UTILIZING GLOTTAL SOURCE PULSE LIBRARY FOR GENERATING IMPROVED EXCITATION SIGNAL FOR HMM-BASED SPEECH SYNTHESIS

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

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

ABSTRACT

This paper describes a source modeling method for hidden Markov model (HMM) based speech synthesis for improved naturalness. A speech corpus is first decomposed into the glottal source signal and the model of the vocal tract filter using glottal inverse filtering, and parametrized into excitation and spectral features. Additionally, a library of glottal source pulses is extracted from the estimated voice source signal. In the synthesis stage, the excitation signal is generated by selecting appropriate pulses from the library according to the target cost of the excitation features and a concatenation cost between adjacent glottal source pulses. Finally, speech is synthesized by filtering the excitation signal by the vocal tract filter. Experiments show that the naturalness of the synthetic speech is better or equal, and speaker similarity is better, compared to a system using only single glottal source pulse.

Index Terms— speech synthesis, HMM, source modeling, glottal inverse filtering, pulse library

1. INTRODUCTION

Hidden Markov model (HMM) based speech synthesis [1] has gained much popularity due to its flexibility and small footprint, but the quality of the synthetic speech has remained poorer compared to that of high-quality unit-selection based text-to-speech (TTS) systems. The degradation in quality stems from three factors: over-simplified vocoder techniques, acoustics modeling accuracy, and over-smoothing of the generated speech parameters [1]. This paper concentrates on the first factor, the over-simplified vocoder techniques.

Most HMM-based TTS systems are based, in one form or another, on the source-filter theory [2]. The reliable modeling of real speech production has been challenging and the quality of the synthetic speech has remained low due to the over-simplified excitation models. Recently, several studies have been conducted to develop better excitation techniques. For example, mixed excitation [3] has clearly improved the quality of the synthesis compared to that given by simple impulse train excitation. In another approach, the so called residual modeling [4], the impulse and the noise parts of the excitation signal are modified with dedicated filters to better model the excitation waveform. Liljencrants-Fant (LF) model [5] of the differentiated volume velocity waveform has also been utilized as a technique for source modeling [6]. While the previous methods effectively solve the buzziness problem of simple excitation, these models still leave room for improvement in terms of naturalness and reproduction of speaker characteristics.

The accurate modeling of the glottal source signal has proven to be very difficult. Thus, the use of specific glottal flow models has been replaced in several recent studies by a different approach, the utilization of the estimated glottal source waveform *per se*. For example, [7, 8] have used principal component analysis (PCA) to model the rough glottal source waveform and used it in HMM-based speech synthesis. In [9], a codebook of pitch-synchronous residual frames is constructed, and the excitation signal is reconstructed by selecting frames according to the down-sampled waveform trained into the HMM. The results have been clearly better compared to conventional excitation models.

In our recent approach for source modeling, glottal source pulses computed from real speech have been used for generating the excitation signal [10]. Although only one glottal source pulse waveform is selected and modified for generating a sentence in this method, the quality has been much better compared to that of conventional methods [10, 11]. However, even with spectral modifications, the use of single pulse makes a strong assumption about the correct pulse shape, and cannot properly model, for example, diplophony, nor phones with strong unvoiced component. Neither can we assume that a single pulse would be able to cover the voice characteristics of the modeled speaker.

Thus, in this paper, we extend the use of a single glottal source pulse per sentence to the use of a library of various pulses. In the analysis stage, we first decompose the speech signal into the glottal source signal and the model of the vocal tract filter using glottal inverse filtering. After the decomposition, we extract pulses from each analysis frame and map these pulses according to source signal parameters. Thus, the large dynamics in the voice source is retained in the pulse library. 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 the joint cost, consisting of a target and concatenation costs.

In what follows, we describe the new synthesis system, its experimental evaluation, and discuss both the benefits and drawbacks of the method.

