Circuits Syst Signal Process (2017) 36:3650–3673
DOI 10.1007/s00034-016-0476-3

Parameterization of Excitation Signal for Improving the Quality of HMM-Based Speech Synthesis System

N. P. Narendra[1] · K. Sreenivasa Rao[1]

Received: 31 August 2015 / Revised: 9 December 2016 / Accepted: 10 December 2016 /
Published online: 3 January 2017

Abstract This paper proposes a new approach of parameterizing the excitation signal for improving the quality of HMM-based speech synthesis system. The proposed method tries to model the excitation or residual signal by segregating the regions of the residual signal based on their perceptual importance. Initially, a study on the characteristics of the residual signal around glottal closure instant (GCI) is performed using principal component analysis (PCA). Based on the present study, and from the previous literature (Adiga and Prasanna in Proceedings of Interspeech, pp 1677–1681, 2013; Cabral in Proceedings of Interspeech, pp 1082–1086, 2013), it is concluded that the segment of the residual signal around GCI which carries perceptually important information is considered as the deterministic component and the remaining part of the residual signal is considered as the noise component. The deterministic component is compactly represented using PCA coefficients (with about 95% accuracy), and the noise component is parameterized 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 of quality for both female and male speakers' speech synthesized by the proposed method, compared to three existing excitation modeling methods. Accurate parameterization of the segment of the residual signal around GCI resulted in the improvement of the quality of the synthesized speech. Synthesized speech samples of the proposed and existing source models are made available online at http://www.sit.iitkgp.ernet.in/~ksrao/parametric-hts/pcd-hts.html.

✉ N. P. Narendra
narendrasince1987@gmail.com

K. Sreenivasa Rao
ksrao@iitkgp.ac.in

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

Keywords HMM-based speech synthesis · Deterministic plus noise model · Excitation model · Residual frame · PCA

1 Introduction

A text-to-speech (TTS) synthesis system converts a given text to corresponding speech output [26,43]. Nowadays, TTS technology has been extensively used in several consumer applications such as speech-to-speech translation systems, mobile phones, household devices, assistive aids for visually challenged people, navigation and personal guidance gadgets [18,27,35]. Hidden Markov model (HMM)-based TTS system has become a very good choice in these applications because of flexibility, reduced memory footprint and high performance with reduced computational resources [43]. In HMM-based speech synthesizer, speech is modeled based on source–filter representation [34]. The source refers to the excitation signal produced due to the vibration of vocal folds while the filter refers to the sequence of time-varying resonators formed by the vocal tract. The vocal tract filter and 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 been degraded by two factors, namely (i) *buzziness* caused due to improper parameterization of the excitation signal and (ii) *muffledness* caused by over-smoothing of the generated parameters due to statistical modeling. This paper addresses the first issue and aims at developing an efficient method for representing and modeling the excitation signal.

In the literature, several excitation or source modeling approaches have been proposed for improving the quality of HMM-based speech synthesis system (HTS). One of the initial approaches to model the excitation was reported by Yoshimura et al. [45]. It consists of modeling the excitation parameters used in mixed excitation linear prediction (MELP) [23] algorithm by HMMs. During synthesis, the generated excitation parameters were used to construct the mixed excitation in the same way as in MELP algorithm. Later, Zen et al. [47] have used mixed excitation approach for speech transformation and representation using adaptive interpolation of the weighted spectrum (STRAIGHT) [16] to the HTS. They modeled F0 and aperiodicity parameters by HMMs in order to enable the generation of excitation signal during synthesis stage [47]. An approach in which the excitation signal is constructed by state-dependent filtering of pulse trains and white noise sequences is proposed in [21]. During training, filters and pulse trains are jointly optimized through a procedure that resembles analysis-by-synthesis speech coding algorithms. Wen et al. [44] proposed the pitch-scaled spectrum-based method 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 HTS. In [3], Liljencrants-Fant (LF) model is proposed for modeling the glottal source signal in HTS. The LF parameters are modeled by HMMs, and during synthesis, the generated LF parameters are used to control the shape of the glottal pulse. Raitio et al. [36] proposed an approach of generating the excitation signal by modifying a single natural instance of glottal flow pulse according to the source parameters generated from the HMM. The glottal flow pulse is obtained by iterative adaptive glottal inverse filtering

[2]. Uniform concatenation excitation model is proposed in [4] to generate the excitation signal in both voiced and unvoiced speech. This model generates the residual signal by concatenating two consecutive segments. The first segment is a part original residual waveform around the pitch mark and the second segment is modeled by the parameters of amplitude envelope and energy of the residual waveform.

