HMM-Based Speech Synthesis Utilizing Glottal Inverse Filtering

Tuomo Raitio, Antti Suni, Junichi Yamagishi, Hannu Pulakka, Jani Nurminen, Martti Vainio, and Paavo Alku

***Abstract*—This paper describes an HMM-based speech synthesizer that utilizes glottal inverse filtering for generating natural sounding synthetic speech. In the proposed method, speech is first decomposed into the glottal source signal and the model of the vocal tract filter through glottal inverse filtering, and thus parametrized into excitation and spectral features. The source and filter features are modeled individually in the framework of HMM and generated in the synthesis stage according to the text input. The glottal excitation is synthesized through interpolating and concatenating natural glottal flow pulses, and the excitation signal is further modified according to the spectrum of the desired voice source characteristics. Speech is synthesized by filtering the reconstructed source signal with the vocal tract filter. Experiments show that the proposed system is capable of generating natural sounding speech, and the quality is clearly better compared to two HMM-based speech synthesis systems based on widely used vocoder techniques.**

***Index Terms*—Speech synthesis, glottal inverse filtering, hidden Markov model.**

I. Introduction

THE ultimate goal of speech synthesis is to create natural sounding spoken expression from arbitrary text. This calls for the ability to synthesize high quality speech, but also provides a means to involve the appropriate variation of the speech characteristics according to the speaker, context, and emotion. The first criterion can be met with a synthesis scheme that concatenates segments of pre-recorded speech. However, these so-called unit selection-based systems are known to suffer from limitations in their ability to vary the speech characteristics [1]. Hidden Markov model (HMM)-based parametric speech synthesis techniques [1]–[4], in turn, are very flexible and can be adapted [5] or modified [6] to generate speech according to virtually any criterion related to varying vocal characteristics. This flexibility is due to the parametric representation of speech, which enables the easy modification of the parameters and the reconstruction of speech from them. However, the quality and naturalness of the parametric HMM-based speech synthesis has remained poorer than that of unit selection methods. This degradation is mainly caused by two factors: oversimplified vocoder techniques that are unable to mimic natural speech pressure waveforms and statistical modeling that causes over-smoothing of the parameters, both resulting in inadequate reconstruction of speech. Research exists on fixing the second problem [7], [8]. This paper will concentrate on the first factor, the inadequate modeling of the real speech production mechanism.

Synthesis methods utilizing parametric representation of speech are largely based on the source-filter theory of speech production [9]. This theory assumes that the production of speech can be interpreted as a linear cascade of three processes: $S(z) = G(z)V(z)L(z)$, where $S(z)$ denotes speech, and $G(z)$, $V(z)$, and $L(z)$ denote the voice source, the vocal tract filter, and the lip radiation effect, respectively. In the real human voice production mechanism, the voice source is represented for the voiced sounds by the glottal volume velocity waveform generated by the vibrating vocal folds. The voice source is known to be the origin for several essential acoustical cues used in spoken communication [10], [11]. In addition to determining the fundamental frequency (F_0) of speech, the voice source also contributes to various spectral and temporal features that are related to voice quality and prosodic variation in speech. In combination, the voice source depicts attitude and emotion, and is also related to acoustical cues underlying the speaker identity. Thus, it can be concluded that a large part of what can be characterized as naturalness in speech emerges from the voice source and its context-dependent variation. Therefore, the search for methods that aim at the accurate modeling of the voice source is justified.

While the vocal tract can be relatively well approximated by a digital filter, the reliable modeling of the voice source has remained more challenging. Therefore, the potential of the voice source modeling as a technique to improve the naturalness of synthetic speech in HMM-based parametric speech synthesizers is substantial. The simplest model for the voice source, a train of impulses [12], [13], is clearly greatly different from the real glottal flow signal, but still widely used in speech synthesis. Improvements to the signal generation techniques of the parametric HMM systems have been introduced, such as mixed excitation [14], residual modeling [15], and two-band excitation [16], and they have been shown to improve the quality compared to systems using the traditional impulse train excitation model. However, the quality of these

Manuscript received May 26, 2009; revised October 29, 2009. This work was supported by Nokia and the Academy of Finland (LASTU programme, project 135003, projects 107606, 128204, 125940). J. Yamagishi is funded by the Engineering and Physical Sciences Research Council (EPSRC) and the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME project). The associate editor coordinating the review of this manuscript for publication was Prof. Richard Gael.

T. Raitio, H. Pulakka, and P. Alku are with the Department of Signal Processing and Acoustics, Helsinki University of Technology (TKK), Espoo, FIN-02015 TKK, Finland, e-mail: Tuomo.Raitio@tkk.fi; Hannu.Pulakka@tkk.fi; Paavo.Alku@tkk.fi

A. Suni and M. Vainio are with the Department of Speech Sciences, University of Helsinki, Finland, e-mail: Antti.Suni@helsinki.fi; Martti.Vainio@helsinki.fi

J. Yamagishi is with the Centre for Speech Technology Research, University of Edinburgh, EH8 9AB, United Kingdom, e-mail: jyamagis@inf.ed.ac.uk

J. Nurminen is with Nokia Devices R&D, Tampere, Finland, e-mail: Jani.K.Nurminen@nokia.com

systems still remains relatively far from the quality of natural speech.

The real voice source generated by the vocal folds has naturally attracted interest in speech research and synthesis, and many techniques have been proposed to mimic the glottal excitation of the human voice production mechanism. Various parametric models for the glottal flow exist, of which the Liljencrants-Fant (LF) model [17] is the most widely used. Previously, LF model pulses have been used in speech synthesis experiments [18], [19], and recently the LF model has been used within an HMM-based speech synthesizer [20], [21]. However, the use of the glottal flow models in HMM-based speech synthesis has been limited and experiments conducted have not succeeded in providing substantial advantages in parametric speech synthesis.

The natural voice source itself is a complex signal and the accurate modeling of the voice source has proven to be very difficult. As an alternative to the artificial voice source models, the idea of utilizing glottal flow pulses extracted from natural speech with the help of glottal inverse filtering has been proposed. For example, natural glottal flow pulses have been used in formant speech synthesis [22], [23], and in creating natural sounding speech stimuli for brain research [24]. However, previous studies based on glottal flow pulses extracted from natural speech are limited to special purposes and have not provided a general synthesis method for utilizing the natural glottal flow.

At present, the most widely used high-quality vocoding technique for HMM-based speech synthesis is the speech manipulation tool STRAIGHT [25]. STRAIGHT was originally proposed as a method to analyze, modify, and resynthesize high-quality speech. Recently, STRAIGHT was adopted for HMM-based speech synthesis [26], and it is currently considered to be one of the best vocoding methods for HMM-based speech synthesis [27]. Despite great improvements, STRAIGHT-based HMM systems cannot yet compete with state-of-the-art unit selection methods in terms of naturalness of synthetic speech [27].

Recently, an idea was proposed to utilize automatic glottal inverse filtering in HMM-based speech synthesis [28]. The proposed concept combines two existing areas of speech technology in a novel way to enable high quality speech synthesis: HMM-based statistical modeling and a physiologically oriented model of speech production based on glottal inverse filtering. Briefly, the idea comprises (a) utilization of automatic glottal inverse filtering in order to decompose and parametrize speech into the glottal source and vocal tract filter components corresponding to the functioning of the real speech production mechanism, (b) individual modeling of these characteristics in the framework of HMM, and (c) utilization of real glottal flow pulses extracted from natural speech for creating the excitation signal. The preliminary results from the first experiments in this work were presented in [29], where the proposed method was compared, with encouraging results, to a conventional HMM-based speech synthesizer [3] using a simple impulse train excitation model. In the present study, the refined and complete algorithm of the proposed method is presented in detail. Most importantly, the proposed new system is compared to a state-of-the-art STRAIGHT-based system in a series of formal listening tests.

This paper is organized as follows. As a background for the proposed method, glottal inverse filtering is described first in Section II. The underlying principles of the new method and a detailed description of the system are presented in Section III. The performance of the proposed system is demonstrated in Section IV through a comparison of the method with natural speech and two other speech synthesis systems. The resulting benefits of the method are depicted and discussed in Section V. Finally, Section VI summarizes the findings and concludes the paper.

