The GlottHMM Entry for Blizzard Challenge 2011: Utilizing Source Unit Selection in HMM-Based Speech Synthesis for Improved Excitation Generation

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

[1]Department of Speech Sciences, University of Helsinki, Helsinki, Finland
[2]Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland
antti.suni@helsinki.fi, tuomo.raitio@aalto.fi

Abstract

This paper describes the GlottHMM speech synthesis system for Blizzard Challenge 2011. GlottHMM is a hidden Markov model (HMM) based speech synthesis system that utilizes glottal inverse filtering for separating the vocal tract and the glottal source from speech signal and models both components individually. In this year's entry, stabilized weighted linear prediction (SWLP) is used to yield more robust estimates of the vocal tract filter of the high-pitched female voice. After the inverse filtering, the resulting source signal is parameterized into excitation features and a glottal flow pulse library, consisting of the variety of different glottal flow pulses. In the synthesis stage, a unit selection scheme is used for reconstructing the source signal: by minimizing the target and concatenation costs, best matching glottal flow pulses are selected from the pulse library in order to create a natural voice source. Finally, speech is synthesized by filtering the excitation signal by the vocal tract filter.

Index Terms: speech synthesis, hidden Markov model, glottal inverse filtering, glottal flow pulse library, unit selection

1. Introduction

GlottHMM text-to-speech (TTS) system [1, 2] is developed in a collaboration between Aalto University and University of Helsinki. In this entry, we have used our speech synthesis system that emphasizes the importance of the speech production mechanism, especially in terms of separating the two distinct parts of it: the glottal excitation and the vocal tract filter.

This year's challenge was reduced in scale, consisting of building only one voice from a large database of American English female speech, designed especially for concatenative synthesis. Although our parametric system was not likely to be very competitive in this kind of task, we decided to participate in order to test and report some new ideas, related to vocoder and HMM modeling. Specifically, we wanted to get listener feedback on the use of a glottal pulse library for generating the excitation signal.

Our TTS system was elaborated with a unit selection type of voice source reconstruction: a glottal flow pulse library is constructed from the speech corpus, and in synthesis stage, best matching pulses are selected in order to create a natural voice source. The glottal inverse filtering method is also refined; stabilized weighted linear prediction (SWLP) is used as a spectral modeling tool in order to yield more robust spectral estimates for the vocal tract filter. SWLP is especially effective for high-pitch voices in which prominent harmonic peaks may bias formant estimates computed by conventional spectral modeling methods such as LPC.

We will first describe our synthesis system, emphasizing the spectral modeling and the use of glottal flow pulse library. This is followed by discussion on voice building, analysis on the results, and conclusions.

2. Overview of the system

Statistical parametric speech synthesis has recently become very popular due to its flexibility. However, the speech quality and naturalness of parametric speech synthesizers are usually inferior compared to state-of-the-art unit selection speech synthesis systems. This degradation is mainly caused by the oversimplified vocoder techniques and over-smoothing of the generated speech parameters [3]. Our GlottHMM text-to-speech (TTS) system tries to overcome especially these problems.

One of the main problems in simplified vocoder techniques is the modeling of the voice source. Recently, the modeling of the voice source has been under intensive research, and several techniques have been proposed to model the source signal [4, 5, 6, 7]. However, the accurate modeling of the glottal source signal has proven to be very difficult. Thus, the use of glottal flow models has been replaced in several studies by the utilization of the estimated glottal source waveform *per se* [8, 9, 10].

In our recent approach, a glottal source pulse is computed from real speech and modified for generating the excitation signal. This has resulted in speech quality that is much better than that of conventional methods [2, 11]. However, a single pulse is unable to cover the wide variety of different voice characteristics. Thus, we have extended the use of a single glottal source pulse to the use of a library of various pulses [12]. This unit selection type of source modeling technique enables the reconstruction of a more natural voice source.

We also use slightly different inverse filtering [13] and spectral modeling approach. Previously, we have used the iterative adaptive inverse filtering (IAIF) method [14, 15] for estimating the vocal tract transfer function from the speech signal. In this work, we have modified the IAIF method to become more robust by reducing the estimation steps. We also use stabilized weighted linear prediction (SWLP) for estimating the vocal tract filter. SWLP applies more weight on the closed phase of the glottis, where the vocal tract filter is more prominent. This reduces the biasing effects of the harmonic peaks on spectral models of the vocal tract.

