HMM-Based Finnish Text-to-Speech System Utilizing Glottal Inverse Filtering

Tuomo Raitio[1], *Antti Suni*[2], *Hannu Pulakka*[1],
Martti Vainio[3], *Paavo Alku*[1]

[1]Department of Signal Processing and Acoustics, Helsinki University of Technology, Espoo, Finland
[2]Department of Speech Sciences, University of Helsinki, Finland
[3]Department of General Linguistics, University of Helsinki, Finland
tuomo.raitio@tkk.fi, asuni@cc.helsinki.fi

Abstract

This paper describes an HMM-based speech synthesis system that utilizes glottal inverse filtering for generating natural sounding synthetic speech. In the proposed system, speech is first parametrized into spectral and excitation features using a glottal inverse filtering based method. The parameters are fed into an HMM system for training and then generated from the trained HMM according to text input. Glottal flow pulses extracted from real speech are used as a voice source, and the voice source is further modified according to the all-pole model parameters generated by the HMM. Preliminary experiments show that the proposed system is capable of generating natural sounding speech, and the quality is clearly better compared to a system utilizing a conventional impulse train excitation model.

Index Terms: speech synthesis, glottal inverse filtering, HMM

1. Introduction

The ultimate goal of text-to-speech synthesis (TTS) is to enable creating natural sounding speech from arbitrary text. Moreover, the current trend in TTS research calls for systems that enable producing speech in different speaking styles with different speaker characteristics and even emotions. In order to fulfill these stringent general requirements, two major synthesis techniques have attracted increasing interest in the speech research community during the past decade. These two alternatives are (1) the unit selection technique and (2) the hidden Markov model (HMM) based approach. The former has been shown to yield synthetic speech of highly natural quality. However, unit selection techniques do not allow for easy adaptation of the TTS system to different speaking styles and speaker characteristics. In addition, their implementation requires databases of extensive sizes, which severely limit the use of this TTS technique, for example, in mobile terminals. HMM-based techniques, in turn, benefit from better adaptability and a clearly smaller memory requirement. However, the current HMM systems often suffer from degraded naturalness in quality. It can be argued that a potential reason for the reduced naturalness in the current HMM-based TTS systems can be explained by the use of signal generation techniques which are oversimplified to properly mimic natural speech pressure waveforms.

A large part of what can be characterized as naturalness in speech emerges from different voice characteristics as well as their context dependent changes. Therefore, it is justified in speech synthesis to search for methods aiming at accurate modeling of different voice characteristics as well as prosodic features of speech. Towards these goals, HMM-based synthesizers have been developed with special emphasis on voice characteristics such as speaker individualities, speaking styles, and emotions [1]. Moreover, some recent studies have introduced improvements to the parametric HMM systems' signal generation techniques by utilizing, for example, mixed excitation [2] and residual modeling [3]. These techniques have been shown to improve the quality of synthetic speech compared to systems utilizing a traditional impulse train excitation model. However, the quality of the systems using these techniques still remains far from the quality of natural speech.

In the real human voice production mechanism, the excitation of (voiced) speech is represented by the glottal volume velocity waveform generated by the vibrating vocal folds. This excitation signal, the glottal source, has naturally attracted interest in speech synthesis and many techniques have been proposed to mimic the glottal source of natural speech. One such technique is the Liljencrants-Fant (LF) model of the differentiated glottal source that has been used both in traditional rule-based synthesis [4, 5] as well as within an HMM-based speech synthesizer [6]. However, the use of artificial glottal flow pulses usually results in a somewhat buzzy quality due to a strong harmonic structure at higher frequencies. To overcome this problem, the idea of utilizing glottal flow pulses extracted from real speech with the help of glottal inverse filtering has been proposed [7, 8]. However, previous studies based on glottal flow pulses extracted from natural speech are limited to special purposes such as the generation of isolated vowels, and the benefits from combining automatic glottal inverse filtering with an HMM-based speech synthesizer have not been utilized.

In this paper, a novel HMM-based speech synthesis system that utilizes glottal inverse filtering for generating natural sounding synthetic speech is presented. The rest of the paper is organized as follows: Section 2 describes the proposed speech synthesis system. The results of the experiments with the new synthesizer are presented in Section 3. Discussion on the proposed speech synthesis system and future plans are presented in Section 4, and final conclusions are presented in Section 5.