Instead of using the parameters derived from statistical models, a hybrid approach was proposed which utilize the real excitation segments for generating the excitation signal [7,8,11,12,37]. In [7,8], the hybrid source models are proposed which generated the excitation signal by selecting suitable residual frames from the codebook based on target residual specification. Raitio et al. [37] utilized the unit selection method to select appropriate glottal source pulses from the database based on target and concatenation costs. The selected glottal source pulses are used to construct the excitation signal. Drugman et al. [11] proposed a hybrid approach based on deterministic plus stochastic model (DSM). 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 the 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 [12].

Some of the recently proposed excitation modeling methods (both parametric and hybrid) perform pitch-synchronous analysis and model the pitch-synchronous residual frames of excitation signal [4,11,12,37]. These methods have shown a better quality of synthesized speech than the traditional way of modeling excitation signal using mixed excitation approach [45,47]. In this work, a new excitation modeling method is proposed where the pitch-synchronous residual frames are decomposed into deterministic and noise components in the time domain. In the proposed method, the deterministic component is parameterized by using PCA coefficients. Even though the proposed and DSM-based source model utilize PCA for parameterization, there exist certain basic differences between the two approaches. The DSM divides the spectrum of the residual frame in two parts, namely low-frequency (termed as the deterministic component) and high-frequency parts (termed as the noise component). Here, PCA is performed on the entire length of the residual frame. The authors of the DSM explored different number of eigenvectors for the excitation generation and concluded that by increasing the number of eigenvectors, significant improvement in the quality synthesis is not observed. Hence, only the first eigenvector is considered for the representing the deterministic component. The proposed method try to model the residual frames based on their perceptual significance. In [1,4], it is observed that in the entire length of the residual frame, the segment around GCI carries important information related to the perception of speech. Based on this motivation, the proposed method performed PCA on the residual frames. From the analysis, it is observed that the segment of the residual frame around GCI which is perceptually important contributes to the major portion of the residual frame. This segment of the residual frame around GCI is considered as the deterministic component, and the remaining segment of the residual frame is considered as the noise component. Another distinctive attribute of the proposed method is that the deterministic component is accurately represented using

PCA coefficients (about 95% accuracy). The other existing methods (including DSM source model) parameterize the residual frames by using PCA coefficients with up to 60–70% accuracy [9,10]. In this paper, the terms source, excitation and residual are used interchangeably.

This paper is organized as follows. Section 2 describes the proposed excitation modeling approach. The steps involved in the synthesis of speech using the proposed excitation model are explained in Sect. 3. Section 4 provides the description of HMM-based speech synthesis system with the proposed excitation model. Evaluation of the proposed method is provided in Sect. 5. Section 6 concludes the present work and presents some guidelines for the future work.

2 Modeling of Deterministic and Noise Components of Excitation Signal in Time Domain

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 [46]. On close observation, the shapes of adjacent residual frames around glottal closure instant (GCI) are very much similar. Most of the existing approaches parameterize the entire residual frame by considering either the time-domain or frequency-domain representation of the signal [8,11,36]. They do not parameterize the residual frames based on its perceptual significance. In [1,4], it is observed that in the entire residual frame, the region around GCI carries important information related to the perceptual characteristics of the voiced speech [1,4]. Motivated by this observation, to further analyze the characteristics of the residual signal around GCI, principal component analysis is performed on the pitch-synchronous residual frames [29]. For analysis, 10,000 residual frames extracted from SLT speaker of CMU Arctic database [5] are considered. PCA is carried out on the database sampled at 16 kHz. The residual frame ($\mathbf{x}$) can be reconstructed by PCA as follows:

$$\tilde{\mathbf{x}} = \sum_{n=1}^{N} \alpha_n \mathbf{u}_n + \bar{\mathbf{x}} \tag{1}$$

where N denotes the number of eigenvectors and $\bar{\mathbf{x}}$ is the sample mean of $\mathbf{x}$. $\mathbf{u_n}$ is the eigenvector of the covariance matrix $\sum_{\mathbf{x}} = E\{\mathbf{x}\mathbf{x}^{\mathrm{T}}\}$ and α_n is the coefficient associated with $\mathbf{u}_n$. It is assumed that the eigenvectors are ordered on the eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_N$. For compact representation, only first $N' < N$ eigenvectors or principal components are used which results in N' PCA coefficients. The principal components represent the directions of the largest variance in the signal space. With different number of eigenvectors (N' = 5, 10, 15, 20 and 25), PCA coefficients are computed. Using different number of eigenvectors and PCA coefficients, the residual frames are reconstructed and their variation are analyzed.