2. SPEECH SYNTHESIS SYSTEM UTILIZING GLOTTAL SOURCE PULSE LIBRARY

The TTS system utilizing the glottal source pulse library is based on our former work. Previously, we have used glottal inverse filtering [12] for decomposing the speech signal into the glottal volume velocity signal and the all-pole model of the vocal tract. The previous system and its quality compared to other well known systems is documented in [10, 11]. In this study, the same speech synthesis system, denoted as GlottHMM, is elaborated to include a glottal source

This research is supported by the Academy of Finland (projects 135003 LASTU programme, 1128204, 1218259), and MIDE UI-ART.

pulse library. In this section, the architecture of the new system is described.

The TTS system used in this work is built on a basic framework of an HMM-based speech synthesis system [13], but the parametrization and synthesis methods (*i.e.* the vocoder part) differ substantially from conventional methods, and therefore are explained in detail below.

2.1. Speech Parametrization

The flow chart of the speech parametrization algorithm is shown in Fig. 1. First, the signal is high-pass filtered (cut-off frequency 70 Hz) in order to remove possible low-frequency fluctuation from the signal. The signal is then 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. The speech features presented in Table 1 are extracted at each time index.

For both speech frames, inverse filtering is performed in order to estimate the glottal volume velocity waveform. An automatic glottal inverse filtering method, iterative adaptive inverse filtering (IAIF) [12], is utilized. IAIF iteratively cancels the effects of the vocal tract and the lip radiation from the speech signal using all-pole modeling. Consequently, the outputs of the IAIF algorithm are the estimated glottal source signal and the all-pole model of the vocal tract. In order to capture the variation in the glottal source due to different phonation or speaking styles, the spectral envelope of the glottal source is parametrized with linear predictive coding (LPC). LPC models of the vocal tract and the voice source are further converted to line spectral frequencies (LSFs). In case of unvoiced frames, conventional LPC is used to evaluate the spectral model of speech.

The fundamental frequency (F_0) is estimated from the glottal source signal with the autocorrelation method. In order to evaluate the degree of voicing in the glottal source 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. In addition, the magnitude difference of the first ten harmonic peaks compared to the first harmonic magnitude of the excitation spectrum is extracted to describe the low-frequency spectral tilt more accurately.

Glottal source 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. After 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 library. Examples of pulse waveforms are shown in Fig. 2. All the voice source parameters (all parameters in Table 1 except the vocal tract spectrum) are also stored to describe each pulse. In addition, a down-sampled constant length (10 ms) version of the pulse is stored to enable the comparison of waveforms between different pulses 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 spectrum	20
Vocal tract spectrum	30

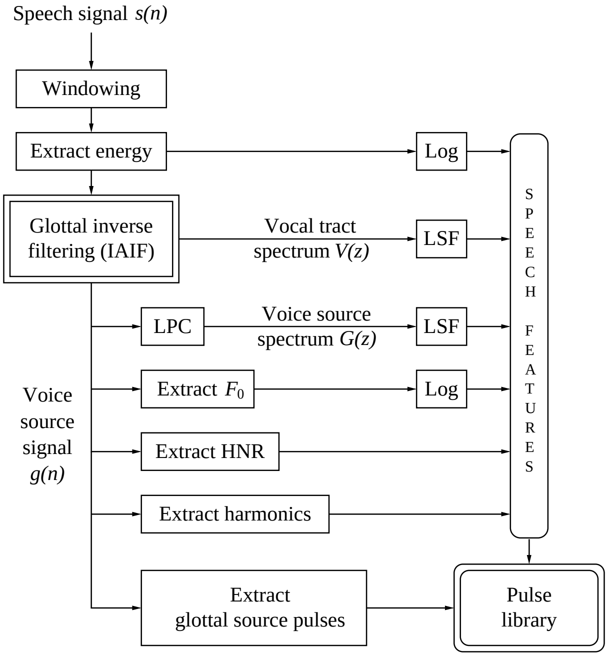

Fig. 1. Illustration of the parametrization stage. The speech signal $s(n)$ is decomposed into the glottal source signal $g(n)$ and the all-pole model of the vocal tract $V(z)$ using glottal inverse filtering. The glottal source signal is further parametrized into several parameters and a glottal source pulse library.