2. HMM-based Speech Synthesis System

The proposed new TTS system aims to produce natural sounding synthetic speech capable of conveying different styles of speaking as well as emotions. In order to achieve this goal, the function of the real human voice production apparatus is modeled with the help of glottal inverse filtering embedded in an HMM framework. Automatic glottal inverse filtering is used in the parametrization stage in order to compute a parametric feature expression for the voice source and the vocal tract transfer function. In the synthesis stage, natural glottal pulses are

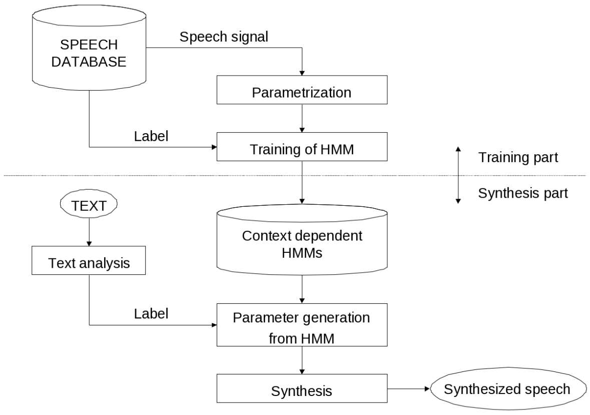

Figure 1: *System overview.*

used for generating the source signal for voiced sounds and the spectral envelope of this glottal excitation waveform is modified with an adaptive IIR filter to imitate the time-varying changes in the real voice source. The current implementation of the system is applied for Finnish, but, in principle, it can be extended to other languages as any data driven synthesizer.

The overview of the system is shown in Figure 1. The system consists of two major parts: training and synthesis. In the training part, speech parameters computed by glottal inverse filtering are extracted from sentences of a speech database. This parametrization stage is a major innovation in the proposed TTS system in comparison to previous HMM-based synthesizers and therefore it is explained in detail below in Section 2.1. The obtained speech parameters are then modeled in the framework of the HMM. In the synthesis part, the HMMs are concatenated according to the analyzed input text and speech parameters are generated from the HMM. The parameters are then fed into the synthesis module for creating the speech waveform.

2.1. Speech Parametrization

The parametrization stage tries to compress the information of the speech signal into a few parameters which would describe the essential characteristics of the original speech signal as accurately as possible. The flow chart of the parametrization stage is shown in Figure 2. First, the speech signal is high-pass filtered with a cut-off frequency of 60 Hz in order to remove any distorting low-frequency fluctuations. The high-pass filtering is especially important for glottal inverse filtering, where even weak low-frequency components may cause extensive fluctuations in the estimated glottal flow. The signal is windowed with a rectangular window to 25-ms frames at 5-ms intervals. The parameters are then extracted from each frame.

The extracted features are presented in Table 1. The core of the parametrization stage is the glottal inverse filtering that estimates the glottal volume velocity waveform from the speech pressure signal. An automatic inverse filtering method, Iterative Adaptive Inverse Filtering (IAIF) [9, 7], is utilized in the system. IAIF iteratively cancels the effects of the vocal tract and the lip radiation from the speech signal using adaptive all-pole modeling. The outputs of the inverse filtering block are the estimated glottal flow signal and the LPC (Linear Predictive Coding) model of the vocal tract (denoted by Voiced spectrum in Table 1). The spectral envelope of the glottal flow is parametrized with LPC (denoted by Voice source spectrum in Table 1). Additionally, an LPC model (denoted by Unvoiced spectrum in Table 1) is computed for unvoiced speech sounds directly from the speech frame. All the obtained LPC models are converted to Line Spectral Frequencies (LSF) [10], a parametric representation of LPC information well-suited to be used in a statistical HMM system. LSFs of voiced and unvoiced spectrum are further converted to the mel scale.