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

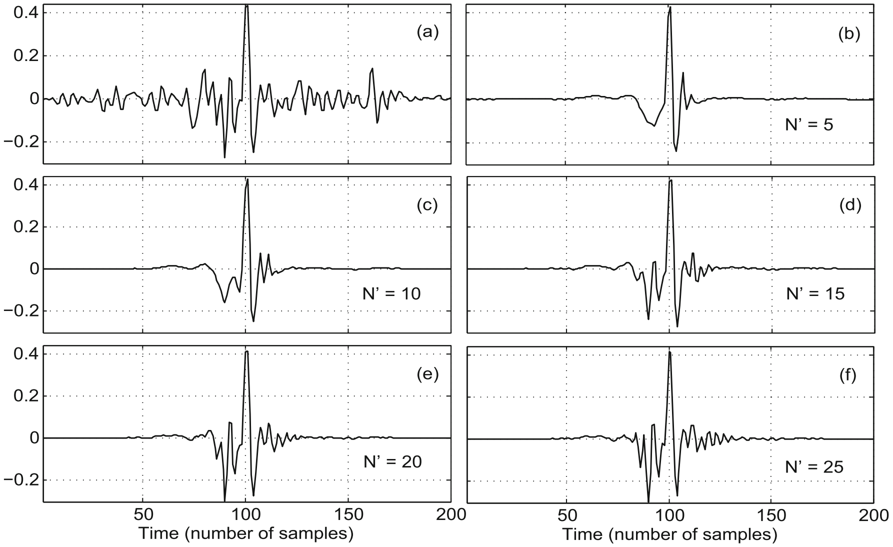
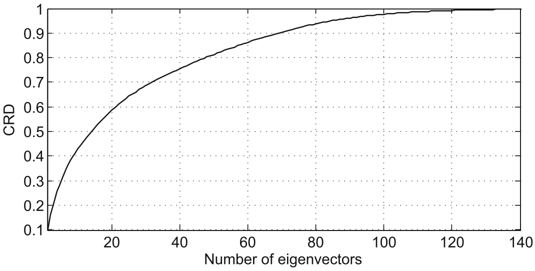


Fig. 1 **a** Original residual frame. Residual frame reconstructed using **b** 5, **c** 10, **d** 15, **e** 20 and **f** 25 eigenvectors

Fig. 2 Evaluation of cumulative relative dispersion (CRD) as a function of number of eigenvectors for SLT speaker. Total number of eigenvectors = 200



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. Evaluation of cumulative relative dispersion (CRD) for the 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. Drugman et al. [9,11], have shown that the segment of the residual signal around GCI is closely related to LF model [3].

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 which is perceptually significant, and the second part is the remaining segment of the residual signal.

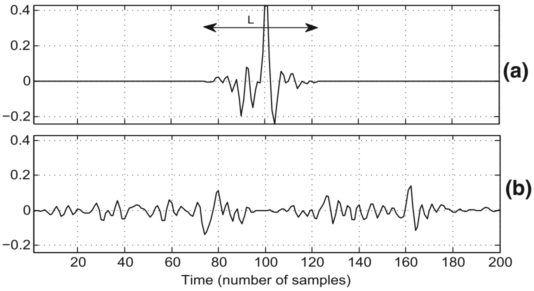


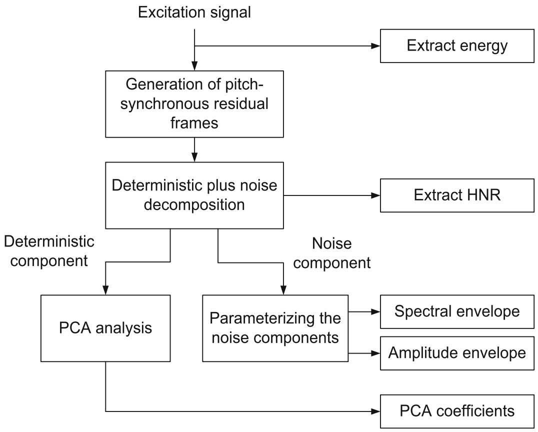
Fig. 3 **a** Deterministic and **b** noise components extracted from the residual frame given in Fig. 1a

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 points, the small segment of the residual frame around GCI is Hanning windowed. The Hanning windowed segment was subtracted from the residual frame 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 of the residual frame. The second part (i.e., other than the deterministic component) requires a large number of eigenvectors (about 100) for accurate estimation, and hence, it can be considered as the noise component of the residual frame. Figure 3 shows the deterministic and noise components extracted from the residual frame shown in Fig. 1a.