After the parameterization and construction of the pulse library, the parameters described in Table 1 are trained with the HTS system [13] with one stream assigned for each parameter type.

2.2. Synthesis

The flow chart of the synthesis stage is shown in Fig. 4. The excitation signal consists of voiced and unvoiced sound sources. The voiced sound source is composed of glottal source pulses selected from the pulse library. The best pulse for each time index is selected by minimizing the joint cost composed of target and concatenation costs. The target cost consists of the error in the voice source parameters between the ones generated by the HMM and the ones stored for each pulse in the library. The concatenation cost consists of the root mean square (RMS) error between the downsampled versions of

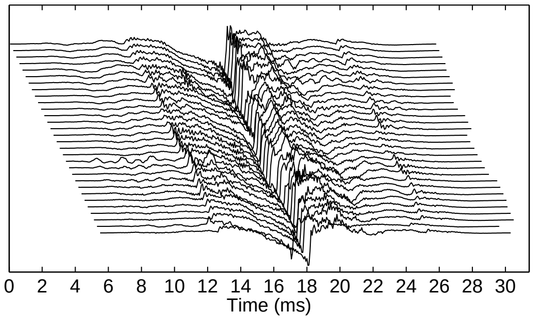

Fig. 2. Windowed two-period glottal volume velocity pulse derivatives from the pulse library of a male speaker extracted with the automatic analysis scheme.

the pulse candidates. The selection process of the pulses is tuned by defining individual weights for each voice source parameter, and additional weights for target and concatenation errors. However, with a large number of concatenation points, the full Viterbi search is not computationally feasible. Therefore, the minimization of the joint cost is performed only over thee consecutive pulses.

Minimizing the target cost of the voice source parameters ensures that a pulse with desired properties (fundamental period, spectral tilt, the amount of noise) is most likely to be chosen. Since F_0 is included in the target cost, a pulse with approximately correct fundamental period will be chosen, and thus further processing, such as interpolation, of the pulse is unnecessary, as opposed to the single pulse technique. Only the energy of the pulse is equalized to the energy measure given by the HMM. 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. After the selection process, the excitation signal is generated by overlap-adding the selected pulses according to the current F_0 value. Thus a pulse train comprising a series of individual glottal source pulses is generated.

The unvoiced excitation is composed of white noise, whose gain is determined according to the energy measure generated by the HMM system. The vocal tract parameters are enhanced [14] in order to alleviate for the over-smoothing, and the LSFs are then interpolated and converted to LPC coefficients, and used for filtering the excitation signal.

Examples of speech spectrograms comparing the single pulse and the pulse library techniques are shown in Fig. 3. The original speech spectrogram is shown in the uppermost figure for comparison. A representative set of speech samples is available online at http://www.helsinki.fi/speechsciences/synthesis/samples.html.

3. EVALUATION

In order to assess the quality of the new system utilizing the glottal source pulse library, several subjective listening tests were conducted. First, a database of 600 sentences spoken by a 39-year-old Finnish male speaker (*mv* database) was used to train the systems. From the first 30 sentences of the database, a pulse library, comprising a total of 22044 glottal source pulses, was constructed. No preselection of the pulses was performed. Speech was synthesized with both the single pulse and the pulse library techniques.

A comparison category rating (CCR) test was used to assess the quality of the two systems. In the CCR test, the listeners were presented with pairs of speech samples and they were asked to assess the difference in quality between the samples on the comparison mean opinion score (CMOS) scale. Ten sentences from the held-out data were used for generating the test samples. Ten Finnish listeners compared a total of 30 speech sample pairs. The ranking of the methods was evaluated by averaging the scores of the CCR test for each method.