II. Glottal Inverse Filtering

A. General

Glottal inverse filtering [30] is a procedure in which the source of voiced speech, the glottal volume velocity waveform, is estimated from speech pressure signals. Conceptually, glottal inverse filtering corresponds to obtaining the glottal volume velocity $G(z)$ from the equation

$$G(z) = \frac{S(z)}{V(z)L(z)}, \tag{1}$$

where $S(z)$, $V(z)$, and $L(z)$ denote z-transforms of the speech signal, vocal tract, and lip radiation effect, respectively. Since the lip-radiation effect $L(z)$ can be modeled as a fixed differentiator [31], only the parameters of the vocal tract need to be estimated to compute the glottal flow from the speech pressure signal.

It is worth noting that the source signal to be solved in Eq. 1 is different from the excitation used in the conventional linear predictive (LP) model, the LP residual. When a speech signal is decomposed into a source and a filter with a single LP analysis, the spectral effects of the three main processes of the real human voice production, the glottal excitation, the vocal tract, and the lip radiation, are combined into a single digital all-pole filter that is excited by an input resembling an impulse train or a noise sequence that is spectrally white. In contrast to this, the excitation given in Eq. 1 is not white but is permitted to show spectral envelopes of varying decays (see Fig. 1). This phenomenon reflects the vibratory patterns of the vocal folds when a speaker regulates adduction of his or her vocal folds, for example, in adjusting the phonation type or in creating different emotional vocal cues.

There are several methods that have been developed during the past decades for estimating the glottal flow from speech (e.g., [32]–[38]). In this study, an automatic glottal inverse filtering method, iterative adaptive inverse filtering (IAIF) [24], [32], is used as a computational tool to implement glottal inverse filtering. IAIF has three methodological and computational features that make it an attractive choice for the present study. First, the method needs only a single input, the speech pressure signal, and there is no need for additional information signals such as the electroglottography (EGG). EGG is widely used in closed phase covariance analysis, a method that is one of the most prevalent glottal inverse filtering algorithms (e.g., [39]–[41]). Second, the IAIF method can

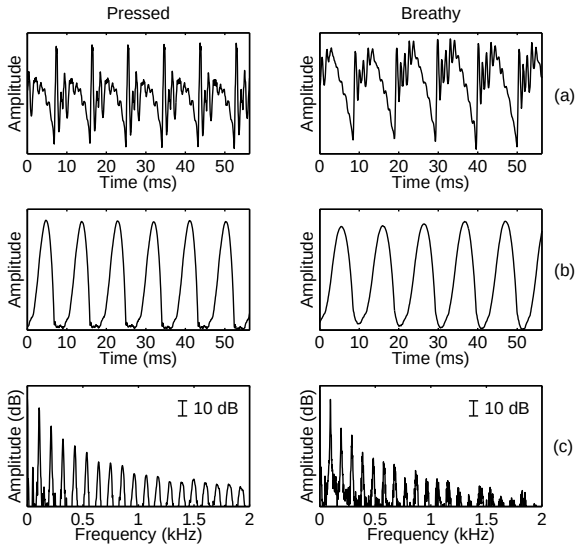


Fig. 1. (a) Waveform of a sustained vowel [a] produced by a male speaker using pressed (left panel) and breathy (right panel) phonation. (b) Corresponding glottal flow signals estimated with the iterative adaptive inverse filtering (IAIF) method [24], [32]. In pressed phonation (left), the glottal flow pulses are shorter compared to breathy phonation and there is a clear closed phase between the pulses. (c) Spectra of the estimated glottal flow signals (only shown for 0–2 kHz). The spectrum of the breathy phonation (right) shows a clear emphasis on the fundamental frequency component. The spectral envelope is also steeper and there is more noise at the higher frequencies.

be implemented in a completely automatic manner and its computational complexity is low. Both of these features are pre-requisites when glottal inverse filtering is used in HMM-based speech synthesis and other applications that require processing of extensive amounts of speech data in the training phase. Third, the IAIF method can be implemented in such a manner that it utilizes the autocorrelation method of linear prediction in modeling of the vocal tract. This enables the use of all-pole filter structures that are guaranteed to be stable, a requirement that is not met, for example, by closed phase covariance analysis.

B. Iterative Adaptive Inverse Filtering

IAIF [24], [32] is a method that automatically decomposes voiced speech into the vocal tract transfer function and the glottal source. In general, the algorithm assumes that the contribution of the glottal excitation to the speech spectrum can be estimated as a low-order (monotonically decaying) all-pole process, which is computed pitch-asynchronously over several fundamental periods. By canceling this effect, a parametric model for the vocal tract is obtained, also in a pitch-asynchronous manner, which is then used to cancel the effect of formants. The method has two main phases. In the first one, a first-order all-pole model is computed to get a preliminary estimate for the glottal contribution. The second phase applies a higher-order all-pole model, which, in principle, is able to yield a more accurate estimate for the contribution of the glottal source. Various spectral modeling tools can be used in the IAIF method, such as digital all-pole modeling (DAP) [42] or linear predictive coding (LPC) [43].

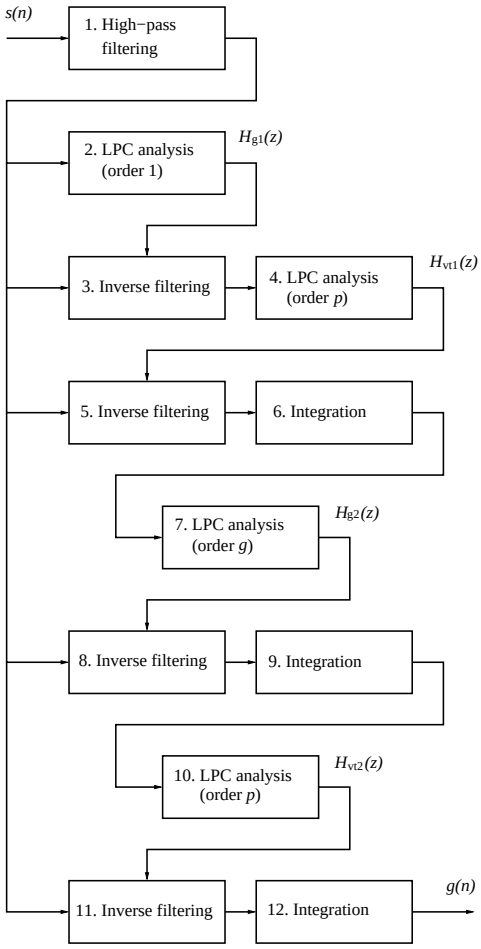


Fig. 2. Block diagram of the IAIF method. The glottal flow signal $g(n)$ is estimated through an repetitive procedure of canceling the effects of the vocal tract and the lip radiation from the speech signal $s(n)$.

Due to its computational efficiency and simplicity, LPC was chosen in this work.

The detailed structure of the IAIF method is shown by the block diagram in Fig. 2. The estimation of the glottal flow with the IAIF method consists of the following stages. First (block 1), the speech signal $s(n)$ is high-pass filtered with a linear phase finite impulse response (FIR) filter (300-tap, linear phase, cut off frequency 70 Hz) in order to remove any low-frequency ambient noise picked up by the microphone. High-pass filtering is a standard pre-processing procedure in glottal inverse filtering because the analysis calls for using high-quality microphones whose frequency response goes down to ca. 5–20 Hz and, consequently, the recordings typically involve low frequency noise. Since canceling the lip radiation effect requires sound integration (blocks no. 6, 9, and 12), the low frequency noise components could distort the glottal flow estimate by introducing a low frequency bias [33]. Next (block 2), a first-order LPC filter is computed, yielding a preliminary estimate of the combined effect of the glottal flow and the lip radiation. The transfer function of the LPC inverse filter is denoted by H_{g1} in Fig. 2. First order LPC analysis yields only a very rough estimate of the contributions of the glottal flow and the lip radiation, because the spectral model