The overview of the system is shown in Figure 1. In the training stage, we first decompose the speech signal into the glottal source signal and the model of the vocal tract filter using glottal inverse filtering. Then we extract pulses from each

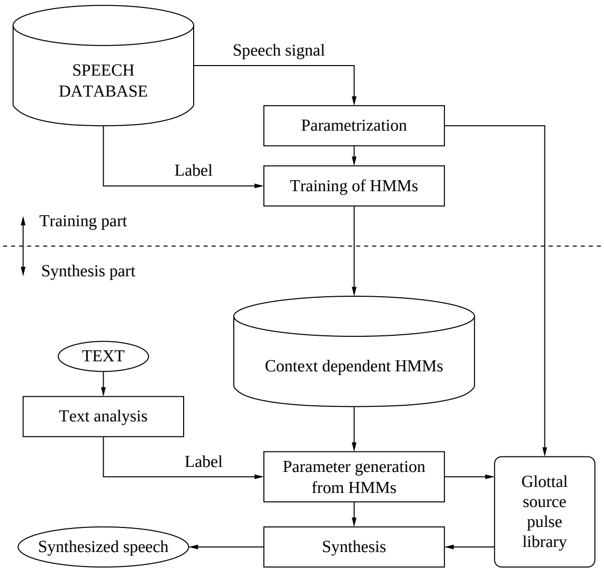

Figure 1: *Overview of the TTS system.*

analysis frame and map these pulses according to excitation parameters. After the analysis stage, the spectral and excitation parameters are trained in the framework of HMMs. In the synthesis stage, the source signal is generated by selecting appropriate pulses from the library according to excitation parameters. Finally, the vocal tract filter is used to filter the excitation to generate speech.

3. Vocoder architecture

The GlottHMM speech synthesis system is built on a basic framework of an HMM-based speech synthesis system [16], but the parametrization and synthesis methods differ from conventional vocoders and are therefore explained in detail below.

3.1. Speech parametrization

The flow chart of the speech parametrization algorithm is shown in Figure 2. First, the signal is windowed with a rectangular window to two types of frames at 5-ms intervals: a 25-ms frame for extracting speech spectrum and energy and a 44-ms frame for extracting the voice source parameters and the glottal source pulses. Additionally, for unvoiced segments, a shorter frame (12.5 ms) is used in order to better capture the transients and noise bursts. The speech features are presented in Table 1.

The log-energy of the windowed speech signal is evaluated first, after which glottal inverse filtering is performed in order to estimate the glottal volume velocity waveform from the speech signal. The inverse filtering method cancels the effects of the vocal tract and the lip radiation from the speech signal. A modified version of the automatic glottal inverse filtering method, iterative adaptive inverse filtering (IAIF) [14, 15], is utilized. While the original IAIF method yields accurate estimates of the voice source signal at its best, in adverse conditions the estimates may vary significantly from frame to frame. In order to prevent such behavior, we have reduced the number of estimation steps in the modified IAIF method from two to one. Thus, the modified IAIF method yields more robust estimates of the glottal flow, although the estimates may not be as detailed as with the original IAIF method. The modified IAIF method is illustrated in Figure 3.

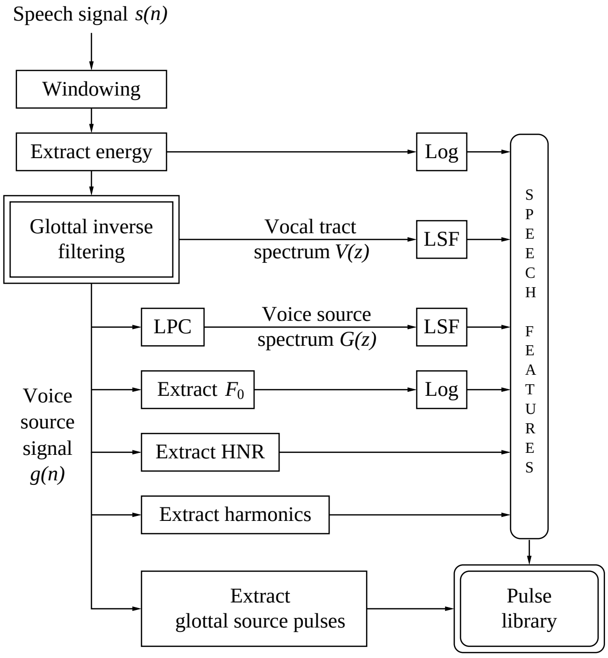

Figure 2: *Illustration of the parametrization stage. The speech signal $s(n)$ is decomposed into the glottal source signal $g(n)$ and the all-pole model of the vocal tract $V(z)$ using the modified IAIF method. The glottal source signal is further parametrized into the all-pole model of the voice source $G(z)$, the fundamental frequency F_0, the harmonic-to-noise ratio (HNR), and the differences of the first ten harmonic magnitudes. A glottal source pulse library is constructed from the extracted glottal flow pulses and the corresponding voice source parameters.*