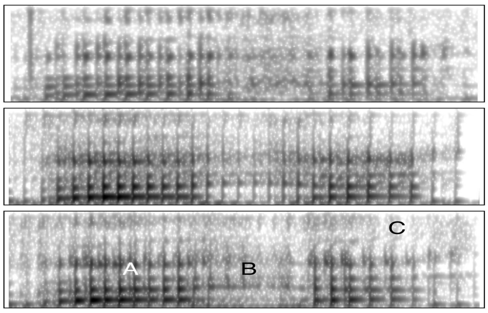

Fig. 3. Spectrograms (0–8000 Hz) of utterance-final word “vähän” (little). Top: original, middle: single pulse, bottom: pulse library. Note the improved modeling of A) diplophony B) voiced fricatives C) high frequencies.

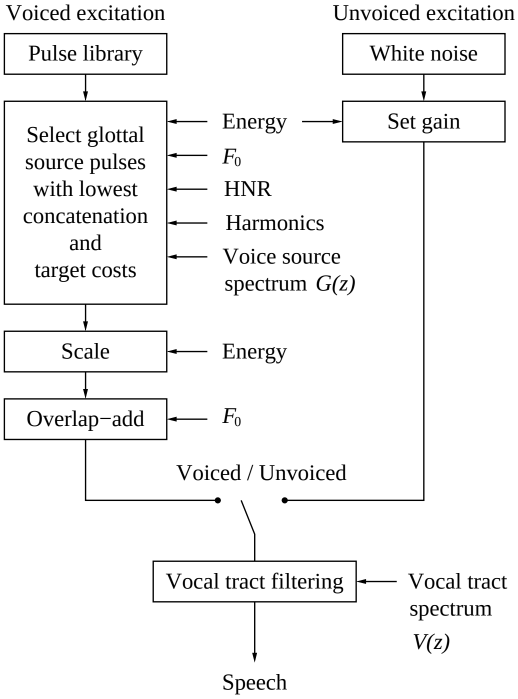

Fig. 4. Illustration of the synthesis stage. The voiced sound source is composed of glottal source pulses selected from the pulse library. Unvoiced excitation is composed of white noise. The excitation signals are combined and filtered with the vocal tract filter $V(z)$ to generate speech.

Second, a database of English male professional speaker (*rjs* database) was used to train the systems. From the first 30 sentences of the database, a pulse library, comprising a total of 23332 glottal source pulses, was constructed. In the case of the *rjs* database, natural speech samples were also included to test the performance of the synthesizers compared to natural speech as well. The CCR method was used again to compare the systems. Ten sentences were used for generating the test samples, and ten Finnish listeners compared a total of 70 speech sample pairs. The listeners were not native speakers of English, but they were all experienced users of English.

Finally, the similarity of the synthetic speech of the *rjs* database was assessed. In this test, listeners were presented three speech samples, *A*, *B*, and *Ref*, and they were asked to choose, which one of the two samples, *A* or *B*, sounded more similar to the speaker in the reference sample. Listeners also had the option of no preference between the two samples. Ten Finnish listeners compared a total of 20 speech sample pairs.

The results of the CCR tests for the *mv* and the *rjs* databases are shown in Fig. 5. In both tests, the pulse library technique has higher average scores. However, the differences are not statistically significant due to the ambivalent results; some listeners yet preferred the uniform quality given by the single pulse technique. The results of the similarity test for the *rjs* database are shown in Fig. 6. The pulse library technique is rated slightly more similar to the reference speaker.

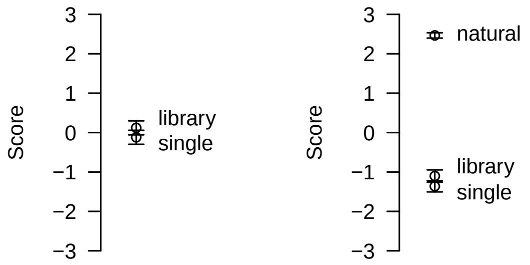

Fig. 5. Left: Ranking of the CCR test for the *mv* voice. Right: Ranking of the CCR test for the *rjs* voice. Sample types: pulse library (library), single pulse (single), natural (natural). The 95 % confidence intervals are presented for each score.