has only one adaptive pole on the real axis in the z-domain. However, using a higher-order linear predictive analysis would most likely result in modeling of the resonant structure of the speech sound, an effect that is to be deliberately avoided if the combined effect of the glottal flow and the lip radiation is to be estimated. Thirdly (block 3), the estimated effect of the glottal flow and the lip radiation is canceled out from the speech signal through inverse filtering. The output is analyzed with a pth order LPC (block 4) in order to obtain the first estimate of the vocal tract filter, denoted by H_{vt1}. (The order p of LPC analysis is typically 20 for speech sampled at 16 kHz.) Next (block 5), the estimated vocal tract transfer function is canceled out from the original speech signal through inverse filtering. The lip radiation is eliminated through integration (block 6), and the resulting signal forms the first estimate of the glottal flow. Next, a gth order LPC analysis (block 7) is computed to get a parametric model of the spectral envelope of the estimate of the glottal flow. For this purpose, since the estimated flow was arrived at by removing (in block 5) the contribution of the vocal tract from speech, it is possible to use a more accurate, higher-order LPC analysis than in block no. 2 without obtaining models that exhibit spurious formant-like peaks. (The value of g is typically between 4 and 8.) The glottal contribution and the lip radiation are canceled out again through inverse filtering and integration (blocks 8 and 9), and the final estimate of the vocal tract (H_{vt2}) is obtained through pth order LPC. Finally, the effects of the vocal tract and the lip radiation are canceled out from the original speech signal, yielding the glottal flow signal $g(n)$. Examples of glottal flow signals estimated with the IAIF method are shown in Fig. 1.

In practice, the computation of the glottal volume velocity from the speech pressure signal is an inverse problem that is extremely difficult, if not impossible, to solve accurately, particularly in the case of continuous speech. Even for sustained phonations there are challenging types of utterances, such as high-pitch speech and nasals, for which the estimation of the glottal flow is known to be problematic (e.g., [44], [45]). In addition, a fundamental problem in trying to evaluate the accuracy of glottal inverse filtering from natural speech is the fact that one is unable to assess in detail how closely the obtained waveform corresponds to the true glottal flow because the latter cannot be measured non-invasively from the human larynx. Therefore, glottal inverse filtering methods are typically tested using synthetic speech (usually vowels) that has been created using a known, artificial waveform of the glottal excitation. This kind of evaluation, however, is not truly objective because speech synthesis and inverse filtering analysis are typically based on similar models of the human voice production apparatus, e.g., the source filter model [9]. In contrast to this, the accuracy of the IAIF method has been assessed recently using a different strategy based on physical modeling of the vocal folds and the vocal tract [46], [47]. This approach is different from ones where synthetic speech excited by an artificial form of the glottal excitation is used, because the glottal flow waveform results from the interaction of the self-sustained oscillation of the vocal folds with subglottal and supraglottal pressures. Results reported in [47] for four different vowels of ten different F_0 values indicate that IAIF yields satisfactory accuracy in the estimation of the glottal flow: the relative error between the original and the estimated flow was less than 10% in the far majority of the analyses when parameterization of the flow was conducted with a normalized amplitude quotient [48]. The estimation error in [47] however, increased when F_0 was raised above 400 Hz and was most severe when a high-pitch was combined with a low value of the first formant. In summary, given the benefits of the IAIF method described in section II. A and its previously evaluated satisfactory accuracy, it was considered justified to use the IAIF method as a computational tool for glottal inverse filtering in the present study.

III. Proposed Speech Synthesis System

A. Overview

The proposed speech synthesis system aims to produce high quality synthetic speech capable of conveying various styles of speaking, speaker characteristics, and emotions. To achieve this goal, the human voice production mechanism is modeled with the help of glottal inverse filtering embedded in an HMM framework. Automatic glottal inverse filtering is used in order to compute a parametric feature expression for the voice source and the vocal tract transfer function. The voice source and vocal tract features are then modeled with multi-stream HMMs. In the synthesis stage, natural glottal flow pulses are used to create the excitation signal for speech. This excitation signal is further modified to reproduce the time-varying changes in the natural voice source. The proposed novel procedure enables both synthesis of natural sounding speech and easy control over individual speech features that contribute to the perceived quality of speech [11], [49], [50].

The overview of the system is shown in Fig. 3. The system consists of two main stages: training and synthesis. In the training part, speech parameters computed by glottal inverse filtering are extracted from each utterance of a training speech database, and the obtained speech parameters are modeled in the framework of HMM. In the synthesis part, speech parameters are generated from the HMMs corresponding to the subword units used in a given input text, and speech waveform is synthesized using the generated speech parameters and natural glottal flow pulses. Although the proposed system is built on a basic framework of an HMM-based speech synthesis system [51], the novel parametrization and synthesis methods are significantly different from previous HMM-based synthesizers, and therefore they are explained in detail below.

B. Speech Parametrization

The parametrization stage attempts to compress the information of the speech signal into a few parameters that would describe the essential characteristics of the original speech signal as accurately as possible. The core of the parametrization method is the automatic glottal inverse filtering method IAIF [24], [32], which decomposes voiced speech into the glottal source signal and the vocal tract transfer function.

The flow chart of the speech parametrization algorithm is shown in Fig. 4. The speech is windowed with a rectangular window to 25-ms frames at 5-ms intervals and the parameters

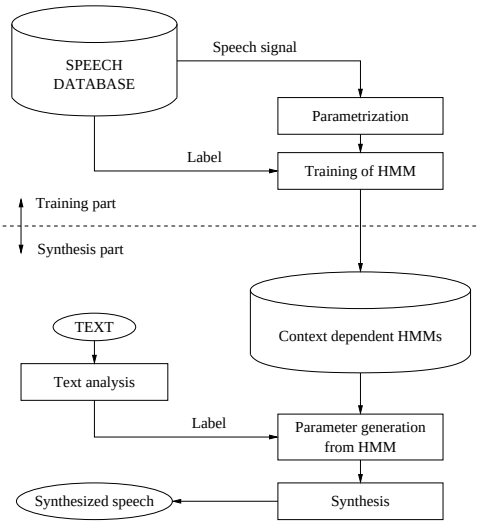


Fig. 3. Overview of the described system. The system consists of two main stages, training and synthesis. In the training stage, utterances of a speech database are parametrized with a glottal inverse filtering based method and trained in a framework of HMM. In the synthesis stage, speech parameters are generated according to the text input and speech is synthesized from the parameters.

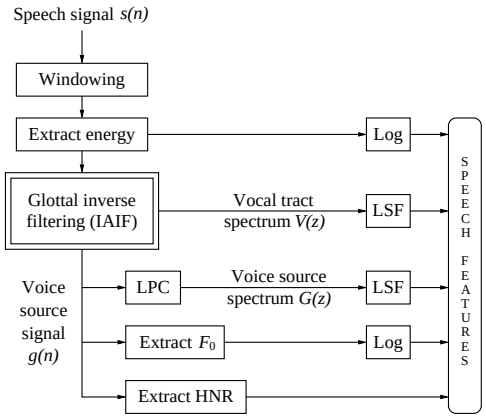


Fig. 4. Flow chart of the parametrization stage. The speech signal $s(n)$ is decomposed into the glottal source signal $g(n)$ and the all-pole model of the vocal tract $V(z)$ using the IAIF method. The glottal source signal is further parametrized into the all-pole model of the voice source $G(z)$, the fundamental frequency F_0, and the harmonic-to-noise ratio (HNR). The obtained parameters are converted to a suitable representation for the HMM system. For clarity, the parametrization of unvoiced speech segments, using basic LPC for the estimation of spectrum, is excluded from the flow chart.

are extracted from each frame. The extracted features are presented in Table I. First, the log-energy of the windowed speech is evaluated, after which glottal inverse filtering is performed. This decomposes a voiced speech signal into the glottal source signal $g(n)$ and the pth order all-pole model of the vocal tract $V(z)$. The order p is set to 30, slightly higher than the IAIF method basically requires for describing the formant structure. The use of higher-order all-pole model will alleviate the problem of coarticulation in continuous speech by adding more poles to represent the more complex spectra. Auditory analyses with the proposed synthesis technique have shown that slightly increasing the order p improves the quality of synthetic speech.