2.1 Proposed Excitation Model

The proposed excitation model represents the excitation signal as deterministic and noise components of the residual signal. 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 residual frames that are synchronous with the GCI and whose length is set to two pitch periods (described in Sect. 2.2). From the pitch-synchronous residual frames, the deterministic and noise components are computed by using the proposed approach. The deterministic component is accurately represented using 20 PCA coefficients (explained in Sect. 2.3), and the noise component is parameterized in terms of spectral and amplitude envelopes (explained in Sect. 2.4). Harmonic to noise ratio (HNR) is computed as the ratio of the energy of deterministic and noise components. The modeling of the HNR ensures that the energy of the deterministic and noise components are properly fixed without any error. Energy, PCA coefficients, HNR, spectral and amplitude envelopes are considered as the excitation parameters. At the time of synthesis, the deterministic component waveform is reconstructed from the generated PCA coefficients, and the noise component is obtained by imposing the target spectral and amplitude envelopes on the white Gaussian noise. The deterministic and noise components are pitch-synchronously overlap-added to generate the excitation signal (described in Sect. 3).

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



2.2 Generation of Pitch-Synchronous Residual Frames

The pitch-synchronous residual frames are extracted from the excitation signal using the knowledge of GCIs. Using GCIs, the boundaries of pitch cycles are marked on the excitation signal. GCIs are estimated from the speech signal using zero-frequency filtering (ZFF) method [25]. The main reason for choosing the ZFF method is that it has good identification rate and accuracy. Using GCI positions as anchor points, two-pitch period long residual signals are extracted, and they are Hanning windowed. During extraction residual frames, only those residual frames are extracted which are having GCI at the centre of the residual frame and whose GCI coincides with the peak of the residual frame. This ensures the selection of the residual frames with correctly detected GCIs and the rejection of the residual frames with the wrongly detected GCIs. The extracted residual signals are normalized both in pitch period and energy. The pitch periods of the residual frames are normalized to the maximum pitch period of the speaker. The energy of the residual frame is normalized by fixing the total energy to 1. These operations make the residual signals comparable so that they can be analyzed under a common framework. GCI centered two-pitch period long, and Hanning windowed residual signals are viewed as the pitch-synchronous residual frames. Figure 5a–c shows the segment of the speech signal, its corresponding residual signal and an example of extracted pitch-synchronous residual frame, respectively. The locations of GCIs are shown by downward arrows in Fig. 5b.

2.3 Parameterization of Deterministic Component

Before parameterizing the deterministic component, the length of the deterministic component (L), i.e., the length of the segment of residual signal around GCI as shown in Fig. 3a should be fixed. The length should be appropriately chosen such that the deterministic component can be accurately represented with M number of eigenvectors. First, by varying the length L from 2 to twice the normalized pitch period (in the number of samples) in steps of 2 samples, the deterministic components are

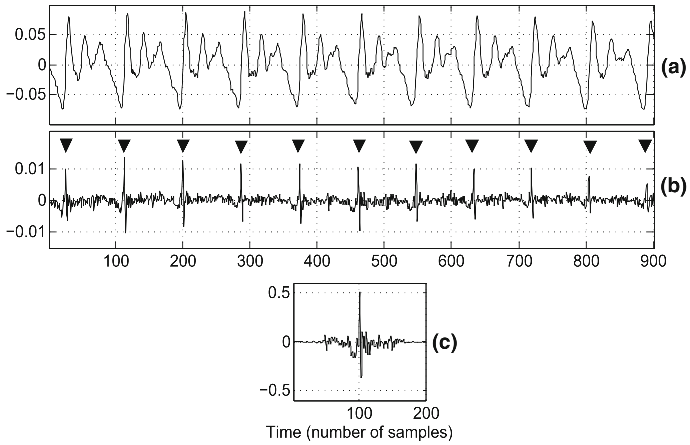


Fig. 5 **a** Speech signal, **b** residual signal and **c** pitch-synchronous residual frame. The locations of GCIs are shown by *downward arrows*

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 is performed. For every L, the CRD value is computed for M number of eigenvectors. The largest possible L which results in the CRD value $\geq$95% is considered as the appropriate length of the deterministic component. Choosing L with CRD value $\geq$95% ensures accurate representation of the deterministic component.

Before finding the appropriate length of the deterministic component, the number of eigenvectors M used for representing the deterministic component should be fixed. By varying M from 1 to 200, the length of the deterministic component is computed which results in the CRD value $\geq$95%. 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 the 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 optimal (about one-third the length of the residual frame). Hence, in this study, M is fixed to 20.

With $M = 20$, CRD values computed for the different lengths of the deterministic components are shown in Fig. 6. The CRD value is close to 100% for the smaller lengths of the deterministic components. From the figure, it can be observed that the largest possible L with CRD value $\geq$95% is 56. With $L = 56$, the deterministic components are extracted from the residual frames of SLT speaker and PCA is performed. Each deterministic component is compactly represented by using 20 PCA coefficients. The deterministic component waveform mean vector and the first three eigenvectors are shown in Fig. 7. The mean vector captures the average shape of the deterministic component waveform, and other components model the rising and decaying patterns just before and after GCI. With the number of eigenvectors M fixed to 20, the length of the deterministic component is computed separately for every speaker. Generally, the length of the deterministic component is higher for male speakers (due to higher pitch period) compared to female speakers.

