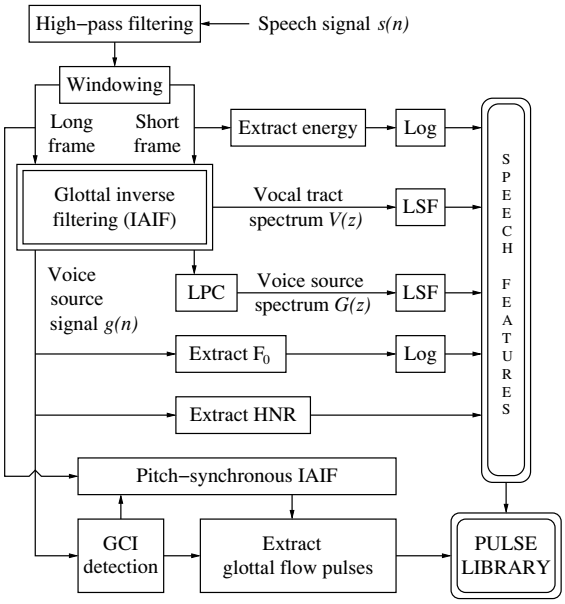

Fig. 1. Illustration of the parametrization of speech with the GlottHMM vocoder.

high-pass filtered with a cut-off frequency of 70 Hz in order to remove possible low frequency fluctuations. Signal is then analyzed in two types of windows. A 25-ms long frame is used to extract energy and spectra of the vocal tract and the voice source. A longer frame, whose length depends on the F0 range of the speaker, is used to extract voice source features that require several full pitch periods for successful analysis. Both frames are processed with iterative adaptive inverse filtering (IAIF) [24, 25], a glottal inverse filtering method that separates the speech signal into the estimated glottal flow signal and the vocal tract filter. LPC is used inside IAIF for estimation of the spectrum. The modified version of the IAIF method [22] is used in order to make the glottal flow signal estimation more robust. The use of IAIF in statistical parametric speech synthesis is described in detail in [15]. The estimated vocal tract and voice source spectra are both parametrized with line spectral frequencies (LSFs), parametric representation of LPC information well-suited for statistical parametric speech synthesis [26], providing stability and low spectral distortion. The longer frame, representing the voice source signal after IAIF, is used to extract the F0 of speech using the autocorrelation method. The relative amplitudes of the periodic vibratory glottal excitation and the aperiodic noise component of the excitation is represented by the harmonic-to-noise ratio (HNR), indicating the degree of voicing. HNR is based on the ratio between the upper and lower smoothed spectral envelopes (defined by the harmonic peaks and interharmonic valleys, respectively) and averaged across five frequency bands according to the equivalent rectangular bandwidth (ERB) scale.

The extraction of individual glottal flow pulses from the estimated glottal flow signal is of special interest in this study. First, the glottal closure instants (GCIs) are detected by a simple peak picking algorithm that searches for the negative excitation peaks of the glottal flow derivative approximately at fundamental period intervals (T0). Although more sophisticated GCI detection methods have been developed (for a review, see e.g. [27]), the use of the simple method does not incur problems in this application since only pulses

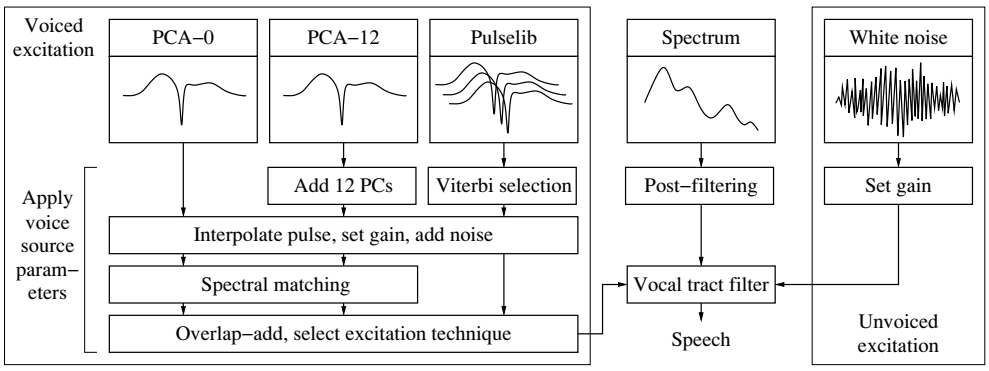

Fig. 2. Illustration of the vocoder and different excitation generation techniques.

that match with T0 are accepted and thus the errors due to GCI detection are minimized. For all the accepted two-pitch period speech segments, the modified IAIF algorithm is applied pitch-synchronously again in order to yield a better estimate of the glottal flow. The re-estimated two-period glottal flow derivative waveforms are finally windowed with the Hann window and linked with the corresponding voice source parameters extracted by the vocoder.

