A DETERMINISTIC PLUS NOISE MODEL OF EXCITATION SIGNAL USING PRINCIPAL COMPONENT ANALYSIS FOR PARAMETRIC SPEECH SYNTHESIS

N. P. Narendra, and K. Sreenivasa Rao

Indian Institute of Technology Kharagpur, Kharagpur - 721302, West Bengal, India

ABSTRACT

This paper proposes a new approach of modeling the excitation signal as deterministic and noise components. Initially, a study on characteristics of excitation or residual signal around glottal closure instant (GCI) is performed using principal component analysis (PCA). Based on the study, the segment of residual signal around GCI is considered as the deterministic component and the remaining part of the residual signal is considered as the noise component. The deterministic component is parameterized using PCA coefficients, and the noise component can be represented in terms of spectral and amplitude envelopes. The proposed excitation modeling approach is incorporated in the HMM-based speech synthesis system. Subjective evaluation results show a significant improvement in the quality of speech synthesized by the proposed method, compared to three existing methods.

Index Terms— HMM-based speech synthesis, deterministic plus noise model, excitation model, residual frame, PCA.

1. INTRODUCTION

Statistical parametric speech synthesis based on hidden Markov models (HMMs) has gained much popularity due to its flexibility, reduced memory footprint and high-performance [1]. In this approach, speech is modeled based on source-filter representation. The source refers to the excitation signal produced due to the vibration of vocal folds, and the filter refers to the sequence of time-varying resonators formed by the vocal-tract. The vocal-tract filter and the excitation signal are parameterized and modeled by HMMs in a unified framework. Even though much research has been carried out in recent years, the quality of synthesized speech still seems to have degraded due to the buzziness caused by improper parameterization of the excitation signal. This paper aims at improving the quality of synthesized speech by developing an efficient method for representing and modeling the excitation signal.

In literature, several excitation modeling approaches have been proposed for improving the quality of HMM-based speech synthesis system (HTS). One of the initial approach to model the excitation signal was based on mixed excitation (ME) approach [2]. In ME, voiced excitation is composed of both periodic and aperiodic components, where their relative magnitudes are controlled by band-pass voicing strengths. Later, Zen et al., have used ME approach for speech transformation and representation using adaptive interpolation of the weighted spectrum (STRAIGHT) [3] to the HTS [4]. In [5], Glott-HMM has derived the excitation signal by modifying a single natural instance of glottal flow pulse according to the generated source parameters. Liljencrants-Fant (LF) model has been used to generate excitation source signal in HTS [6]. The LF parameters are modeled by HMMs, and during synthesis the generated LF parameters are used to control the shape of the glottal pulse. A hybrid approach proposed in [7] utilizes a codebook of pitch-synchronous residual frames for generating the source of excitation during synthesis. Drugman et. al., proposed a hybrid approach based on deterministic plus stochastic model (DSM) [8][9].

In order to reduce buzziness and improve the quality of synthesized speech, a new approach of modeling the excitation signal is proposed. In this approach, the residual signal is decomposed into deterministic and noise components based on principal component analysis. The deterministic component is parameterized using PCA coefficients, and the noise component is represented in terms of spectral and amplitude envelopes. This approach is simple and computationally less intensive, as deterministic and noise components are extracted directly from time-domain representation, without transforming to any other domain. In this paper, the terms source, excitation, and residual are used interchangeably. This paper is organized as follows. Section 2 describes the proposed excitation model. Speech synthesis using the proposed approach is described in Section 3. Evaluation of proposed method is provided in section 4. Section 5 concludes the present work and gives some guidelines for future work.

2. PROPOSED EXCITATION MODEL

The excitation signal is obtained by inverse filtering the speech signal. The filter parameters model the vocal-tract transfer function. The excitation signal is pitch-synchronously decomposed into a number of residual frames. The number of pitch-synchronous residual frames varies from one phone to other. Adjacent pitch-synchronous residual frames exhibit strong correlation [10]. On close observation, the shapes of