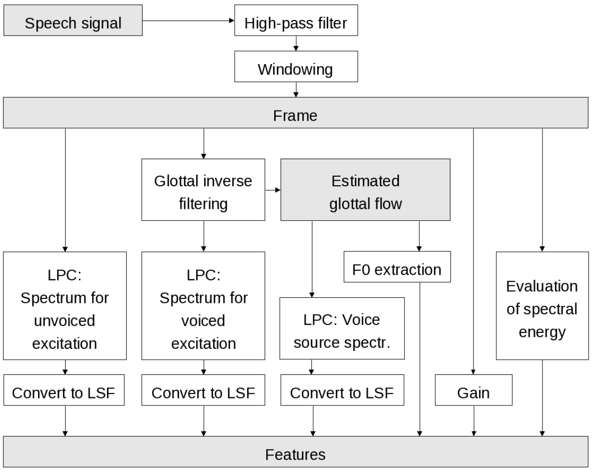
Speech signal
High-pass filter
Windowing
Frame
Glottal inverse filtering
Estimated glottal flow
LPC: Spectrum for unvoiced excitation
LPC: Spectrum for voiced excitation
F0 extraction
LPC: Voice source spectr.
Evaluation of spectral energy
Convert to LSF
Convert to LSF
Convert to LSF
Gain
Features

Figure 2: *Flow chart of the parametrization stage.*

The fundamental frequency is determined from the glottal flow with the autocorrelation method, and the energy of the frame is computed. In addition, the spectral energy of five bands (0–1000 Hz, 1000–2000 Hz, 2000–4000 Hz, 4000–6000 Hz, and 6000–8000 Hz) is calculated from the speech frame with FFT for determining the unvoiced excitation.

2.2. Synthesis

The flow chart of the synthesis stage is presented in Figure 3. The excitation signal consists of voiced and unvoiced sound sources. The basis of the voiced sound source is a glottal flow pulse extracted from a natural vowel produced by a male speaker. In comparison to artificial glottal flow pulses one may argue that the use of a real glottal pulse helps in preserving the naturalness and quality of the synthetic speech. By interpolating and scaling in magnitude this real glottal flow pulse, a pulse train comprising a series of individual glottal pulses with varying period lengths and energies is generated. In order to mimic the natural variations in the voice source, the desired voice source all-pole spectrum ($H_{\mathrm{orig}}(z)$) generated by the HMM is applied to the pulse train. This is achieved by first evaluating the LPC spectrum of the generated pulse train ($H_{\mathrm{synth}}(z)$), and

Table 1: Speech features and the number of parameters.

Feature	Parameters per frame
Fundamental frequency	1
Energy	1
Spectral energy	5
Voice source spectrum	10
Voiced spectrum	20
Unvoiced spectrum	20

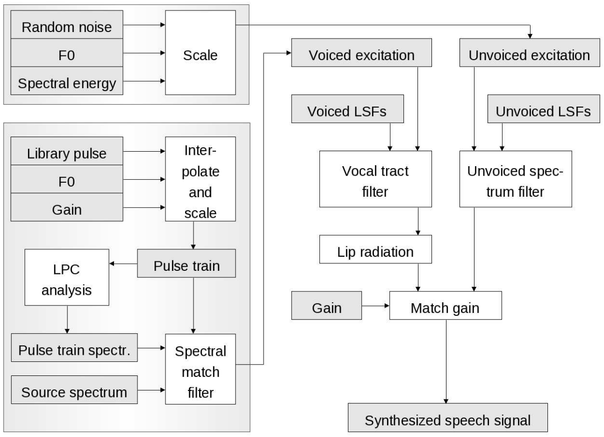

Figure 3: *Flow chart of the synthesis stage.*

then filtering the pulse train with an IIR filter

$$H_{\text{match}}(z) = \frac{H_{\text{orig}}(z)}{H_{\text{synth}}(z)}, \tag{1}$$

which first flattens the spectrum of the pulse train and then applies the desired spectrum. The unvoiced sound source is represented by white noise which is weighted according to the energies of the five frequency bands.