In addition, stabilized weighted linear prediction (SWLP) [17] is used for spectral modeling in the modified IAIF method. SWLP was developed from weighted linear prediction (WLP) [18], but, differently from WLP, the filter stability is always guaranteed in SWLP, hence making its use justified in applications where all-pole synthesis is needed. In SWLP analysis, the autocorrelation is weighted by the short time energy window of the signal, thus emphasizing high energy parts. SWLP has two benefits compared to conventional linear prediction (LP) analysis. First, SWLP spectrum is less distracted by the harmonics of the excitation signal since the high energy parts are located in the glottal closed phase instants, thus giving less weight to the excitation instants. For the same reason, the inverse filtering is more accurate as the excitation is given less weight when determining the vocal tract spectrum. Thus, the spectral tilt of the excitation has less effects on the vocal tract spectrum, and the separation between the vocal tract spectrum and the voice source is more accurate.

The outputs of the modified IAIF algorithm are the estimated glottal flow signal and the all-pole model of the vocal tract. In order to capture the variations in the glottal flow due to different phonation or speaking style, the spectral envelope of the excitation signal is further parametrized with conventional linear predictive coding (LPC). This spectral model of the glottal excitation captures mainly the spectral tilt, but also the more detailed spectral structure of the source.

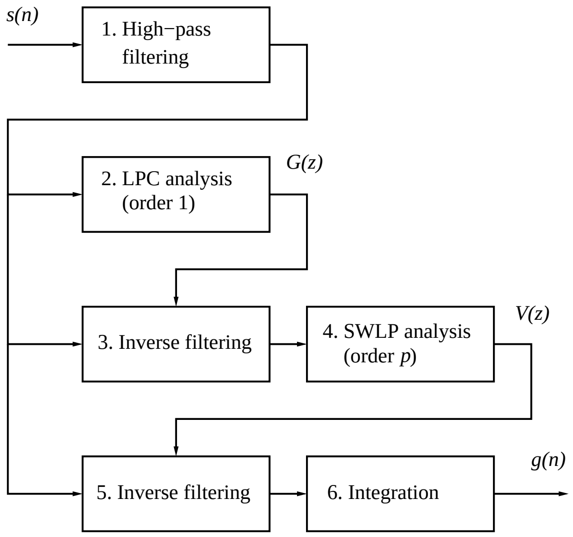

Figure 3: *Block diagram of the modified IAIF method.*

The fundamental frequency is estimated from the glottal flow signal with the autocorrelation method. In order to evaluate the degree of voicing in the glottal flow signal, a harmonic-to-noise ratio (HNR) is determined 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 [19]. In addition, the magnitude difference of the first ten harmonic peaks compared to the first harmonic magnitude of the excitation spectrum is parametrized to describe the low-frequency source spectrum more accurately.

LPC models of the vocal tract and the voice source are further converted to line spectral frequencies (LSFs) [20], which provides stability [20] and low spectral distortion [21]. In case of unvoiced speech, conventional LPC is used to evaluate the spectral model of speech. In order to preemptively alleviate for the over-smoothing of the vocal tract parameters in HMM training, a formant enhancement technique [22] is used in the parametrization stage instead of post-filtering after the parameter generation.

For constructing a glottal source pulse library, pulses are extracted from the differentiated glottal volume velocity signal. First, glottal closure instants (GCIs) are determined by searching for the minima of the glottal source signal at fundamental period intervals. This simple GCI detection method, when applied to the glottal inverse filtered signal, works sufficiently well for the purpose. After the GCI detection, each complete two-period glottal source segment is extracted and windowed with the Hann window. The energy of each pulse is normalized and the pulses are stored to the pulse library. All the voice source parameters (all parameters in Table 1 except the vocal tract spectrum) are also stored to the library in order to describe the characteristics of each pulse. In addition, a down-sampled constant length (10 ms) version of each pulse is stored to enable the evaluation of the concatenation cost in the synthesis stage.