The spectrum of the glottal source signal is estimated with a 10th order LPC to capture the spectral properties of the excitation $G(z)$, mainly the spectral tilt, but also the more detailed spectral structure of the source. Several previous studies have utilized LPC-analysis for computing an estimate of the spectral tilt of the glottal flow. Some of these studies have used a very small LPC order, such as two [52] or three [11]. Results in [53], however, indicate that the spectral dynamics of the glottal flow cannot be modeled properly with LPC analysis of this small a prediction order. Auditory analyses with the proposed synthesis technique corroborated this finding, and, consequently, 10th order LPC analysis was selected to be used in modeling of the spectral decay of the glottal flow. Both obtained all-pole models are converted further to line spectral frequencies (LSF) [54], a parametric representation of LPC information well-suited to be used in a statistical HMM system [55], providing stability [56] and low spectral distortion [57].

The source signal estimated by the glottal inverse filtering reflects the acoustic excitation generated by the vibrating vocal folds. Therefore, the estimated glottal flow is used in the estimation of the fundamental frequency F_0. The autocorrelation method [58] is used to extract the fundamental frequency from the glottal source signal and the values are further converted to a logarithmic scale. The voiced-unvoiced decision is made based on the energy of the low frequency band (0–1 kHz) and the number of zero-crossings in the frame. The frames determined as unvoiced are marked as zeros. Additionally, a range of possible F_0 values is defined based on the speaker's F_0 range in order to reduce gross errors.

In order to capture the degree of voicing in the excitation, the harmonic-to-noise ratio (HNR) [59] of the glottal source signal is analyzed in four bands (0–2, 2–4, 4–6, 6–8 kHz). The HNR measure indicates the proportion of the periodic vibratory glottal excitation compared to the aperiodic noise excitation of the voice source. The HNR is determined by first evaluating the fast Fourier transform (FFT) of the windowed speech signal, and evaluating the cepstrum separately for each frequency band. The degree of the harmonicity is then indicated by the strength of the cepstral peak, whose location is defined by the fundamental period. Finally, the HNR is defined for each band as the ratio of the maximum value of the cepstral peak to the averaged value of other quefrencies of the cepstrum.

In case of unvoiced speech, conventional LPC is used to evaluate the spectral model of speech, though the inverse filtering is continuously performed in order to get the F_0 estimate. In unvoiced segments, F_0 and HNR are set to zero.

TABLE I
SPEECH FEATURES AND THE NUMBER OF PARAMETERS.

Feature	Parameters per frame
Fundamental frequency	1
Energy	1
Harmonic-to-noise ratio	4
Voice source spectrum	10
Vocal tract spectrum	30

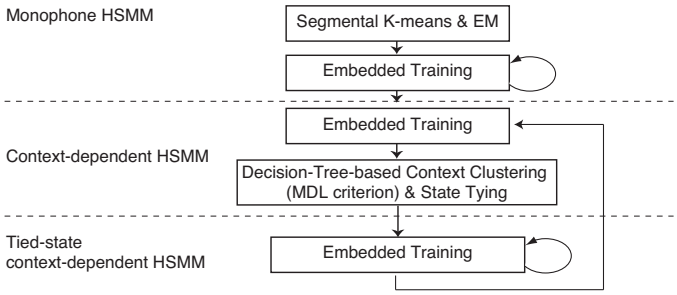


Fig. 5. Overview of the training stages of HMMs. First, monophone HSMMs are trained, converted into context-dependent HSMMs, and re-estimated. Then, decision-tree-based context clustering is applied to the HSMMs and the model parameters of the HSMMs are tied. Finally, the clustered HSMMs are re-estimated again.

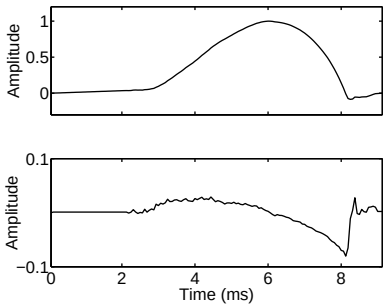


Fig. 6. Example of a library pulse and its time derivative. The pulse is extracted from a sustained vowel [a] produced by a male speaker using normal phonation. The closed phase of the original pulse derivative was forced to zero in order to remove minor fluctuations present in the waveform produced by the glottal inverse filtering.

The voice source spectrum is continuously extracted, although only the values for voiced segments are used in the synthesis stage.

C. HMM Training

The basic steps for the HMM training are similar to that in [26]. In order to model the extracted features together with their duration in a unified modeling framework, context-dependent multi-stream and multi-space distribution (MSD) [60] hidden semi-Markov models [61] (for short, MSD-HSMMs [62]) are utilized as acoustic units for speech synthesis. The multi-stream model structure is used for simultaneous and synchronous modeling of the extracted features. The MSD structure is used for statistical modeling of the fundamental frequency as mixture sequences of continuous real numbers for voiced regions and symbol strings for unvoiced regions. The rest of the features are modeled as continuous probability distribution (CD) streams. The harmonic-to-noise ratio and the voice source spectrum would also be natural candidates for MSD modeling, but in the following experiment CD was used and the parameters generated for unvoiced regions were simply omitted. The contexts used include not only phonetic information but also linguistic information such as morpheme features, accentual features, and even utterance-level features [63], [64].

Five-state left-to-right MSD-HSMMs without skip paths are used for all speech synthesis described in this paper. Each state has a single Gaussian probability distribution function (pdf) with a diagonal covariance matrix as the state output pdf and a single Gaussian pdf with a scalar variance as the state duration pdf.

The overview of the training process is illustrated in Fig. 5. First, monophone MSD-HSMMs are trained using the segmental K-means and expectation-maximization (EM) algorithms based on phonetic labels having time alignment information. These are converted into context-dependent MSD-HSMMs and the model parameters are re-estimated again. Then, decision-tree-based context clustering using a minimum description length (MDL) criterion [65] is applied to the HSMMs and the model parameters of the HSMMs at each leaf node of the decision trees are tied. The clustered MSD-HSMMs are then re-estimated. For the proposed system, only the vocal tract LSFs and F_0 features were considered during the alignment step of the re-estimation; the weights of the other streams were set to zero.

D. Parameter Generation from HMMs

In the parameter generation step, an input text is first transformed into a sequence of context-dependent phoneme labels. A sentence MSD-HSMM corresponding to the label sequence is then constructed by concatenating the parameter-tied context-dependent MSD-HSMMs. Then, a feature sequence trajectory is statistically generated from the sentence MSD-HSMM itself. Here the duration pdfs automatically determine the duration of each state of the sentence MSD-HSMM. A trajectory sequence that satisfies the global variance of the whole corpus [8] is used to generate the parameters.

E. Speech Synthesis

After the generation of the speech parameters from the HMM system, the speech waveform is synthesized from the parameters. There are two main differences in this synthesis stage compared to conventional synthesis methods. First, natural glottal flow pulses are used to create the voiced excitation signal, and second, the spectral properties of the excitation signal are modified with an adaptive infinite impulse response (IIR) filter with the aim of reproducing the time-varying changes in the real voice source and preserving the original voice quality.

At present, a single glottal flow pulse, called the library pulse, extracted from real speech with glottal inverse filtering, is used for the given speaker. The pulse is extracted from a glottal flow signal of a sustained vowel, and a pulse showing distinct open and closed phases is selected as a representative sample of the excitation. An example of a library pulse and its derivative is shown in Fig. 6. A pulse based on real speech is expected to be more natural than previous artificial excitations, since various natural properties are absent from previously available excitation models, especially from the simplest one: a train of impulses. For example, the temporal and spectral properties of the noise contained by the library pulse are a consequence of real physical movement of the vocal folds and

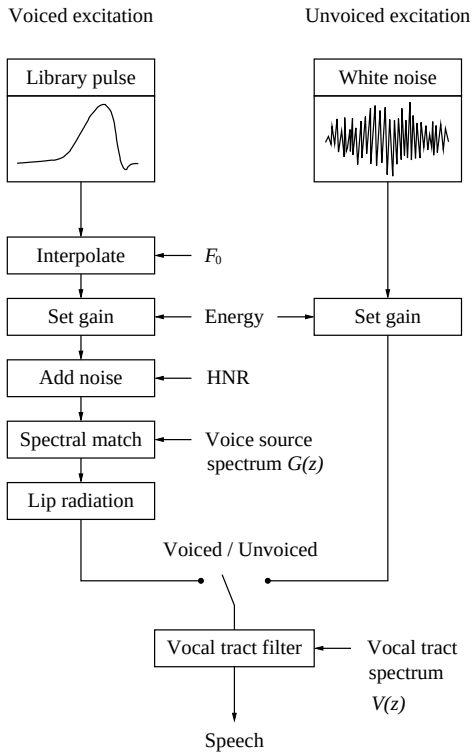


Fig. 7. Flow chart of the synthesis stage. The basis of the voiced excitation signal is a library glottal flow pulse, which is modified according to the voice source parameters. Unvoiced excitation is composed of white noise. The excitation signals are combined and filtered with the vocal tract filter $V(z)$ to generate speech.

the following turbulence in the glottis [66], which cannot be modeled with known techniques.