2.2. PCA Decomposition of the Pulse Library

Principal component analysis (PCA) is a mathematical procedure that uses an orthogonal transformation to convert a set of observations into a set of linearly uncorrelated variables called principal components (PCs). The use of PCA for voice source waveform modeling was first proposed in [28] and a speech analysis/synthesis scheme was elaborated in [29, 30]. It has also been used successfully in statistical parametric speech synthesis [17, 19].

In this work, the extracted two-period glottal flow derivative waveforms of the pulse library are first interpolated to a constant length (25 ms) and normalized in energy. The pulses are then normalized by evaluating and subtracting the mean glottal flow pulse, after which PCA of order 12 is applied. The results of the analysis thus consists of the mean glottal flow signal, the 12 PCs, and the 12 PC weights for the library pulses. The PCA decomposition of the pulse library is performed separately from the analysis of the whole speech corpus. Thus, the size of the pulse library can be kept relatively small with the assumption that the pulse library is a representative set from the corpus.

After applying PCA to the pulse library, the actual speech analysis and PCA decomposition takes place, where the extracted pulses from each frame are converted to corresponding PC weights. This process consists of the same normalization scheme as in the case of the pulse library, i.e., length, energy, and mean normalization, after which PC weights are calculated according to the PC vectors.

Table 1. Trained (T) and static (S) speech features.

Type	Feature	No. of parameters
T	Vocal tract spectrum	24/30 (female/male)
T	Energy	1
T	Fundamental frequency	1
T	Harmonic-to-noise ratio	5
T	Voice source spectrum	5/10 (female/male)
T	Principal component weights	12
S	Mean pulse	1 vector
S	Principal components	12 vectors
S	Pulse library	∼7500 pulses + params

2.3. Training of the Synthesizer

The resulting speech features, including the 12 PC weights, are trained within the HTS speech synthesis framework [31, 32]. Both F0 and PC weights are trained within multi-space distribution (MSD) streams [33], while other speech features are trained in continuous streams. All speech features extracted for the current study are depicted in Table 1.

2.4. Speech Synthesis with the Excitation Techniques

There are three excitation generation techniques that are experimented in this study:

1. *PCA-0:* Mean pulse + spectral matching
2. *PCA-12:* Mean pulse + 12 PCs + spectral matching
3. *Pulselib:* Pulse library

The techniques are illustrated in Figure 2. *PCA-0* is related to the method in [19], using only the mean of the pulse library (actually first eigen-residual is used in [19] which is, in principle, slightly different). However, there are two important differences. First, the variation of the excitation is largely captured by the LPC spectrum in [19], thus making the spectral envelope of the LPC residual rather constant. In this study, glottal flow signal is used instead of LPC residual, which gives the excitation more room for natural variation. Thus, to allow the single mean glottal flow pulse to vary, a spectral matching scheme [15] is used, i.e., the excitation signal is filtered with an infinite impulse response (IIR) filter that flattens the pulse and applies the modeled voice source spectrum.

PCA-12 is otherwise identical to the *PCA-0* technique but the glottal flow pulse is allowed for more variation by adding the 12 principal components according to the modeled PC weights. Spectral matching is still applied in order to normalize the overall spectral tilt, as is done in *PCA-0*.

In *Pulselib* [21], pulses are selected from the library according to manually tuned target cost consisting of the voice source features (F0, energy, voice source spectrum, HNR, and PC weights) and the concatenation cost consisting of the root mean square (RMS) error between adjacent downsampled pulse candidates. The selection process is optimized with the Viterbi search. Spectral matching is not

required since it is assumed that the selection process will automatically select pulses with the desired spectral tilt.

In all techniques, the degree of voicing is controlled identically. The amount of noise in the voiced excitation is matched by manipulating the phase and magnitude of the spectrum of each pulse according to HNR at each ERB band. Finally, the pulses are overlap-added to create a continuous excitation signal, which is filtered with the formant enhanced (post-filtering) vocal tract filter to create speech.

3. EXPERIMENTS

In order to compare the methods, both quality and similarity of the speech samples generated by the techniques were assessed in subjective listening tests. Tests were performed both with male and female speakers. The male database (mv) consists of 600 phonetically rich utterances spoken by a low-pitched Finnish male [34]. The female database (heini) consists of 513 phonetically rich utterances spoken by a young Finnish female.

First 20 sentences of the databases were used to build the pulse library for each speaker. The pulse libraries consisted of 7528 and 7500 pulses for the male and female speaker, respectively. PCA was applied to the pulse libraries in order to get the mean pulses, PCs and PC weights.

The synthesis times between the *PCA-0* and *PCA-12* methods were similar, but the synthesis time for the *Pulselib* method was almost double with the male speaker and over ten times longer with the female compared to PCA-based methods. The significant difference between the male and female speaker with *Pulselib* is due to the F0 of the speaker; the voiced sections in female speech consists of a large number of pulses, and thus the computational cost of the pulse selection algorithm increases exponentially.