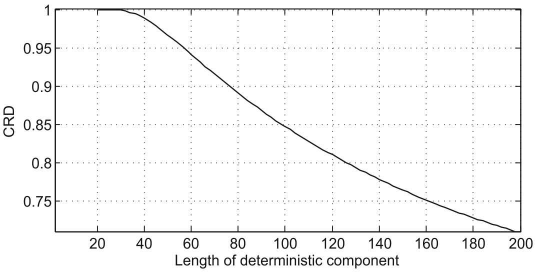


Fig. 6 Cumulative relative dispersion (CRD) values computed for the different lengths of the deterministic component (L) for SLT speaker

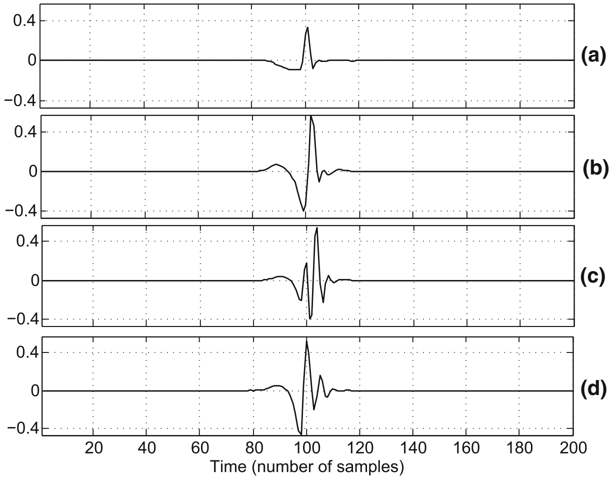


Fig. 7 **a** Mean vector and **b**, **c**, **d** first three eigenvectors of the deterministic component

2.4 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). A typical criterion to select the order of the LPC analysis is to use 1 complex pole per each kHz of the total bandwidth (equal to half the sample rate) plus 2–4 additional poles [14,22]. For the sampling frequency of 16 kHz, 10–14 poles are typically used for LPC analysis. Hence, in this work, the order of LPC is chosen to be 10. The LPC coefficients are converted to line spectrum frequency (LSF) coefficients. The LSFs have better quantization properties and result in low spectral distortion than the conventional LPC coefficients [31,39]. The amplitude envelope ($a(n)$) is obtained by filtering the absolute value of noise component ($u(n)$) with a moving average filter of order $2N + 1$. N is chosen to be 8. The amplitude envelope is given by:

$$a(n) = \frac{1}{(2N+1)} \sum_{k=-N}^{N} |u(n-k)|. \tag{2}$$

Table 1 Source features and the number of parameters

Features	Parameters per frame
Pitch	1
Energy	1
HNR	1
PCA coefficeints	20
Noise spectrum	10
Noise amplitude envelope	15

Normalization of the envelope is performed by setting the maximum value to 1. This method of amplitude envelope estimation was previously performed by Pantazis et al. [32]. Due to smoothening by the moving average filter, the amplitude envelope shows slow variation. The overall shape of the amplitude envelope is represented by a small number of samples. In our case, the amplitude envelope is represented by downsampling it into 15 samples. If the amplitude envelope of the noise component is not modeled along with the spectral envelope, then the noise component is not properly fused into the deterministic component. This can lead to the perception of background noise in the synthesized speech [24,40].

PCA coefficients, spectral and amplitude envelopes of the noise component and HNR are computed for every pitch-synchronous residual frame. As it is convenient to model the parameters at the frame size of 25 ms with the frame shift of 5 ms, the parameters extracted from pitch-synchronous residual frames present in the frame are averaged and assigned as the parameters of that frame. In the case of unvoiced speech, except energy of excitation signal, all other excitation parameters are set to zero. The source features considered in this work are given in Table 1. The source features are modeled under the framework of HMM (discussed in Sect. 4).