The flow chart of the synthesis stage is shown in Fig. 7. The excitation consists of voiced and unvoiced sound sources. The basis of the voiced sound source is the library pulse. This glottal flow pulse is interpolated in the time domain with a cubic spline interpolation algorithm [67], [68] in order to achieve a specific fundamental period, and the energy (gain) of the pulse is equalized to the energy measure given by the HMM.

Next, the harmonic-to-noise ratio of the pulse is measured by creating a frame that consists of pulses interpolated to the given F_0 and by using the same HNR estimation method as in the analysis phase. Then, noise is added separately to four bands (0–2, 2–4, 4–6, 6–8 kHz) in the frequency domain. The amount of noise in each band is determined by the ratio of the HNR measure of the pulse and the measure generated by the HMM system. Noise is created by evaluating the FFT of the pulse and adding a random component both to the real and imaginary parts of the FFT vector. The strength of the random component is determined by the band-wise HNR ratios.

Since the spectrum of the excitation generated by a single library pulse is constant over time and does not correspond to the target spectrum, the spectrum of the excitation needs to be modified. The target spectrum is the all-pole model of the glottal source $G(z)$ generated by the HMM. This spectrum is achieved by first evaluating the all-pole model of the library pulse with LPC (the order is equal to the order of $G(z)$), and then filtering the pulse train with an IIR filter constructed from these two all-pole models, compensating for the differences between the spectra.

The lip radiation effect is modeled with a fixed differentiator. The unvoiced excitation is white noise, the gain of which is determined by the energy measure generated by the HMM system. A formant enhancement procedure [69] is applied to the LSFs generated by the HMM system to compensate for the averaging effect of the statistical modeling. The LSFs are then interpolated and converted into LPC coefficients $V(z)$ and used to filter the excitation signal.

IV. Evaluation

The quality of the proposed system was compared to two alternative HMM-based speech synthesis systems for evaluation purposes. The first system uses the popular STRAIGHT vocoding technique, which has been shown to be able to generate high quality synthetic speech [70]. The second system uses an older vocoding technique, mel-cepstral analysis and synthesis [71] with a simple impulse train excitation model. In addition, natural speech samples were included in the test. In order to understand the differences between the systems and the underlying reasons for the possible differences in quality, the two reference systems are described next.

A. STRAIGHT-Based System

The speech manipulation tool STRAIGHT [25] mainly consists of F_0-adaptive spectral smoothing carried out in the time-frequency domain to remove signal periodicity, mixed excitation, and group delay manipulation [72]. The STRAIGHT-based system extracts and models three kinds of parameters required for the STRAIGHT vocoder with mixed excitation: spectral coefficients calculated from the smoothed spectra [26], fundamental frequency, and aperiodicity measures [72], [73].

In STRAIGHT analysis, speech signals are first high-pass filtered with a cut-off frequency of 70 Hz. They are then sampled at a rate of 16 kHz and windowed by an F_0-adaptive Gaussian window with a 5-ms shift. The vocal tract filter is estimated with a 39th-order mel-cepstrum [71]. In this system C_0 is used instead of log-energy. The extraction of F_0 values and voiced-unvoiced decision are performed by using voting of outputs of an instantaneous-frequency-amplitude-spectrum-based algorithm [74], a fixed-point analysis called TEMPO [75], and the ESPS get-F_0 tool [76], [77]. The aperiodicity measures for mixed excitation are based on a ratio between the lower and upper smoothed spectral envelopes [72], [73] and averaged across five frequency sub-bands (0–1, 1–2, 2–4, 4–6, and 6–8 kHz). The total number of speech parameters per frame is 46, which is exactly the same as in the proposed system.

In STRAIGHT synthesis, an excitation signal is generated using mixed excitation consisting of impulses and a noise component weighted by band-pass filtered aperiodicity parameters. At each frequency bin, the aperiodicity parameter is converted to the weight for a noise signal by using a sigmoid mapping function adopted in [73]. The pitch-synchronous overlap add (PSOLA) [78] method is used to reconstruct the excitation

signal, which is then used to excite a mel log spectrum approximation (MLSA) filter [71], [79], corresponding to the STRAIGHT mel-cepstral coefficients. The vocoder modules of the STRAIGHT-based system are the same as in [26].

B. Impulse Train Based System

The second reference for the experiment was the *de facto* standard HMM-based speech synthesizer, which is included in the current HTS (HMM-based Speech Synthesis System) release [51], [80]. This system uses mel-cepstral coefficients to model the spectrum of speech, and the excitation is modeled only by the fundamental frequency. Voiced speech is excited by an impulse train controlled by the fundamental frequency and unvoiced speech is excited by white noise. Synthesis from mel-cepstral coefficients is performed with the mel log spectrum approximation (MLSA) filter [71], [79]. This vocoder is known for its smooth speech quality, but also for its buzziness, inherent to the simple excitation scheme.

For the experiment, 25 mel-cepstral coefficients (including the zeroth coefficient for the gain) were extracted using a 25-ms Blackman window with a 5-ms frame shift. The log-F_0 values were estimated with the proposed system. The total number of parameters per frame is thus 26, substantially less than the other two systems.

While this vocoder could be considered outdated, it is a useful reference for the test, because its properties are widely known and the quality compared to the STRAIGHT-based system is documented [26].

C. System Configurations

In order to ensure comparability between the proposed and the two reference methods, identical linguistic and acoustic data were used to train each system. The same linguistic front-end was utilized to extract 67 contextual features for each phone in the training corpus, and the same question set was used to guide the model clustering.

The HMM configuration of the three systems was generally similar. However, the HMM training and parameter generation involves several tunable parameters, such as stream weights, MDL factors, and global variance factors, whose optimal values are dependent on the speech feature representation. Thus, several parameters related to the HMM system were optimized independently according to the requirements of each parametrization scheme.

D. Speech Material

The proposed method was tested by training all the systems with a prosodically annotated database of 600 phonetically rich Finnish utterances spoken by a 39-year-old Finnish male speaker, comprising approximately one hour of speech material [81]. The utterances were from two to eleven seconds of duration and they contained from two to 22 words, averaging 9.34 words. The number of total phone instances (including silences) was 43210. Ninety-two utterances held out from the database were used as stimuli for the evaluation.

Glottal inverse filtering is known to be sensitive to the recording conditions, especially to the phase response of the microphone [33]. Although phase information is not crucial in the proposed method (only the magnitude spectrum is used), a high quality recording was performed to ensure correct glottal inverse filtering. The recordings were done directly to a computer hard drive in an anechoic chamber using a high-quality condenser microphone (AKG CK 61-ULS). The speech was sampled at 16 kHz.

E. Test Setup

Three separate subjective listening tests were conducted to assess the performance of the proposed system compared to other speech synthesizers. All the tests were carried out in an acoustically modified multipurpose room with low background noise level. Each listener performed the tests individually using a graphical user interface on a computer terminal. Test samples were played to both ears with high-quality headphones (Sennheiser HD580). The three tests took approximately one hour on average per person.

Fifteen listeners (13 males, 2 females) with no known hearing impairment participated in the test. The listeners were native Finnish speakers between 23 and 35 years of age, averaging 28 years. All the listeners were graduate or post-graduate students working in the field of acoustics or signal processing, but not in speech synthesis.