For measuring the quality of speech, a comparison category rating (CCR) test was used. In CCR test, subjects are presented with a speech sample pairs and the task of the listener is to rate the quality difference between the samples on the comparison mean opinion score (CMOS) scale, which is a discrete seven-point scale ranging from much worse (-3) to much better (3). Ten sentences per speaker were synthesized with each method for the test and a total of 60 comparisons were performed per speaker per test subject. The responses of the CCR test were summarized by calculating the mean score for each method with 95% confidence intervals, which yields the order of preference and distances between the methods.

For measuring the similarity, a forced choice test is used in which the listener is presented with four reference samples of natural speech and one synthetic sample per each method. The task of the listener is to choose the method that is most similar to the speaker in the reference samples. Ten sentences were synthesized per speaker with each method and used in the similarity test. The responses are summarized by evaluating the percentage of choices for each method with 95% confidence intervals.

All tests were performed in quiet listening booths with high-quality headphones. A total of ten listeners participated in the test, thus yielding a total of 600 and 100 data points per speaker for the quality and similarity test, respectively.

The results of the quality test are shown in Figure 3. The results indicate that the glottal flow mean based excitation techniques *PCA-0* and *PCA-12* show no statistically significant differences neither for male nor female speaker. The pulse library method is rated worse than the other two methods. The results of the similarity test are shown in Figure 4. Methods *PCA-0* and *PCA-12* show no statistically significant difference in similarity, while the pulse library method is rated slightly less similar.

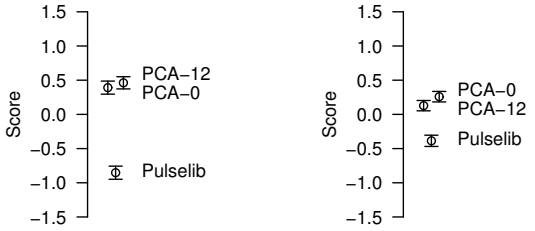

Fig. 3. Quality results for the male (left) and female (right) voices.

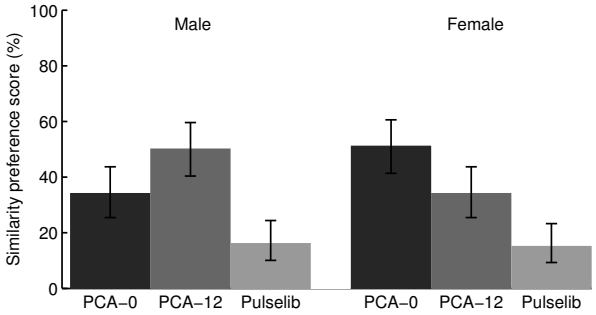

Fig. 4. Similarity results for the male (left) and female (right) voices.

4. DISCUSSION AND CONCLUSIONS

Firstly, this study shows that PCA-based excitation technique can be successfully used for the estimated glottal flow signal, while previously PCA has been only used for the LPC residual signal [17, 18, 19, 20, 28, 29, 30]. Secondly, the study shows that using the PCs in addition to the mean pulse does not increase the quality of synthetic speech, corroborating the results obtained in [19]. Although the use of PCs occasionally makes speech more vivid, especially in the extreme modes of glottal flow, such as in very breathy or tense voice, the differences were small and infrequent, thus making the consistent assessment between the two methods difficult for the listeners. Thirdly, the results show that the pulse library method is currently not robust enough to yield quality comparable to the PCA-based excitation techniques. Although some segments sound very close to original speech, occasional artefacts, reported as "reverberant" or "chorus" type effects in voiced speech, deteriorate the overall quality. This indicates that the smoothness (or regularity) of the resulting speech is of primary importance for the listeners. Thus, attempts on realistic modeling of the contextual variation in excitation are most likely to fail if this prerequisite is compromised. In addition, the complex unit-selection type optimization of the pulse library technique makes the voice building more difficult and unpredictable. For example, the male voice created with the pulse library method was assessed in [21] to be equal to or better than the original GlottHMM [15] technique with apparently more successful tuning of the target and concatenation weights. Nevertheless, the differences between all the systems are rather small due to the identical base system. It is also worthwhile to note that the *PCA-0* technique is very similar to the original GlottHMM implementation [14, 15], in which a single estimated glottal flow pulse is modified to create excitation. The new mean-based excitation scheme eliminates the need for the manual selection of the glottal flow pulse and ensures that the pulse does not have a significant existing noise component.

5. REFERENCES

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

[2] T. E. Tremain, "The government standard linear predictive coding algorithm: LPC-10," *Speech Technology*, vol. 1, pp. 40–49, Apr. 1982.