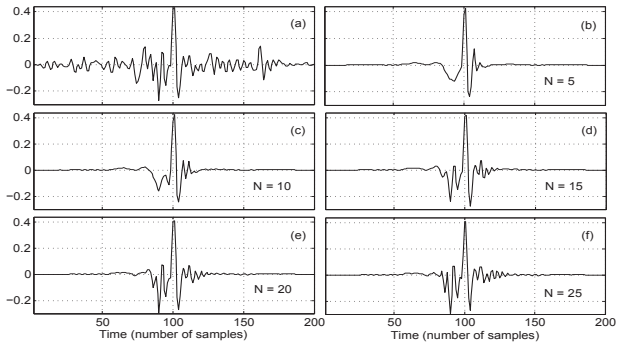

Fig. 1. (a) Original residual frame. Residual frame reconstructed using (b) 5, (c) 10, (d) 15, (e) 20 and (f) 25 eigenvectors

adjacent residual frames around glottal closure instant (GCI) are very much similar. To further analyze the characteristics of residual signal around GCI, principal component analysis is performed on the pitch-synchronous residual frames. For analysis, we considered 10,000 residual frames extracted from SLT speaker of CMU Arctic database [11]. The residual frame ($\mathbf{x}$) can be reconstructed by PCA analysis as follows: $\tilde{\mathbf{x}} = \sum_{n=1}^{N} \alpha_n \mathbf{u_n} + \bar{\mathbf{x}}$. N denotes the number of eigenvectors and $\bar{\mathbf{x}}$ is the sample mean of $\mathbf{x}$. $\mathbf{u_n}$ and α_n denote the eigenvector corresponding to the n-th eigenvalue and the coefficient associated to $\mathbf{u_n}$, respectively. Original residual frame and residual frames reconstructed using first 5, 10, 15, 20 and 25 eigenvectors are shown in Fig. 1. From the figure, it can be observed that by considering lower order eigenvectors (5 and 10), only the region around GCI (middle portion of the residual frame) is reconstructed. Finer details present at other regions are captured, as the order of eigenvectors is increased. Evolution of cumulative relative dispersion (CRD) for different number of eigenvectors is shown in Fig. 2. CRD is defined as the ratio of variance represented by the first M eigenvectors to the total variance. From Fig. 2, it can be seen that about 59% of the variance is represented by the first 20 eigenvectors which mainly corresponds to the region around GCI of the residual frame. To represent the remaining part of the residual frame, 100 higher order eigenvectors are required. The region around GCI represents most of the variance and hence can be regarded as dominant part of the residual frame. The region around GCI also carries important information related to perceptual characteristics of the voiced speech [12][13]. In [14][8], it is stated that the segment of residual signal around GCI is closely related to LF model [6].

Based on the above observation, the residual signal can be divided into two parts. The first part is the small segment of the residual signal around GCI and the second part is the remaining segment of the residual signal. The segment of residual signal around GCI is considered to have equal length on either side of GCI. To ensure smooth continuity at the joining

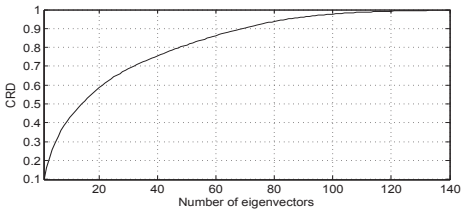

Fig. 2. Evolution of CRD as a function of number of eigenvectors for SLT speaker. Total number of eigenvectors = 200

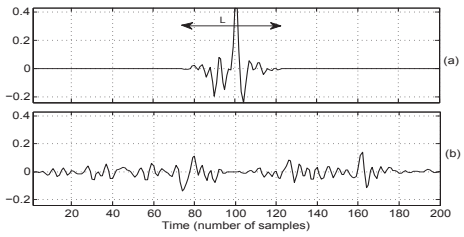

Fig. 3. (a) Deterministic and (b) noise components extracted from the residual frame given in Fig. 1(a)

points, the segment of residual around GCI is Hanning windowed. The Hanning windowed segment is subtracted from the residual signal to obtain the second part. The first part can be predicted from a small number of eigenvectors (about 20) and hence it can be considered as the deterministic component. The second part requires a large number of eigenvectors (about 100) for accurate estimation and hence it can be regarded as the noise component. Fig. 3 shows the deterministic and noise components extracted from the residual frame shown in Fig. 1(a).