The test method used for the first two tests was similar to the comparison category rating (CCR) test described in [82]. In this test type, a listener is presented with a pair of speech samples for each trial. A sample pair consists of a sentence synthesized with two different methods (or natural speech). The task of the listener is to assess the quality of the second sample compared to the first one on the comparison mean opinion score (CMOS) scale, which is a discrete seven-point scale ranging from much worse (-3) to much better (3). The listener responses in a CCR test can be summarized concisely by calculating the mean score for each evaluated method. The mean yields the order of preference and distances between all the methods (i.e., the amount of preference relative to each other), but the numerical values of the mean scores do not have explicit meaning.

In the first test, the CCR method was used to assess the quality of the proposed method in comparison to natural speech and synthetic speech generated by the STRAIGHT-based system. The purpose of this test was to evaluate the overall performance of the two synthesis methods compared to natural speech. Ten randomly chosen sentences from held-out data were used to generate the test samples for each method. All possible combinations of the three sample types were assessed in both orders for each sentence. Ten null pairs with identical samples were also included. Altogether, the test consisted of 70 cases.

Second, a similar CCR test was performed, but the natural speech samples were replaced with synthetic speech generated by the system using impulse train excitation. The purpose of this test was to evaluate the overall performance of the proposed method compared to two well known speech synthesizers. Again, ten randomly chosen sentences from held-out data were used to generate the test samples for each method and a total of 70 test cases were assessed.

The pair comparison test method was used in the third test. In each test case, the subjects listened to a pair of samples and selected the one they preferred. They also had the option no preference between the two samples. The listeners could listen to the samples any number of times before making their choice. Only the synthetic speech samples generated by the proposed system and the STRAIGHT-based system were involved in the pair comparison test. The purpose of this test was to evaluate the general preference between the two high-quality speech synthesis systems. Ten randomly chosen sentences from held-out data were used to generate the test samples for each method. A total of 24 speech sample pairs were assessed.

After the listening tests, the listeners were asked to report and describe possible artifacts they noticed in the samples in order to obtain information about the most salient aspects that degrade the quality of synthetic speech.

It is worth noticing that the prosody of the synthesized utterances was not normalized between the test samples. Normalization of the fundamental frequency and durations was considered, but since there were various features in each system, normalization of all differences between the systems would not have been possible. The prosody of natural speech was clearly different from that of the synthesized utterances. The synthetic speech samples resembled each other highly in terms of prosody.

Finally, the intelligibility of the proposed and the STRAIGHT-based system was evaluated using a standard semantically unpredictable sentence (SUS) intelligibility test [83]. A total of 41 native Finnish speakers participated in a web-based listening test containing 30 semantically unpredictable sentences. In the test, subjects listened the sentences and typed in what they heard. The letter error rate (LER) was calculated from the results and the Wilcoxon signed rank test [84] was used to evaluate the statistical significance.

F. Results

The result of the first test, the ranking of natural speech and synthetic speech generated by the proposed and the STRAIGHT-based system, is shown in Fig. 8. The figure shows that the quality of natural speech is still superior compared to synthetic speech. This large difference results partly from the degraded naturalness and occasional artifacts in the synthetic speech samples, but also from the prosodic discrepancies between the synthetic and natural speech samples. It is worth emphasizing that in this experiment, where the prosody was not normalized, natural speech was considered far superior compared to either of the speech synthesis systems, but despite that, the difference between the proposed system and the STRAIGHT-based system was statistically significant.

The result of the second CCR test is shown in Fig. 9. The three speech synthesis methods, the proposed system, the STRAIGHT-based system, and the impulse train based system, are ranked according to the evaluation. The figure shows a clear preference for the proposed method over the two other methods. The STRAIGHT-based system is also considered clearly better than the impulse train based system. The result

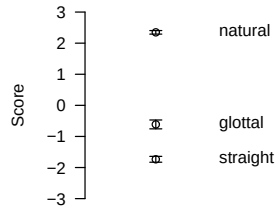


Fig. 8. Ranking of the first CCR test for the following speech samples: natural speech (natural), proposed system (glottal), and STRAIGHT-based system (straight). The mean score has no explicit meaning, but the distances between the scores define the amount of preference relative to each other. The 95% confidence intervals are presented for each score.

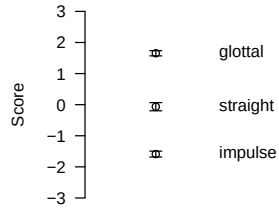


Fig. 9. Ranking of the second CCR test for the following speech samples: proposed system (glottal), STRAIGHT-based system (straight), and impulse train based system (impulse). The mean score has no explicit meaning, but the distances between the scores define the amount of preference relative to each other. The 95% confidence intervals are presented for each score.

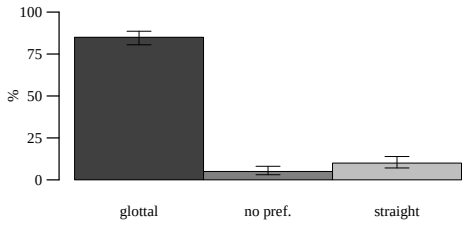


Fig. 10. Results of the pair comparison test applied for the proposed system (glottal) and the STRAIGHT-based system (straight). The bars indicate the percentage of the total number of answers to the question “Which one would you rather listen to?”. The center bar (no pref.) indicates no preference for either of the methods. The 95% confidence intervals are presented for each bar.

can also be roughly interpreted so that the improvement in quality from the STRAIGHT-based system to the proposed system is as significant as the improvement from the impulse train based system to the STRAIGHT-based system.

The result of the third test is shown in Fig. 10, which shows that the proposed system is almost always preferred over the STRAIGHT-based system, answering the question “Which one would you rather listen to?”.

The listeners described the proposed system as smooth, warm, natural sounding, and having clear characteristics of a person and showing some emotion, but it was also criticized for having some artifacts near consonants. The STRAIGHT-based system was described as very clean, but also slightly artificial and nasal, and some listeners compared it to band-limited speech. The system using impulse train excitation was mostly described as artificial and unnatural. The prosody of all the synthetic methods was criticized, but all the methods were

assessed to have approximately the same amount of errors in prosody.

The standard SUS test showed that the intelligibility of the proposed system was better than that of the STRAIGHT-based system. The letter error rate was 1.6% and 2.6% for the proposed and the STRAIGHT-based method, respectively. The difference between the systems was statistically significant ($p < 0.0001$).

A representative set of test samples is available online at http://www.helsinki.fi/speechsciences/synthesis/samples.html.

V. Discussion

Over the years, speech synthesis has largely been based on the use of source-filter models, in which the production of speech is represented in terms of the excitation and filter characteristics. The separation of speech into these major components has been mainly conducted by utilizing techniques based on impulse-type waveforms for the excitation. This is greatly different from the functioning of the real human voice production mechanism, in which the excitation is represented by the glottal volume velocity waveform, a smooth quasi-periodic signal generated by the fluctuation of two physiological organs, the human vocal folds. The artificial impulse-type excitation used in typical source-filter models is also extensively different in terms of its spectral behavior when compared to its real, physiologically-oriented counterpart; the spectral envelope of the impulse-train is always flat while that of the real glottal excitation varies depending on, for example, the phonation type and vocal intensity of the spoken message. One can argue that, in speech synthesis, the use of speech models unable to model properly the function of the real human voice production, such as the impulse-based excitation, may lead to a loss of valuable information and limit the flexibility of the synthesizer.

Utilizing glottal inverse filtering-based modeling of the human voice production mechanism in HMM-based synthesis is attractive because it also provides, in comparison to traditional signal models, a better correspondence between the synthetic speech and its underlying physiological parameters. Although the imitation of the human vocal apparatus itself may not have intrinsic value, this approach could provide several benefits. For example, through the decomposition of speech according to a glottal inverse filtering-based model, each component, representing a specific physical phenomenon and thus representing specific features of speech, can be individually adapted or modified based on the knowledge of the speech production mechanism [85]. In a similar way to [85], this approach would enable easy control over the features that can be directly attributed to the perceived voice characteristics by modifying the glottal source signal or the vocal tract transfer function.

