Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 77 (2016) 65–83

www.elsevier.com/locate/specom

Time-domain deterministic plus noise model based hybrid source modeling for statistical parametric speech synthesis

N.P. Narendra*, K. Sreenivasa Rao

School of Information Technology, Indian Institute of Technology Kharagpur, Kharagpur West Bengal 721302, India

Received 8 April 2015; received in revised form 6 December 2015; accepted 9 December 2015
Available online 23 December 2015

Abstract

This paper proposes a time-domain deterministic plus noise model based hybrid source modeling framework for improving the quality of statistical parametric speech synthesis system. In the proposed approach, the excitation signal is modeled as a combination of deterministic and noise components. Time-domain pitch-synchronous analysis is performed on the excitation or residual signal. From the pitch-synchronous residual frames of a phone, the deterministic and noise components are estimated. The deterministic components of all phones are systematically arranged in the form of a decision tree. The spectrum and amplitude envelope of noise components are modeled using hidden Markov models (HMMs). During synthesis, the suitable deterministic component is chosen from the leaf of a decision tree. The noise component is obtained after imposing the target spectrum and amplitude envelopes generated from the HMMs. The sum of deterministic and noise components are pitch-synchronously overlap added to construct the excitation signal of a phone. The proposed hybrid source modeling approach is incorporated in the statistical parametric speech synthesis system. Performance evaluation results show that the proposed method is capable producing natural sounding synthetic speech and the quality is clearly better than the state-of-the-art statistical parametric speech synthesis systems. Synthesized speech samples of the proposed and the state-of-the-art methods used for the comparison are made available online at http://www.sit.iitkgp.ernet.in/~ksrao/HSM-SPSS/hsm.html.

Keywords: Statistical parametric speech synthesis; Deterministic plus noise model; Hybrid source modeling; Pitch-synchronous residual frame.

1. Introduction

Statistical parametric speech synthesis (SPSS) based on hidden Markov models (HMMs) is growing in popularity over the past few years (Tokuda et al., 2013). In statistical parametric speech synthesis, speech can be parameterized using different approaches such as harmonic plus noise model, sinusoidal model and source-filter representations. In this work, speech is parameterized based on source-filter representation. The source refers to the excitation signal produced due to the vibration of vocal folds while the filter refers to the cascade of resonators realized by the shape of the vocal tract system. The vocal tract filter and the excitation signal are estimated directly from the speech signal. The spectral envelope of speech represents the vocal tract filter. The excitation signal is obtained by removing the contribution of the spectral envelope from the speech signal by inverse filtering. The spectral envelope of speech and the excitation signal are parameterized and modeled by the HMMs in a unified framework. Improper modeling of the spectral envelope of speech and the excitation signal results in the generation of erroneous parameters. Speech synthesized using these erroneous parameters leads to unnaturalness, and hence results in degradation of overall voice quality. This paper aims at developing an efficient method for representing and modeling the excitation signal.

In literature, several source modeling approaches have been proposed for improving the quality of SPSS. One of the initial approaches to model the excitation was reported by Yoshimura et al. (2001). It consists of modeling the excitation parameters used in Mixed Excitation Linear Prediction (MELP) (McCree et al., 1996) algorithm by hidden Markov models. During synthesis, the generated excitation

* Corresponding author. Tel.: +91 9547144956.
E-mail addresses: narendrasince1987@gmail.com, narendranp666@gmail.com (N.P. Narendra), ksrao@iitkgp.ac.in (K. Sreenivasa Rao).

http://dx.doi.org/10.1016/j.specom.2015.12.002
0167-6393/

parameters were used to construct the mixed excitation in the same way as in MELP algorithm. Later, Zen et al. used speech transformation and representation using adaptive interpolation of the weighted spectrum (STRAIGHT) (Kawahara et al., 1999) based source model for generating excitation signal (Zen et al., 2007). They modeled F0 and aperiodicity parameters by HMMs in order to enable the generation of excitation signal during synthesis stage (Zen et al., 2007). In Maia et al. (2007), the excitation signal is constructed by state-dependent filtering of pulse trains and white noise sequences. During training, filters and pulse trains are jointly optimized through a procedure that resembles analysis-by-synthesis speech coding algorithms. Liljencrants-Fant (LF) model is used to generate glottal source signal in SPSS (Cabral et al., 2011). The LF parameters are modeled by HMMs, and during synthesis, the generated LF parameters are used to control the glottal pulse shape. In Raitio et al. (2011b), the excitation signal is constructed by modifying a single natural instance of glottal flow pulse according to the source parameters generated by HMM. The glottal flow pulse is obtained by iterative adaptive glottal inverse filtering (Alku, 1992). In Cabral (2013), the excitation signal is generated as a combination of two segments. The first segment is only a small fraction of real residual frame around glottal closure instant and the second part is obtained from the model generated source parameters that represent the amplitude envelope and the energy of the residual waveform. In Wen et al. (2013), pitch-scaled spectrum is used to derive the periodic and aperiodic parts of the excitation signal. The periodic spectrum is compressed to reduce the dimensionality, and the aperiodic measure is fitted to a sigmoid function for integration into SPSS.