Table 1: Speech features and the number of parameters.

Feature	Parameters per frame
Fundamental frequency	1
Energy	1
Harmonic-to-noise ratio	5
Harmonic magnitudes	10
Voice source spectr. (filter ord.)	7
Vocal tract spectr. (filter ord.)	25

3.2. Synthesis

The flow chart of the synthesis stage is shown in Figure 4. The excitation signal consists of voiced and unvoiced sound sources. The voiced excitation is constructed by utilizing a unit selection scheme for the source signal: appropriate glottal flow pulses are selected from the glottal flow pulse library in order to generate a natural voice source signal. The pulses are selected by minimizing the joint cost, consisting of target and concatenation costs. The target cost is composed of the root mean square (RMS) error between the voice source parameters of the pulse and the ones generated by the HMMs. Individual weights for each voice source parameter are experimentally set. The target cost assures that an appropriate pulse is selected with desired voice source characteristics. The concatenation cost is composed of the RMS error between the down-sampled pulse waveforms of the consecutive pulses in each full voiced section. Minimizing the concatenation cost ensures that the adjacent pulse waveforms do not differ substantially from each other, possibly producing abrupt changes in the excitation signal leading to a harsh voice quality. The best matching pulses, in terms of target and concatenation costs are selected for each voiced section at a time, and the process is optimized with the Viterbi search among all pulses. Individual weights for the target and concatenation costs are tuned by hand.

After selecting the pulses for a voiced sections, the pulses are scaled in amplitude according to the energy measure given by the HMMs. Then, the pulses are overlap-added according to F_0 values in order to create a continuous voiced excitation. Since the fundamental frequency is included in the target cost, pulses with approximately correct fundamental period will be chosen, and no further processing of the pulses is necessary.

The unvoiced excitation is composed of white noise, whose gain is determined according to the energy measure generated by the HMMs. The voiced and unvoiced excitations are then combined and filtered with the vocal tract filter for generating speech.

4. Voice building

4.1. Front end

Perhaps the most interesting aspect of this year's challenge was the unconventional labeling provided with the speech data. The annotation consisted of so called lessemes [23], phonemes augmented with detailed information about speech melody and other phonetic details. As the voice talent was familiar with this notation and the text was annotated prior to reading, the accuracy of F_0 movement labeling was high above normal TTS level.

While the authors of the notation had used lessemes as atomic units in TTS, it seemed sensible to break the features apart for use in a conventional context-dependent label format. In addition to lesseme features, typical positional and quantitative features were extracted, as well as unigram probabilities of

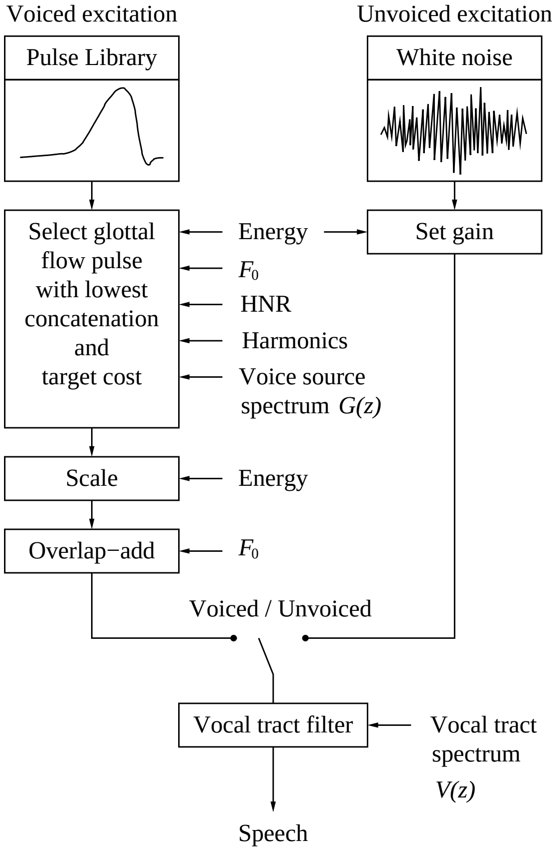

Figure 4: *Illustration of the synthesis stage. The voiced sound source is composed of glottal source pulses selected from the pulse library. Unvoiced excitation is composed of white noise. The excitation signals are combined and filtered with the vocal tract filter $V(z)$ to generate speech.*

the words to help with rhythm and phrasing.