The proposed novel method addresses two important issues in parametric speech synthesis. First, the concept of separating the voice source from the vocal tract filter in HMM-based speech synthesis is expected to make speech synthesis more flexible and natural compared to conventional methods. Second, the utilization of a natural glottal flow pulse addresses the problem of voice source modeling by preserving some of the detailed structure of the natural excitation, which cannot be easily modeled. However, since the current system uses only a single glottal flow pulse for an utterance, the natural variation in the excitation from one pulse to another cannot be well reproduced. The utilization of multiple pulses is a future direction of this work.

This study demonstrates that the modeling of the real speech production mechanism through the utilization of glottal inverse filtering in HMM-based speech synthesis can improve the quality of synthetic speech. The formal evaluation with one male speaker, expected to be easiest for the glottal inverse filtering, showed that the proposed method can synthesize very natural sounding speech. Informal experiments on the proposed system with several speakers, male and female, have yielded encouraging results[1]. Based on the preliminary results and the theoretical benefits of the method, it is expected that high-quality synthetic speech in different speaking styles and with different speaker characteristics can be reproduced. In order to thoroughly test the performance of the proposed synthesis technique, one should also use expressive speech, where the spectral dynamics of the voice source are expected to be larger. Another direction for future research is speaker and speaking style adaptation, where the proposed method could provide substantive improvements compared to traditional methods.

VI. Conclusions

In this paper, a new HMM-based text-to-speech system utilizing glottal inverse filtering was described. The study presented a method to extract and model individual parameters for the voice source and the vocal tract, and a method to reconstruct a realistic voice source from the parameters using real glottal flow pulses. These novel procedures enable the generation of high quality synthetic speech. Subjective listening tests showed that the proposed method is able to generate highly natural synthetic speech, and the quality of the proposed system is considerably better compared to two other commonly used HMM-based speech synthesizers.

Acknowledgment

The authors would like to thank Professor Hideki Kawahara of Wakayama University for permission to use the STRAIGHT vocoding method. We also thank Mr. Vasilis Karaiskos of University of Edinburgh for his contribution to our listening tests. We also thank the three anonymous reviewers for their constructive feedback and helpful suggestions.

References

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. Eurospeech*, Sep. 1999, pp. 2374–2350.

[2] K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi, and S. Imai, "An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features," in *Proc. Eurospeech*, vol. 1, 1995, pp. 757–760.

[1] Audio examples for the additional speakers are also available online via the URL mentioned before.

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

[4] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, 2009, (In Press).

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

[6] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing," *IEICE Trans. Inf. & Syst.*, vol. E88-D, no. 11, pp. 2484–2491, Nov. 2005.

[7] Y.-J. Wu and R.-H. Wang, "Minimum generation error training for HMM-based speech synthesis," in *Proc. ICASSP*, vol. 1, 2006, pp. 889–892.

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

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

[10] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. America*, vol. 87, no. 2, pp. 820–857, Feb. 1990.

[11] D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *J. Acoust. Soc. America*, vol. 90, no. 5, pp. 2394–2410, Nov. 1991.

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

[13] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice-Hall, 1978.

[14] 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.

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

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

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

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

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

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

[21] ——, "Glottal spectral separation for parametric speech synthesis," in *Proc. Interspeech*, 2008, pp. 1829–1832.

[22] 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.

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

[24] 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.

[25] 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.

[26] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of the Nitech HMM-based speech synthesis for Blizzard Challenge 2005," *IEICE Trans. Inf. & Syst.*, vol. E90-D, pp. 325–333, 2007.

[27] V. Karaiskos, S. King, R. A. J. Clark, and C. Mayo, "The Blizzard Challenge 2008," in *Blizzard Challenge Workshop*, 2008.

[28] T. Raitio, "Hidden Markov model based Finnish text-to-speech system utilizing glottal inverse filtering," Master's thesis, Helsinki University of Technology, Finland, 2008.

[29] 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.

[30] R. L. Miller, "Nature of the vocal cord wave," *J. Acoust. Soc. America*, vol. 31, no. 6, pp. 667–677, Jun. 1959.

[31] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*, 2nd ed. Springer-Verlag, 1972.

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

[33] D. Wong, J. Markel, and A. Gray Jr., "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 27, no. 4, pp. 350–355, Aug. 1979.

[34] H. Strube, "Determination of the instant of glottal closure from the speech wave," *J. Acoust. Soc. America*, vol. 56, no. 5, pp. 1625–1629, 1974.

[35] M. Plumpe, T. Quatieri, and D. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 5, pp. 569–585, 1999.

[36] M. Fröhlich, D. Michaelis, and H. Strube, "SIM – simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals," *J. Acoust. Soc. America*, vol. 110, no. 1, pp. 479–488, 2001.

[37] O. Akande and P. Murphy, "Estimation of the vocal tract transfer function with application to glottal wave analysis," *Speech Commun.*, vol. 46, pp. 15–36, 2005.

[38] Q. Fu and P. Murphy, "Robust glottal source estimation based on joint source-filter model optimization," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 492–501, 2006.

[39] D. Childers and C. Ahn, "Modeling the glottal volume-velocity waveform for three voice types," *J. Acoust. Soc. America*, vol. 97, no. 1, pp. 505–519, 1995.

[40] A. Krishnamurthy and D. Childers, "Two-channel speech analysis," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 730–743, 1986.

[41] H. Strik and L. Boves, "On the relation between voice source parameters and prosodic features in connected speech," *Speech Commun.*, vol. 11, pp. 167–174, 1992.

[42] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *IEEE Trans. Signal Processing*, vol. 39, no. 2, pp. 411–423, Feb. 1991.

[43] J. Makhoul, "Linear prediction: A tutorial review," *Proc. of the IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.

[44] D. Veeneman and S. BeMent, "Automatic glottal inverse filtering from speech and electroglottographic signals," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 369–377, 1985.

[45] J. Walker and P. Murphy, "Advanced methods for glottal wave extraction," in *Nonlinear Analyses and Algorithms for Speech Processing*, M. Faundez-Zanuy *et al.*, Eds. Springer Berlin/Heidelberg, 2005, pp. 139–149.

[46] P. Alku, J. Horáček, M. Airas, F. Griffond-Boitier, and A.-M. Laukkanen, "Performance of glottal inverse filtering as tested by aeroelastic modelling of phonation and FE modelling of vocal tract," *Acta Acustica united with Acustica*, vol. 92, pp. 717–724, 2006.

[47] P. Alku, B. Story, and M. Airas, "Estimation of the voice source from speech pressure signals: Evaluation of an inverse filtering technique using physical modelling of voice production," *Folia Phoniatrica et Logopaedica*, vol. 58, no. 2, pp. 102–113, 2006.

[48] P. Alku, T. Bäckström, and E. Vilkman, "Normalized amplitude quotient for parameterization of the glottal flow," *J. Acoust. Soc. America*, vol. 112, no. 2, pp. 701–710, 2002.

[49] C. Gobl and A. Ní Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Commun.*, vol. 40, no. 1-2, pp. 189–212, 2003.

[50] A.-M. Laukkanen, E. Vilkman, P. Alku, and H. Oksanen, "On the perception of emotions in speech: the role of voice quality," *Logopedics Phoniatrics Vocology*, vol. 22, no. 4, pp. 157–168, 1997.

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

[52] D. Klatt, "Review of text-to-speech conversion for English," *J. Acoust. Soc. America*, vol. 82, no. 3, pp. 737–793, 1987.

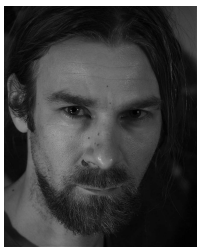
[53] J. Deller, "On the time domain properties of the two-pole model of the glottal waveform and implications for LPC," *Speech Commun.*, vol. 2, no. 1, pp. 57–63, 1983.

[54] F. K. Soong and B.-H. Juang, "Line spectrum pair (LSP) and speech data compression," in *Proc. ICASSP*, vol. 9, Mar. 1984, pp. 37–40.

[55] 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).