Instead of using the parameters derived from statistical models for generating the excitation signal, a hybrid approach of utilizing the real excitation segments was followed in Drugman et al. (2009a); Raitio et al. (2011a); Drugman and Dutoit (2012) and Drugman and Raitio (2014). In Drugman et al. (2009a), a codebook of pitch-synchronous residual frames was constructed. The source signal was generated by selecting suitable residual frames from the codebook based on target residual specification. Raitio et al. (2011a) proposed unit selection method to select appropriate glottal source pulses from the database based on target and concatenation costs. Drugman and Dutoit proposed a hybrid approach based on deterministic plus stochastic model (DSM) (Drugman and Dutoit, 2012). The excitation signal is divided into two bands delimited by a maximum voiced frequency. The deterministic component is the first eigenvector obtained by Principal Component Analysis (PCA) of low-frequency components of residual frames. The stochastic component is the spectrum and the amplitude envelope modulated white Gaussian noise. The spectrum and the amplitude envelopes are obtained from high-pass filtered residual frames. Instead of using fixed maximum voiced frequency, DSM-based source model is enhanced by using time-varying maximum voiced frequency (Drugman and Raitio, 2014).

Recently proposed hybrid approaches have shown a better quality of synthesized speech than the parametric approach of source modeling. In most of these hybrid approaches, phone specific characteristics of residual frames are not thoroughly explored in modeling and generation of excitation signal (Drugman et al., 2009a; Raitio et al., 2011a; Drugman and Dutoit, 2012). In Drugman and Dutoit (2012), certain preliminary studies have been performed to check the phone dependency of residual frames. Even though there are variations in the residual frames across different phonetic classes, it is grossly concluded that these variations are not significant. Our intuition is that the quality of synthesized speech may be more natural, if phone specific characteristics of excitation signal are incorporated in the synthesized speech. In this paper, phone dependent characteristics of excitation signal are analyzed and a hybrid approach of source modeling capable of generating excitation signal specific to phones is developed. In the proposed approach, the excitation signal of a phone is modeled based on time-domain deterministic plus noise model. The pitch-synchronous residual frames of a phone are modeled as a sum of deterministic and noise components. The deterministic components estimated from all phones are systematically arranged in the form of a decision tree. The noise components are parameterized in terms of spectral and amplitude envelopes. During synthesis, for the given input phone, the suitable deterministic component is chosen from the leaf of the decision tree. The noise component is generated by imposing the target spectral and amplitude envelopes on the natural instance of the noise signal. Phone specific excitation signal results in the generation of synthetic speech close to natural quality.

This paper is organized as follows. Section 2 provides the detailed description of proposed hybrid source modeling approach. This section describes the analysis of deterministic and noise components for different phonetic classes. The steps involved in modeling and generation of excitation signal under the time-domain deterministic plus noise model framework are also explained in Section 2. Section 3 provides the description of SPSS system incorporating the proposed source model. Evaluation of the proposed system is performed in Section 4 by comparing with three other state-of-the-art excitation generation methods. The significance of proposed method and relevance of the evaluation results are discussed in Section 5. Finally, Section 6 provides the summary of the contributions of the paper and presents some guidelines for future work.

2. Time-domain deterministic plus noise model of excitation signal

Speech is obtained by filtering excitation signal through time-varying vocal-tract system response. The source-filter model of speech production is given by

$$s(t) = e(t) * v(t). \tag{1}$$

where $s(t)$ is the speech signal, $e(t)$ is the excitation signal and $v(t)$ is the impulse response of vocal tract system. The excitation signal is pitch-synchronously decomposed in time-domain into a number of residual frames. The residual