Separate spectra for voiced and unvoiced excitation are used since the vocal tract transfer function does not apply for unvoiced speech sounds. In order to incorporate an unvoiced component also when the speech sounds are voiced (e.g. breathy sounds), both streams are produced concurrently throughout the frame. A formant enhancement procedure [12] is applied to the LSFs generated by the HMM to compensate for the averaging effect of the statistical modeling. The voiced and unvoiced LSFs are then interpolated and converted to LPC coefficients, and used for filtering the excitation signals. For voiced excitation, the lip radiation effect [11] is modeled as a first-order differentiation operation. Finally the gain of the combined signal is matched according to the energy measure generated by the HMM.

3. Experiments

The proposed method was tested by training the system with a prosodically annotated database of 600 phonetically rich sentences spoken by a 39-year-old Finnish male speaker, comprising approximately one hour of speech material. The speech was sampled at 16 kHz. 20th-order LPC was used in parametrizing the spectra of voiced and unvoiced speech, and 10th-order LPC was used in parametrizing the voice source spectrum. Features described in previous section were extracted together with their delta and delta-delta features. A 7-state left-to-right model structure with 5 emitting states was used. Each new feature type was assigned to an individual stream, resulting in a model of 8 streams with a feature order of 171 in total. In the current system, all streams except the fundamental frequency were modeled by a single Gaussian distribution with a diagonal covariance matrix.

For evaluation purposes, a de facto standard HTS model structure described in [1] was used as a baseline system. This previously developed HHM system uses the mel-cepstral analysis technique [13] for spectrum modeling and a simple impulse train excitation model for excitation generation. Instead of using more sophisticated excitation models, the simple one was selected for the comparison because its quality is generally known in the field of speech synthesis.

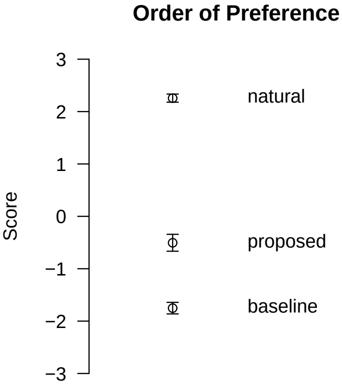

Figure 4: *Ranking of the CCR test for the following speech samples: natural speech (natural), proposed system (proposed), baseline system with an impulse train excitation model (baseline) [1]. The mean score has no explicit meaning, but the distances between the scores are essential. The 95 % confidence intervals are presented for each score.*

The training procedure for both systems was similar to that described in [1]. First, monophone models were trained and then converted to context dependent models. In order to robustly estimate the model parameters, decision tree state-tying was performed for each stream. For decision tree clustering, a rich set of contextual features was extracted by a proprietary front-end, ranging from phone level to higher-level phonological features such as word prominence, clause type, and whether the sentence starts a new topic.

3.1. Subjective Evaluation

Two subjective listening tests were conducted to evaluate the quality of the proposed TTS system. First, a Comparison Category Rating (CCR) test [14] was used to assess the quality of the proposed method in comparison to natural speech and synthetic speech generated by the baseline system. In the CCR test, the listeners were presented with a pair of speech samples on each trial, and they were asked to assess the quality of the second sample compared to the quality of the first one on the Comparison Mean Opinion Score (CMOS) scale. Although the CCR test is designed for slightly different purposes, the test was considered suitable for obtaining preliminary data. Ten randomly chosen sentences from held-out data were used for generating the test samples for each method. Eleven Finnish naive listeners (9 men and 2 women) compared a total of 70 speech sample pairs. The ranking of the methods was evaluated by averaging the scores of the CCR test for each method. The ranking of the three methods with 95 % confidence intervals according to the CCR test are shown in Figure 4.

In the second test, only the synthetic sounds generated by the two HMM-based TTS systems were involved. A pair comparison test method was used, where subjects listened to samples A and B, and selected the one they would rather listen to. They were also given an option to choose that the samples