[56] F. Soong and B.-H. Juang, "Line spectrum pair (LSP) and speech data compression," in *Proc. ICASSP*, vol. 9, Mar. 1984, pp. 37–40.
[57] K. Paliwal and W. Kleijn, "Quantization of LPC parameters," in *Speech Coding and Synthesis*, W. Kleijn and K. Paliwal, Eds. Elsevier, 1995, ch. 12.
[58] L. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 25, no. 1, pp. 24–33, 1977.
[59] P. Murphy, "Perturbation-free measurement of the harmonics-to-noise ratio in voice signals using pitch synchronous harmonic analysis," *J. Acoust. Soc. America*, vol. 105, no. 5, pp. 2866–2881, 1999.
[60] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Trans. Inf. & Syst.*, vol. E85-D, no. 3, pp. 455–464, Mar. 2002.
[61] S. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition," *Computer Speech and Language*, vol. 1, no. 1, pp. 29–45, 1986.
[62] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 825–834, May 2007.
[63] M. Vainio, A. Suni, and P. Sirjola, "Accent and prominence in Finnish speech synthesis," in *Proc. of the 10th International Conference on Speech and Computer (Specom 2005)*, G. Kokkinakis, N. Fakotakis, E. Dermatos, and R. Potapova, Eds. University of Patras, Greece, Oct. 2005, pp. 309–312.
[64] A. Suni and M. Vainio, "Deep syntactic analysis and rule based accentuation in text-to-speech synthesis," in *TSD '08: Proc. of the 11th International Conference on Text, Speech and Dialogue*, 2008, pp. 535–542.
[65] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *J. Acoust. Soc. Japan (E)*, vol. 21, pp. 79–86, Mar. 2000.
[66] F. Alipour and R. Scherer, "Characterizing glottal jet turbulence," *J. Acoust. Soc. America*, vol. 119, no. 2, pp. 1063–1073, 2006.
[67] G. Engeln-Müllges and E. Uhlig, *Numerical Algorithms with C*. Berlin: Springer, 1996.
[68] M. Galassi *et al.*, *GNU Scientific Library Reference Manual*, 3rd ed., 2009.
[69] Z.-H. Ling, Y. Wu, Y.-P. Wang, L. Qin, and R.-H. Wang, "USTC system for Blizzard Challenge 2006: an improved HMM-based speech synthesis method," in *Blizzard Challenge Workshop*, 2006.
[70] H. Zen and T. Toda, "An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005," in *Proc. Interspeech*, Sep. 2005, pp. 93–96.
[71] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *Proc. ICASSP*, vol. 8, Apr. 1983, pp. 93–96.
[72] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," in *2nd International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, Sep. 2001.
[73] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," in *Proc. Interspeech*, Sep. 2006, pp. 2266–2269.
[74] D. Arifianto, T. Tanaka, T. Masuko, and T. Kobayashi, "Robust F0 estimation of speech signal using harmonicity measure based on instantaneous frequency," *IEICE Trans. Inf. & Syst.*, vol. E87-D, no. 12, pp. 2812–2820, Dec. 2004.
[75] H. Kawahara, H. Katayose, A. Cheveigné, and R. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity," in *Proc. Eurospeech*, Sep. 1999, pp. 2781–2784.
[76] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. Kleijn and K. Paliwal, Eds. Elsevier, 1995, pp. 495–518.
[77] *ESPS Programs Version 5.0*, Entropic Research Laboratory Inc, 1993.
[78] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Commun.*, vol. 9, no. 5–6, pp. 453–467, Dec. 1990.
[79] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP*, vol. 1, 1992, pp. 137–140.
[80] HTS, "HMM-based speech synthesis system," Apr. 2009. [Online]. Available: http://hts.sp.nitech.ac.jp
[81] M. Vainio, "Artificial neural network based prosody models for Finnish text-to-speech synthesis," Ph.D. dissertation, University of Helsinki, Finland, Dec. 2001.
[82] ITU, "Methods for subjective determination of transmission quality," International Telecommunication Union, Recommendation ITU-T P.800, Aug. 1996.
[83] C. Benoît, M. Grice, and V. Hazan, "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences," *Speech Commun.*, vol. 18, no. 4, pp. 381–392, 1996.
[84] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics*, vol. 1, pp. 80–83, 1945.
[85] Z.-H. Ling, K. Richmond, J. Yamagishi, and R.-H. Wang, "Articulatory control of HMM-based parametric speech synthesis driven by phonetic knowledge," in *Proc. Interspeech*, Brisbane, Australia, Sep. 2008, pp. 573–576.

Tuomo Raitio was born in 1983. He received the M.Sc. degree in telecommunication technology in 2008 from Helsinki University of Technology (TKK), Espoo, Finland. Since graduation he has continued with post-graduate research at the Department of Signal Processing and Acoustics at TKK. His research is concentrated on voice source modeling in HMM-based speech synthesis.



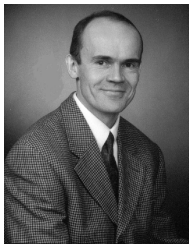
Antti Suni entered the field of speech technology in 2002, working in commercial speech recognition. Since 2005 he has been a project researcher at the Department of Speech Sciences, University of Helsinki. His main research interests are prosody modeling and HMM-based speech synthesis.

Junichi Yamagishi received the B.E. degree in computer science, M.E. and Dr. Eng. degrees in information processing from Tokyo Institute of Technology, Tokyo, Japan, in 2002, 2003, and 2006, respectively. He pioneered the use of speaker adaptation techniques in HMM-based speech synthesis in his doctoral dissertation *Average-voice-based speech synthesis*, which won the Tejima Doctoral Dissertation Award 2007. He held a research fellowship from the Japan Society for the Promotion of Science (JSPS) from 2004 to 2007. He was an intern researcher at ATR spoken language communication Research Laboratories (ATR-SLC) from 2003 to 2006. He was a visiting researcher at the Centre for Speech Technology Research (CSTR), University of Edinburgh, U.K. from 2006 to 2007. He is currently a senior research fellow at the CSTR, University of Edinburgh, and continues the research on the speaker adaptation for HMM-based speech synthesis in an EC FP7 collaborative project called the *EMIME* project (www.emime.org). His research interests include speech synthesis, speech analysis, and speech recognition. He is a member of ISCA, IEICE, and ASJ.

Hannu Pulakka was born in Finland in 1978. He received the M.Sc. degree in computer science and engineering from Helsinki University of Technology (TKK), Espoo, Finland, in 2005. He is conducting postgraduate research and pursuing the D.Sc. (Tech.) degree at the Department of Signal Processing and Acoustics at TKK. His research interests include speech and audio signal processing and speech enhancement.

Jani Nurminen received his M.Sc. degree from Department of Information Technology, Tampere University of Technology, Finland, in 2001. He has worked on speech related technologies since 1999, first in Tampere University of Technology until 2002, and after that in Nokia. He has authored or co-authored about 40 research publications and has over 30 granted or pending patents. Currently, he is a Technology Manager with Nokia Devices R&D. His research interests include speech synthesis, speech and audio processing, language processing, data compression, and multimodal user interfaces.

Martti Vainio received his Ph.D. degree in phonetics from the University of Helsinki, Finland, in 2001. He is currently a Finnish Academy Research Fellow at the Department of Speech Sciences, University of Helsinki. He is an Adjunct Professor in both language technology and general phonetics at the University of Helsinki. He has led research projects ranging from basic phonetics to speech technology and language resources. He is currently one of the leading experts in text-to-speech synthesis in Finland and has published widely in phonetics and speech technology. He is an active member in scientific review committees for conferences and journals related to acoustics, phonetics, and linguistics.



Paavo Alku received the M.Sc, Lic.Tech, and Dr.Sc. (Tech.) degrees from Helsinki University of Technology, Espoo, Finland in 1986, 1988, and 1992, respectively. He was Assistant Professor with the Asian Institute of Technology, Bangkok, Thailand, in 1993, and the University of Turku, Finland, in 1994–1999. Currently, he is professor of speech communication technology at Helsinki University of Technology. His research interests are analysis and parametrization of speech production, spectral modeling of speech, speech enhancement and cerebral processing of speech.