3 Speech Synthesis Using the Proposed Excitation Model

During synthesis, MGC coefficients, F0 including voicing decision and excitation parameters are generated from HMMs using constrained maximum likelihood algorithm [42]. The block diagram showing different synthesis stages are shown in Fig. 8. In the figure, parameters generated by the HMMs are shown in italics. 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 constructed using white Gaussian noise. First, white Gaussian noise is resampled to twice the target pitch period. Then, the target spectral envelope generated from the HMM is imposed on the resampled white Gaussian noise. The target spectral envelope is the all-pole model of noise represented by LSF coefficients. The LSFs are converted to LPCs (a_ks). An IIR filter is constructed which filters the white noise signal to obtain the desired target spectrum. The transfer

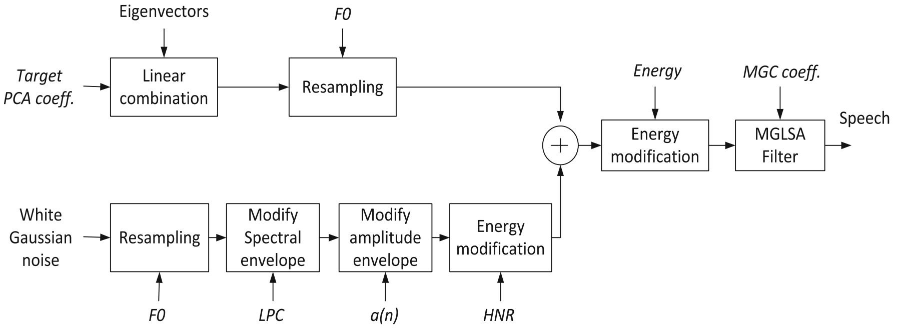


Fig. 8 *Block diagram* showing different stages in synthesis. Parameters generated from the HMM are shown in *italics*

function of IIR filter is given by

$$H(z) = \frac{1}{(1 - G(z))} \tag{3}$$

where $G(z) = \sum_{k=1}^{p} a_k z^{-k}$ is the FIR filter obtained from the LPCs of target spectral envelope. The target amplitude envelope $(a(n))$ generated by the HMM is imposed on the IIR filtered noise signal. The target amplitude envelope which is represented by 15 samples is upsampled to the required target pitch period. The amplitude envelope of the IIR filtered noise signal is also computed. The target envelope is imposed on the IIR filtered noise signal by compensating the difference between two envelopes. The energy of the spectrum and amplitude envelope modified noise signal is changed 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.

In order to understand the perceptual significance of excitation parameters (HNR, noise spectral and amplitude envelopes) on the synthesis quality, the excitation signal is reconstructed by adding the deterministic component with the noise component which is incrementally modified by using each of the excitation parameters. Using the reconstructed excitation signal, the speech signal is reconstructed (analysis–synthesis framework). Figure 9 shows the natural speech, excitation signal, synthesized speech and corresponding excitation signals constructed by adding the deterministic component with the noise component which is incrementally modified using different excitation parameters. Figure 9d shows the excitation signal constructed by using only deterministic components. In our source model, the deterministic component is considered to be a small segment around GCI. Hence, in the figure, we can observe nonzero amplitude values around GCI and other regions of excitation signal are set to zero values. In Fig. 9f, the excitation signal is generated by adding the determin-