4.2. Feature extraction

Parameters described in Table 1 were extracted along with their delta and delta-delta features. Additionally, a pulse library of approximately 15000 pulses was constructed from 20 selected utterances with rich F_0 movement and phonetic content. Examples of the pulse waveforms are shown in Figure 5.

4.3. HMM training

Due to some failed experiments and time constraints, only 3000 randomly selected sentences were used in training the final voice. The models, consisting of seven independent streams, were trained with the standard HTS 2.1 recipe [16] except for changes described below:

4.3.1. Explicit voicing

Having multiple independent streams provides for efficient clustering but introduces problems due to lack of coherence between streams, resulting in fuzzy voicing boundaries and artefacts noted in the previous challenge [11]. We tried to alleviate this problem by introducing explicit state-wise voicing information to contextual labels, to be used in the final clustering step for all streams except F_0. By changing the question “Is the current phoneme voiced?” to “Is the current state voiced in the training data?” we hoped to achieve crisper voicing boundaries and less audible artefacts in the final voice. In parameter generation, F_0 prediction is first performed normally, and the predicted voicing is considered for other streams.

4.3.2. Reducing over-smoothing by extrapolation

It is well known that the effect of dynamic features in parameter generation is not considered in ML-based HMM training, causing over-smoothing in generated parameter trajectories. This problem has been largely solved by introducing minimum generation error (MGE) criterion [24] to HMM training. However the MGE training method is computationally intensive and, importantly for many, is not included in the public HTS framework.

In this year’s challenge, we experimented with MGE inspired method for trajectory sharpening with the available tools. In this method, first, an estimate of the magnitude and direction of over-smoothing for each model is achieved by training an over-smoothed model set with generated parameters and using the difference between the original and the over-smoothed model set to apply a proper amount of sharpening for each model.

The process involves alignment of the training data, generating the training data with the original state alignments, and re-estimation of the original models with the generated parameters. Then, at synthesis stage, the model interpolation framework in HTS_engine is applied with over-smoothed and original models as reference points, to extrapolate away from the over-smoothed models. In the current voice, extrapolation ratios for each parameter type were tuned by hand. Informally, this method seems to provide more detailed trajectories and generally better speech quality than parameter generation considering global variance (GV), but like GV, is subject to artefacts if applied too strongly.

4.4. The resulting voice

The submitted voice was informally assessed and found generally crisp and smooth but somewhat inconsistent, with some utterances containing unit selection type artefacts and hoarseness, due to pulse selection errors. Also, unexplained low frequency clicks occurred on some contexts which could not be fixed before the deadline.

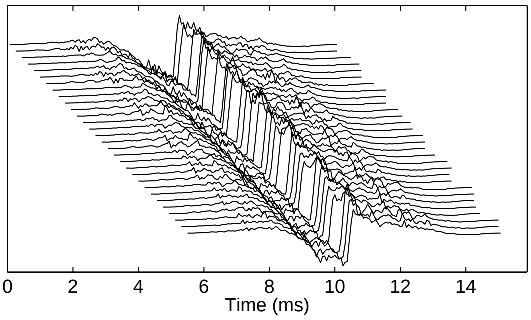

Figure 5: Windowed two-period glottal volume velocity pulse derivatives from the pulse library of the American English female speaker extracted with the automatic speech parametrization method.

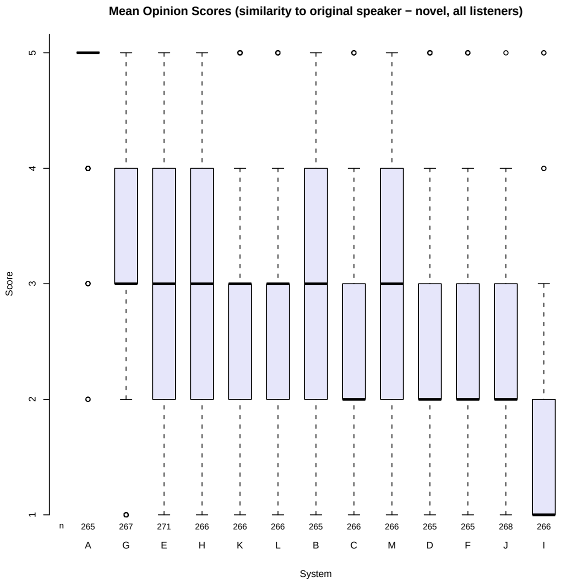

Figure 6: *Similarity scores for the novel sentences for all listeners. Our system is depicted with letter M.*

5. Results and discussion

This year's submissions were mostly very good quality due to large, well-annotated database, and the differences between systems were small. As usual, with many variables, analysis of our own results is difficult. Compared to last year's English hub task, this year's results were slightly worse, not being significantly better than the HMM-based benchmark voice (system C) on any of the measured aspects when all listeners were considered. Closer examination revealed that, for some reason, especially the paid listeners judged our system harshly, while the online listeners preferred our system to system C.