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

[4] S. J. Kim and M. Hahn, "Two-band excitation for HMM-based speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, 2007.

[5] 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, Apr. 1999.

[6] H. Kawahara, Jo 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 International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, Sep. 2001.

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

[8] R. Maia, H. Zen, and M. J. F. Gales, "Statistical parametric speech synthesis with joint estimation of acoustic and excitation model parameters," in *SSW7*, Sep. 2010, pp. 88–93.

[9] H. Zen and T. Toda, "An overview of nitech HMM-based speech synthesis system for Blizzard Challenge 2005," in *The Blizzard Challenge 2005 workshop*, 2005, http://festvox.org/blizzard.

[10] J. Cabral, S. Renalds, K. Richmond, and J. Yamagishi, "Towards an improved modeling of the glottal source in statistical parametric speech synthesis," in *SSW6*, 2007, pp. 113–118.

[11] J. Cabral, S. Renalds, K. Richmond, and J. Yamagishi, "Glottal spectral separation for parametric speech synthesis," in *Proc. Interspeech*, 2008, pp. 1829–1832.

[12] J. Holmes, "The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer," *IEEE Trans. Audio and Electroacoustics*, vol. 21, no. 3, pp. 298–305, Jun. 1973.

[13] K. Matsui, S. D. Pearson, K. Hata, and T. Kamai, "Improving naturalness in text-to-speech synthesis using natural glottal source," in *Proc. ICASSP*, Apr. 1991, vol. 2, pp. 769–772.

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

[15] 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 Lang. Proc.*, vol. 19, no. 1, pp. 153–165, Jan. 2011.

[16] A. Suni, T. Raitio, M. Vainio, and P. Alku, "The GlottHMM speech synthesis entry for Blizzard Challenge 2010," in *The Blizzard Challenge 2010 workshop*, 2010, http://festvox.org/blizzard.

[17] T. Drugman, G. Wilfart, and T. Dutoit, "A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis," in *Proc. Interspeech*, 2009, pp. 1779–1782.

[18] J. Sung, D. Hong, K. Oh, and N. Kim, "Excitation modeling based on waveform interpolation for HMM-based speech synthesis," in *Proc. Interspeech*, 2010, pp. 813–816.

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

[20] T. Drugman, G. Wilfart, A. Moinet, and T. Dutoit, "Using a pitch-synchronous residual codebook for hybrid HMM/frame selection speech synthesis," in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, Apr. 2009, pp. 3793–3796.

[21] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis," in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, 2011, pp. 4564–4567.

[22] A. Suni, T. Raitio, M. Vainio, and P. Alku, "The GlottHMM entry for blizzard challenge 2011: Utilizing source unit selection in hmm-based speech synthesis for improved excitation generation," in *The Blizzard Challenge 2011 workshop*, 2011, http://festvox.org/blizzard.

[23] A. Suni, T. Raitio, M. Vainio, and P. Alku, "The GlottHMM entry for blizzard challenge 2012 - hybrid approach," in *The Blizzard Challenge 2012 workshop*, 2011, http://festvox.org/blizzard.

[24] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Commun.*, vol. 11, no. 2-3, pp. 109–118, 1992.

[25] P. Alku, H. Tiitinen, and R. Näätänen, "A method for generating natural-sounding speech stimuli for cognitive brain research," *Clinical Neurophysiology*, vol. 110, pp. 1329–1333, 1999.

[26] M. Marume, H. Zen, Y. Nankaku, K. Tokuda, and T. Kitamura, "An investigation of spectral parameters for HMM-based speech synthesis," in *Proc. Autumn Meeting of Acoust. Soc. of Japan*, Sep. 2006, (In Japanese).

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

[28] M. Thomas, J. Gudnason, and P. Naylor, "Data-driven voice source waveform modelling," in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, April 2009, pp. 3965 –3968.

[29] J. Gudnason, M. Thomas, P. Naylor, and D. Ellis, "Voice source waveform analysis and synthesis using principal component analysis and Gaussian mixture modelling," in *Proc. Interspeech*, 2009, pp. 108–111.

[30] J. Gudnason, M. Thomas, D.P.W. Ellis, and P.A. Naylor, "Data-driven voice source waveform analysis and synthesis," *Speech Commun.*, vol. 54, no. 2, pp. 199–211, Feb. 2012.

[31] 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 *SSW6*, Aug. 2007, pp. 294–299.

[32] HTS, "HMM-based speech synthesis system," Nov. 2012, http://hts.sp.nitech.ac.jp.

[33] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Trans. Inf. & Syst.*, vol. E85-D, no. 3, pp. 1455–464, 2002.

[34] M. Vainio, *Artificial Neural Network Based Prosody Models for Finnish Text-to-Speech Synthesis*, Ph.D. thesis, University of Helsinki, Finland, Dec. 2001.