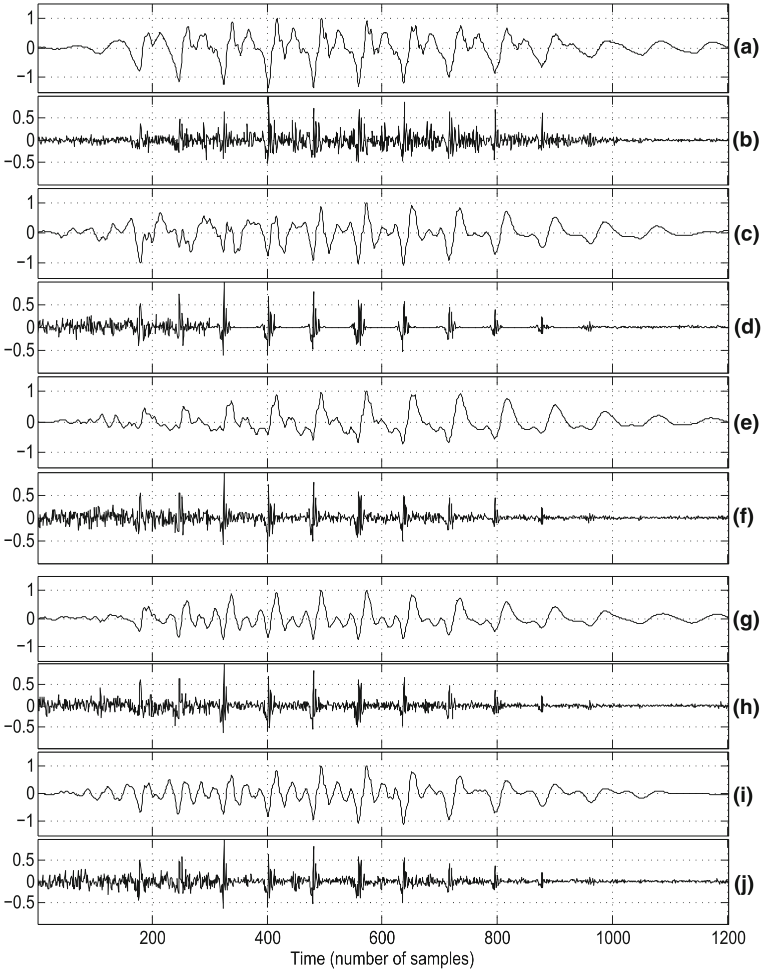


Fig. 9 Illustration of synthesized speech and excitation signals constructed by adding the deterministic component with the noise component which is incrementally modified using different excitation parameters such as HNR, noise spectral and amplitude envelopes. **a** Natural speech, **b** excitation signal, synthesized speech and excitation signal **c**, **d** by using only deterministic component, **e**, **f** by using deterministic component and noise component generated using only HNR, **g**, **h** by using deterministic component, and noise component generated using HNR and noise spectral envelope, **i**, **j** by using deterministic component, and noise component generated using HNR, noise spectral and amplitude envelopes

istic component with the noise component which is constructed by using only HNR parameter. Here, to construct the noise component, the energy of white Gaussian noise is modified according to the HNR value. From Fig. 9f, we can observe that the zero values present in Fig. 9d are filled with nonzero noise values. In Fig. 9h, the excitation signal is generated by adding the deterministic component with the noise component which is obtained by using HNR and spectral envelope. Upon imposing the spectral envelope on the noise component, we can see slight variations in the shapes of the excitation signal (particularly around GCI). Figure 9j shows the excitation signal constructed by adding the deterministic component with the noise component which is constructed by using HNR, spectral and amplitude envelopes. Upon imposing the amplitude envelope on the noise component, we can observe variations in the envelope of every cycle of the excitation signal. We noticed that the secondary excitations are becoming prominently visible after imposing the amplitude envelope on the noise component. On the whole from Fig. 9, it can be noticed that upon adding each of the excitation parameters, the reconstructed speech and excitation signals are observed to be close to the natural speech and excitation signals. We performed informal listening tests on ten speech utterances synthesized by incrementally adding each of the excitation parameters. From informal listening tests, it is observed that by adding each of excitation parameters, the quality of synthesized speech is increased. Among different versions, the speech synthesized by using the excitation signal obtained by combining the deterministic component with the noise component which is generated using HNR, noise spectral and amplitude envelopes is close to natural speech.

4 HMM-Based Speech Synthesis System with the Proposed Excitation Model

The goal of the proposed speech synthesis system is to produce high-quality synthetic speech. The general block diagram of HMM-based speech synthesis system including the proposed excitation model is shown in Fig. 10. The system consists of two main modules: training and synthesis. The HMM-based speech synthesizer is implemented using publicly available HTS toolkit [13].

In the training part, spectrum or vocal tract part and F0 are estimated from speech utterance present in the database. Mel-generalized cepstrum (MGC) parameters which represent the vocal tract part are extracted from the speech utterance. Thirty-fourth-order MGC coefficients are extracted with the parameter values $\alpha = 0.42$ (Fs $=$ 16 kHz) and $\gamma = -1/3$ [48]. In the literature, HMM-based speech synthesis systems are developed with the different orders of MGC coefficients such as 24 [33], 30 [12], 34 [6] and 39 [20]. In [47], Zen et al. have concluded that by increasing the order of MGC coefficients, a small improvement in the quality of synthesis can be observed. In this work, we consider MGC order = 34, as it is considered to be moderate (neither too low nor high for 16 kHz) in the context of HMM-based speech synthesis. F0 estimation with voicing decision is performed using the recently proposed method based on the strength of instants of significant excitation [28]. The excitation signal is obtained by inverse filtering using the MGLSA filter. The excitation signal is modeled using the proposed time-domain deterministic plus noise model approach. Using the proposed source model, set of excitation parameters, namely PCA coefficients, HNR, spectral and amplitude envelopes are estimated. Even though F0 values are also part of excitation parameters, to differentiate between the proposed excitation parameters and the F_0 values generated from the previous method, both parameters are shown as separately obtained from different blocks. MGC coefficients, F0, and excitation parameters are modeled using multi-stream HMMs. Except F0, all parameters are modeled by the continuous probability density HMMs (CD-HMM). The F0 patterns are modeled by an HMM based on the multi-space probability distribution (MSD-HMM). The MSD-HMM consists of a continuous mixture HMM to model one-dimensional continuous