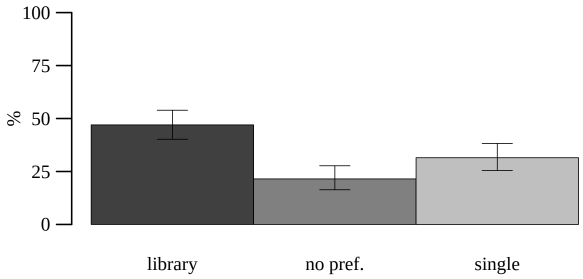

Fig. 6. Results of the similarity test for the *rjs* voice: pulse library (library), single pulse (single). The bars indicate the percentage of the total number of answers to the question *"Which one is more similar to the reference speaker?"*. The 95 % confidence intervals are presented for each bar.

4. CONCLUSIONS

The results show that the new method utilizing the glottal source pulse library yields better or equal quality compared to the single pulse technique. Since the quality of the single pulse technique is shown to be high [10, 11], the proposed technique is clearly better compared to traditional excitation methods. The similarity, or the speaker identity, was rated better with the pulse library technique.

The quality and similarity improvements stem from the properties in the pulse library retained from the natural voice source. Especially partly voiced sounds, such as [v,f,h,z] are produced much more naturally compared to the single pulse technique. This is due to the inability of the pulse modification procedure to modify the pulses to correspond to the extreme modes of natural excitation. However, the listening test results reveal that the two methods divided the opinions of the listeners. While the more lively excitation generated by the pulse library is perceived as more natural for most of the listeners, others tend to rate it less natural due to some irregularity in the excitation. Thus, more sophisticated methods for selecting the pulses from the library are required, for example, by preselecting contextually appropriate pulse candidates based on HMM decision-tree clusters. Furthermore, in order to prevent the selection of the same pulse for too long, possibly resulting in a larger gap to the next appropriate pulse, a bias may be introduced to the concatenation error for identical adjacent pulses. According to small scale experiments, tentatively some of the problems of irregularity are solved.

Compared to the closest similar approach by Drugman *et al.* [9], the proposed method has several aspects that makes it potentially more feasible. First, the source signal utilized here corresponds to the glottal excitation, while LPC residual signal is used in [9]. Thus, in our approach, the real variation of the glottal excitation is retained in the pulse library, whereas part of the voice source variation, most importantly the spectral tilt, is leaked to the LPC spectrum in [9]. Second, [9] assumes that the pulses are best described and selected by comparing the lowpass waveforms of the residuals. In this assumption, the amount of noise present in the glottal source is neglected. Third, no concatenation cost was utilized in [9]. Furthermore, no interpolation of glottal source pulses is required for our approach. In our experiments, interpolation of the pulses introduce some artefacts that are avoided by selecting pulses of appropriate length and overlap-adding them according to F_0.

This is our first attempt to utilize glottal source pulse library for source modeling. The results are promising and they demonstrate the capability of the method for greatly improving the quality of HMM-based speech synthesis; the best instances of synthetic speech are very close to natural ones. However, synthesis with the pulse library requires further investigation to fully utilize its potential.

5. REFERENCES

[1] A.W. Black, H. Zen, and K. Tokuda, "Statistical parametric speech synthesis," in *Proc. ICASSP*, Apr. 2007, vol. 4, pp. 1229–1232.

[2] G. Fant, *Acoustic Theory of Speech Production*, Mouton, The Hague, 1960.

[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] 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.

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

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

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

[13] 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.

[14] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "Comparison of formant enhancement methods for HMM-based speech synthesis," in *SSW7*, Sep. 2010, pp. 334–339.