However, current task was a female voice, difficult for our inverse filtering based approach, so it is better to relate our performance with our Blizzard Challenge 2010 mandarin female voice. Here, in reference to HMM benchmark voice, we see clear improvement on speaker similarity, likely due to the new pulse library method as well as SWLP parameterization. Apparently, the similarity is especially strong when reading novels, as seen in Figure 6, where our entry is labeled with letter M. Improvements in HMM modeling and shorter window for unvoiced LSFs seemed to have benefited the intelligibility of our system, which is now in line with other parametric systems. On the downside, naturalness was not improved, probably attributable to the pulse selection artefacts.

6. Conclusions

In this paper, we have described the novel aspects of the GlottHMM system for the Blizzard challenge 2011, most notably, the use of glottal pulse library in a unit selection framework and the weighted linear prediction based speech parameterization.

While the performance of our submitted voice was not exactly stellar, some optimism is warranted. Progress on modeling female speech was noted comparing our entries from previous and current challenges. Also, most of the methods described in this paper were tested for the first time in this challenge in a rather immature state. Better results can be expected based on further experimentation on the pulse library and vocal tract parameterization.

7. Acknowledgements

The research in this paper is supported by the Academy of Finland (projects 135003, 107606, 1128204, 1218259, research programme LASTU), and MIDE UI-ART.

8. References

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

[2] Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M. and Alku, P., "HMM-based speech synthesis utilizing glottal inverse filtering", IEEE Trans. Audio, Speech, and Language Processing, 19(1):153–165, Jan. 2011.

[3] Zen, H., Tokuda, K. and Black, A. W., "Statistical parametric speech synthesis", Speech Commun., 51(11):1039–1064, 2009.

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

[5] Maia, R., Toda, T., Zen, H., Nankaku, Y. and Tokuda, K., "An excitation model for HMM-based speech synthesis based on residual modeling", Sixth ISCA Workshop on Speech Synthesis, Aug. 2007.

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

[7] Fant, G., Liljencrants, J. and Lin, Q., "A four-parameter model of glottal flow", STL-QPSR, 4:1–13, 1985.

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

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

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

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

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

[13] Miller, R. L., "Nature of the vocal cord wave", J. Acoust. Soc. Am., 31(6):667–677, Jun. 1959.

[14] Alku, P., "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering", Speech Commun., 11(2–3):109–118, Jun. 1992.

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

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

[17] Magi, C., Pohjalainen, J., Backström, T. and Alku, P., "Stabilised weighted linear prediction", Speech Comm. 51(5):401–411, May 2009.

[18] Ma, C., Kamp, Y. and Willems, L., "Robust signal selection for linear prediction analysis of voiced speech", Speech Comm. 12(1):69–81, 1993.

[19] Moore, B. C. J. and Glasberg, B. R., "A revision of Zwicker's loudness model", ACTA Acustica, 82:335–345, 1996.

[20] Soong, F. K. and Juang, B.-H., "Line spectrum pair (LSP) and speech data compression", Proc. ICASSP, 9:37–40, 1984.

[21] Paliwal, K. and Kleijn, W., "Quantization of LPC parameters", Speech Coding and Synthesis, W. Kleijn and K. Paliwal, Eds. Elsevier, ch. 12, 1995.

[22] Ling, Z.-H., Wu, Y.J., Wang, Y.-P., Qin, L. and Wang, R.-H., "USTC system for Blizzard Challenge 2006: an improved HMM-based speech synthesis method", The Blizzard Challenge 2006 workshop, 2006, http://festvox.org/blizzard.

[23] Nitisaroj, R., Wilhelms-Tricarico, R., Mottershead, B., Reichenbach, J. and Marple, G., "The Lessac Technologies System for Blizzard Challenge 2010", The Blizzard Challenge 2010 workshop, 2010, http://festvox.org/blizzard.

[24] Wu, Y.-J. and Wang, R.-H., "Minimum generation error training for HMM-based speech synthesis", Proc. ICASSP, pp. 89–92, 2006.