The proposed excitation model represents the residual signal as deterministic and noise components. The flow diagram indicating different steps in the proposed excitation modeling is shown in Fig. 4. First, energy is extracted from every frame of the excitation signal. Then, the pitch-synchronous analysis is performed on the excitation signal leading to a set of GCI centered two-pitch period long and Hanning windowed residual signals. The pitch periods of residual frames are normalized to maximum pitch period of the speaker. The energy of residual frame is normalized by fixing the total energy to 1. From the residual frames, deterministic and noise components are computed using the proposed approach. The deterministic component is accurately represented using 20 PCA coefficients (explained in Sec 2.1) and the noise component is parameterized in terms of spectral and amplitude envelopes (explained in Sec 2.2). Harmonic to noise ratio (HNR) is computed as the ratio of energy of deterministic and noise components. In addition, 34th order Mel-Generalized Cepstral (MGC) coefficients (with α = 0.42, Fs = 16 KHz and γ = -1/3) and F0 are extracted from speech utterances. The extracted parameters are modeled under HMM framework.

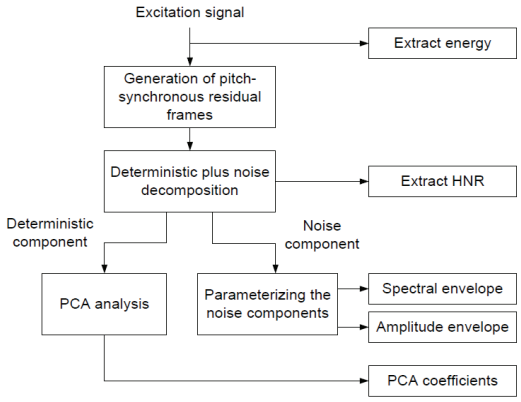

Fig. 4. Flowchart indicating the sequence of steps for proposed excitation modeling

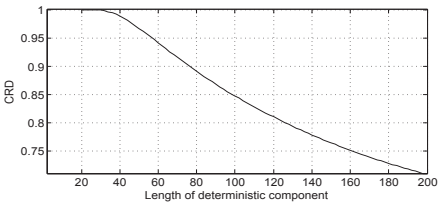

Fig. 5. CRD values computed for different lengths of deterministic component (L) for SLT speaker.

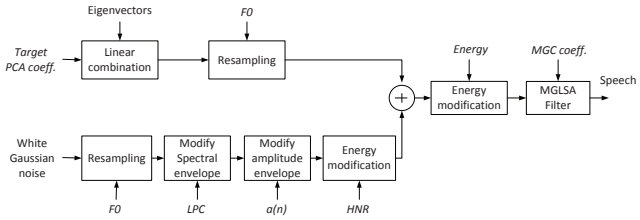

Fig. 6. Block diagram showing different stages in synthesis. Parameters generated by the HMMs are shown in italics.

2.1. Parameterization of deterministic component

Before parameterizing the deterministic component, we need to fix the length (L) of deterministic component (shown in Fig. 3(a)). The length should be appropriately chosen such that the deterministic component is accurately represented with M number of eigenvectors. First, by varying the length L from 2 to twice the normalized pitch period (in number of samples) in steps of 2 samples, the deterministic components are extracted from the residual frames. Here, 10,000 residual frames from SLT speaker are considered. By considering the deterministic components of every length L, PCA analysis is performed. For every L, the CRD value is computed for M number of eigenvectors. The largest possible L which results in CRD value $\geq 95\%$ is considered as the appropriate length of the deterministic component. Before finding the appropriate length L, the number of eigenvectors M should be fixed. By varying M from 1 to 200, deterministic components are computed. Increasing the value of M results in the subsequent increase in the value of L and vice versa. If M is chosen very small, the length L will also be very small, This may not exactly capture the region around GCI and results in reduced quality of speech. If M is chosen very large, then the complexity of model increases and more data is required to capture the actual distribution. For M = 20, the length of the deterministic component is observed to be optimum. For different lengths of deterministic components, CRD values are computed for 20 eigenvectors. Fig. 5 provides the CRD values computed for different lengths of deterministic components for SLT speaker. From the figure, it can be observed that the maximum length of the deterministic component with CRD value greater than 95% is 56. With L = 56, the deterministic components are extracted from the residual frames of SLT speaker and PCA analysis is performed. Each deterministic component is compactly represented by using 20 PCA coefficients.