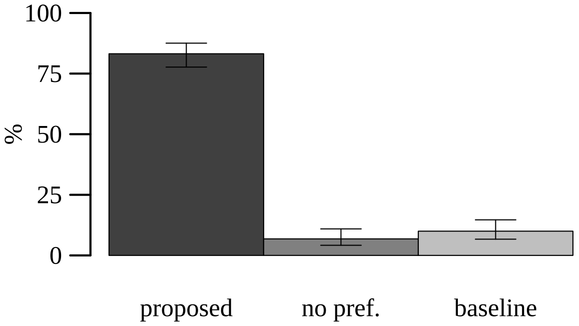

Figure 5: *Results of the pair comparison test applied for the proposed system (proposed) and the baseline system with an impulse train excitation model (baseline) [1]. The bars indicate the percentage of the total number of answers to the question "Which one would you rather listen to". The center bar (no pref.) indicates no preference for either of the methods. The 95 % confidence intervals are presented for each bar.*

sounded about the same, indicating no preference between the two samples. Ten randomly chosen sentences from held-out data (different from the ones used in the CCR test) were used for generating the test samples for each method. Eleven Finnish naive listeners (9 men and 2 women) compared a total of 24 speech sample pairs. The results of the second test with 95 % confidence intervals are shown in Figure 5.

The results of the first test show that the proposed new TTS system utilizing glottal inverse filtering has a considerably better quality than the previously developed HHM-based method. Compared to natural speech, the quality of the proposed system is clearly worse. However, since the prosodic features of the synthetic speech were generated directly from the HMM, the evaluated degradation in quality may partly result from the prosodic discrepancies between the synthetic and natural speech samples. The second test shows that the proposed system is almost always preferred over the baseline system.

4. Discussion

The experimental results show that the proposed system is able to generate natural sounding speech. However, the full potential of the proposed system is not entirely used in the current implementation. For example, the interpolation of the glottal pulse according to the fundamental frequency is far from the natural behavior of the glottal flow, and the use of a single glottal library pulse is unable to mimic the dynamics of glottal flow pulses that exist in natural continuous speech. Experiments made with the system show that the selection of the library pulse and the technique used for changing the fundamental frequency of the voice source have significant effects on the quality of the synthesized speech. Some problems were also discovered in the learning process of the voice source spectrum parameters indicating that they became somewhat oversmoothed and resulted in a lack of desired variation in the voice source characteristics. This may arise from too small training material, or from problems in context clustering.

The development of the presented system continues, and future work will be focused on improving the use and shaping of the natural glottal pulses, and enhancing the use of voice source characteristics obtained by glottal inverse filtering in the synthesis module.

5. Conclusions

In this study, a new HMM-based text-to-speech system utilizing glottal inverse filtering was described. Subjective listening tests showed that the quality of the proposed system was considerably better when compared to an HTS system with a traditional impulse train excitation model. The utilization of glottal inverse filtering in an HMM-based TTS system is justified since a large part of what can be characterized as naturalness in speech emerges from the voice source. Thus, utilizing knowledge that describes the functioning of the real excitation of the human voice production mechanism might lead to improved naturalness of synthetic speech.

6. Acknowledgments

This project is supported by the Academy of Finland (projects 111848, 107606).

7. References

[1] Tokuda, K., Zen, H. and Black, A. W., "An HMM-based speech synthesis system applied to English", Proc. IEEE Workshop on Speech Synthesis, 227–230, 2002.

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

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

[4] Carlson, R., Fant, G., Gobl, C., Granström, B., Karlsson, I. and Lin, Q., "Voice source rules for text-to-speech synthesis", Proc. ICASSP, 1:223–226, 1989.

[5] Carlson, R., Granström, B. and Karlsson, I., "Experiments with voice modelling in speech synthesis", Speech Commun., 10:481–489, 1991.

[6] Cabral, J. P., Renalds, S., Richmond, K. and Yamagishi, J., "Towards an improved modeling of the glottal source in statistical parametric speech synthesis", Sixth ISCA Workshop on Speech Synthesis, Aug. 2007.

[7] 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.

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

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

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

[11] Flanagan, J. L., "Speech analysis, synthesis and perception", Springer-Verlag, 1972.

[12] Ling, Z.-H., Wu, Y.-J., Wang, Y.-P., Qin, L., Wang, R.-H., "USTC system for Blizzard Challenge 2006 an improved HMM-based speech synthesis method", Blizzard Challenge Workshop, 2006.

[13] Imai, S., "Cepstral analysis synthesis on the mel frequency scale", Proc. ICASSP, 8:93–96, 1983.

[14] Recommendation ITU-T P.800, "Methods for subjective determination of transmission quality", International Telecommunication Union, Aug. 1996.