2.2. Parameterization of noise component

The noise component is parameterized in terms of its spectral and amplitude envelopes. The spectral envelope of the noise component is estimated by using linear predictive coding (LPC). The order of LPC is chosen to be 10. The LPC coefficients are converted to line spectrum frequency (LSF) coefficients. The amplitude envelope ($a(n)$) is obtained by filtering the absolute value of noise component ($u(n)$) with a moving average filter of order $2N + 1$. In this work, N is chosen to be 8. Normalization of the envelope is performed by setting the maximum value to 1. Due to smoothening by the moving average filter, the amplitude envelope shows slow variation. The overall shape of the amplitude envelope is represented by downsampling it into 15 samples. PCA coefficients, spectral and amplitude envelopes of the noise component, and HNR are computed for every residual frame. As it is convenient to model the parameters at frame size of 25 ms with frame shift of 5 ms, the parameters extracted from the residual frames present in the frame are averaged and assigned as the parameters of that frame. In case of unvoiced speech, except energy, all other excitation parameters are set to zero.

3. SPEECH SYNTHESIS

During synthesis, MGC coefficients, F0 including voicing decision, and excitation parameters are generated from the HMMs using constrained maximum likelihood algorithm [15]. The block diagram showing different synthesis stages are shown in Fig. 6. The excitation signal is generated separately for voiced and unvoiced frames. For voiced frame, the deterministic component of the residual frame is obtained

from the linear combination of eigenvectors and target PCA coefficients. The deterministic component is zero padded on either side such that its length is twice the normalized pitch period. The zero padded deterministic component is resampled to twice the target pitch period. The noise component of the residual frame is generated by imposing target spectrum and amplitude envelopes on white Gaussian noise. The energy of noise component is modified according to the generated HNR. Both deterministic and noise components are superimposed, and then overlap-added to generate the excitation signal. The energy of excitation signal is modified according to the energy measure generated from the HMM. For unvoiced speech, white noise whose energy is modified according to the generated energy measure is used as the excitation signal. The resulting excitation signal is given as input to the Mel-Generalized Log Spectrum approximation (MGLSA) filter, controlled by MGC coefficients to generate speech.

4. EVALUATION

The proposed method is evaluated using one female (SLT) and one male (AWB) speakers from CMU Arctic speech database [11]. The training set of each of the speaker consists of about 1100 phonetically balanced English utterances. The duration of the training set is about 56 and 79 minutes for SLT and AWB speakers, respectively. 20 sentences that were not part of training data were used for evaluation purpose. Subjective evaluation is conducted with 20 research scholars in the age group of 23-35 years. The subjects have sufficient speech knowledge for proper assessment of the speech signals. The quality of synthesized speech from the proposed method is compared with three existing methods, namely, pulse-HTS, STRAIGHT-HTS [4] and DSM-HTS [8]. In pulse-HTS, a sequence of pulses positioned according to the generated pitch is used as the excitation signal. In STRAIGHT-HTS, the excitation signal consists of a sequence of impulses and noise components weighted by band-pass filtered aperiodicity parameters. In DSM-HTS, the excitation signal is constructed by modifying the previously stored deterministic component and the energy envelope of noise component according to the generated pitch. Before evaluation, the energy of speech samples are normalized to the same level.

Subjective evaluation is performed using two measures, namely, comparative mean opinion scores (CMOS) and preference tests. In CMOS, subjects were asked to listen to two versions, namely, speech synthesized from the proposed method and the other from the existing methods. Two versions were randomly shuffled to avoid the bias towards any specific method. Subjects were asked to grade the overall preference on a 7-point scale (-3 to +3). A positive score indicates that the proposed method is preferred over other method, and negative score implies the opposite. CMOS

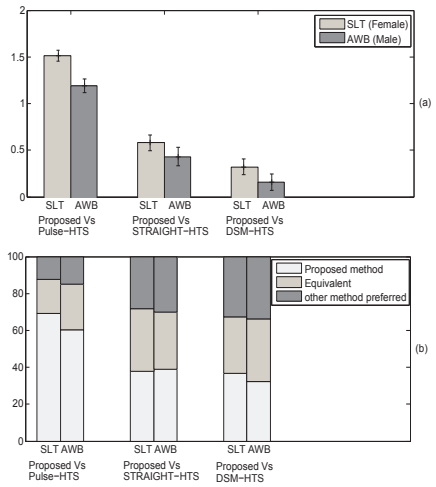

Fig. 7. (a) CMOS with 95% confidence intervals and (b) preference scores.

scores with 95% confidence intervals are shown in Fig. 7(a). CMOS scores of both female and male speakers are varying between 0.1 to 1.5 which indicates that the proposed method is better than the existing methods. CMOS scores of the female speaker are relatively higher compared to the male speaker. Among three methods, CMOS score of DSM-HTS has the lowest value (close to 0). The reason for low value is that both methods parameterize the excitation signal for every glottal cycle. The proposed method is slightly better as the segment of residual signal around GCI which is important for perception of speech is accurately represented. In preference tests, subjects were asked to either prefer one of the synthesized speech utterances or to prefer both as equal. The preference scores are provided in Fig. 7(b). For both female and male speakers, subjects preferred the proposed method compared to other three methods.

5. CONCLUSION

This paper proposed a parametric approach of modeling the excitation signal as deterministic and noise components. The deterministic component is modeled using PCA coefficients, and the noise components are parameterized in terms of spectral and amplitude envelopes. During synthesis, the deterministic and noise components are reconstructed from the parameters generated from HMMs. The evaluation results indicated that the quality of proposed method is considerably better compared to three existing methods. In this work, PCA analysis is performed on the residual frames of all phones. Instead, PCA analysis can be performed on the residual frames of every phone, and the quality of synthesized speech can be analyzed. The relation between time and frequency domain decomposition of excitation signal can also be analyzed.

6. REFERENCES

[1] Keiichi Tokuda, Yoshihiko Nankaku, Tomoki Toda, Heiga Zen, Junichi Yamagishi, and Keiichiro Oura, "Speech synthesis based on hidden Markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[2] 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.

[3] H Kawahara, I Masuda-Katsuse, and A de Cheveigne, "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 Communication*, vol. 27, pp. 187–207, 1998.

[4] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE Transactions on Information and Systems*, vol. E90-D, pp. 325–333, 2007.

[5] Tuomo Raitio, Antti Suni, Junichi Yamagishi, Hannu Pulakka, Jani Nurminen, Martti Vainio, and Paavo Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 153–165, 2011.

[6] J.P. Cabral, S. Renals, J. Yamagishi, and K. Richmond, "HMM-based speech synthesiser using the LF-model of the glottal source," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 4704 –4707.

[7] Thomas Drugman, Alexis Moinet, Thierry Dutoit, and Geoffrey Wilfart, "Using a pitch-synchrounous residual codebook for hybrid HMM/frame selection speech synthesis," in *Proc. International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, 2009, pp. 3793–3796.

[8] Thomas Drugman and Thierry Dutoit, "The deterministic plus stochastic model of the residual signal and its applications," *IEEE Transactions on Audio, Speech, and Language processing*, vol. 20, pp. 968–981, 2012.

[9] Thomas Drugman and Tuomo Raitio, "Excitation modeling for HMM-based speech synthesis: breaking down the impact of periodic and aperiodic components," in *Proc. International Conference on Audio, Speech and Signal Processing (ICASSP)*, 2014, pp. 260–264.

[10] E Yumoto, W Gould, and T Baer, "Harmonics-to-noise ratio as an index of the degree of hoarseness," *Journal of Acoustical Society of America*, vol. 71, no. 6, pp. 1544–1550, 1982.

[11] "CMU ARCTIC speech synthesis databases," *[Online]. Available: http://festvox.org/cmu_arctic/.*

[12] Joao P. Cabral, "Uniform concatenative excitation model for synthesising speech without voiced/unvoiced classification," in *Proc. Interspeech*, 2013, pp. 1082–1086.

[13] Nagaraj Adiga and S. R. Mahadeva Prasanna, "Significance of instants of significant excitation for source modeling," in *Proc. Interspeech*, 2013, pp. 1677–1681.

[14] Thomas Drugman, Geoffrey Wilfart, and Thierry Dutoit, "A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis," in *Proc. Interspeech*, 2009, pp. 1779–1782.

[15] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*, 2000, vol. 3, pp. 1315–1318.