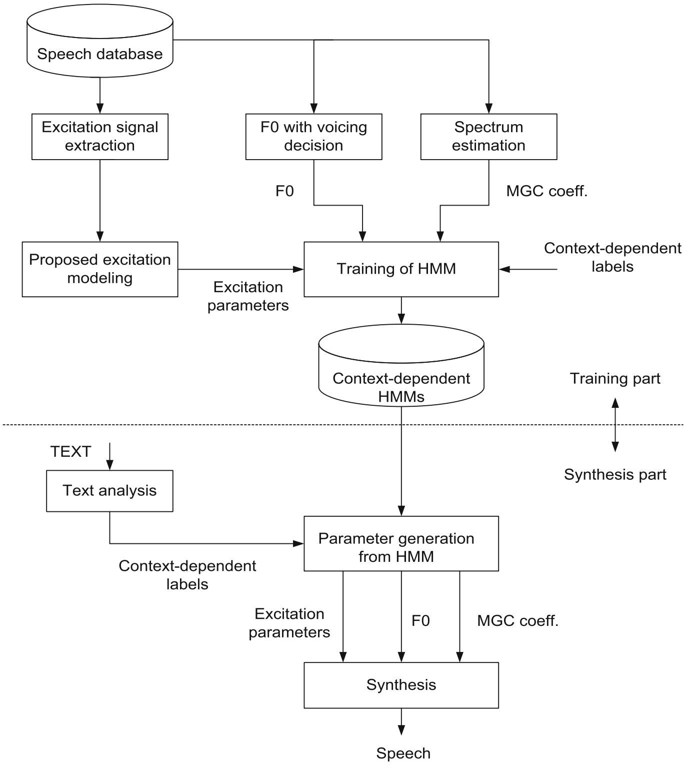


Fig. 10 Block diagram of HMM-based speech synthesis system including proposed excitation model

F0 values that describe voiced region and a discrete HMM to model discrete symbols that represent unvoiced regions. The proposed excitation parameters can also be modeled using MSD-HMM, but in this work CD-HMM was used, and the parameters generated for the unvoiced region are simply discarded. We have also examined modeling the proposed excitation parameters by using MSD-HMM. But, we have not observed any noticeable difference in the synthesized speech by modeling the proposed excitation parameters using CD-HMM and MSD-HMM. The HMMs used in this work consist five emitting states. The output probabilities of each state are modeled using a single Gaussian distribution with diagonal covariance. The temporal structure of the speech is modeled by state duration densities of HMMs. The state durations of each phoneme HMM are modeled using a single Gaussian distribution with diagonal covariance.

First, using the phonetic labels having time alignment information, the monophone HMMs are trained using the segmental K-means and expectation–maximization (EM) algorithm. The monophone HMMs are converted into context-dependent HMMs, and the model parameters are reestimated again. Decision tree-based context clustering technique [30,38] is applied to context-dependent HMMs. The question set consists of a standard list of 53 positional and contextual features provided in basic HTS implementation [13]. At each leaf node of the decision tree, the model parameters are tied and reestimated again. In the proposed system, only MGC coefficients and F0 streams were considered during the alignment step of reestimation; weights of other streams are set to zero.

During synthesis, the input text is converted into a sequence of context-dependent phoneme labels. According to the label sequence, a sentence HMM is constructed by concatenating context-dependent HMMs. Then, a sequence of parameters is generated from the sentence HMM. To alleviate the problem of over-smoothing of generated parameters due to statistical processing, the global variance technique is used [41]. From the generated source parameters, the excitation signal is constructed. Speech waveform is synthesized using the generated MGC coefficients and the excitation signal.

5 Evaluation

The proposed method is evaluated using two female (SLT and CLB) and two male (AWB and KSP) speakers from CMU Arctic speech database [5]. 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, 64, 79 and 59 min for SLT, CLB, AWB and KSP speakers, respectively. All experiments carried out in this work use the speech database sampled at 16 kHz. Utilizing 16 kHz sampling rate can preserve up to 8 kHz speech bandwidth which is sufficient to preserve most of the spectral energy. In the literature, many speech synthesis systems developed with 16 kHz sampling rate can produce pleasant and intelligible synthetic speech [11,36]. Twenty 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, as all of them have taken a full semester course on speech technology. Each of the subjects was given a pilot test about the perception of speech signals by playing samples of synthesized speech files. Once they were comfortable with judging, they were allowed to take the tests. The tests were conducted in the laboratory environment by playing the speech signals through headphones.

The quality of synthesized speech from the proposed method is compared with three existing excitation modeling methods, namely pulse-HTS, STRAIGHT-HTS [47] and DSM-HTS [11]. Pulse-HTS is known for its simple excitation scheme, and it is mostly used as a reference for testing the proposed methods. In pulse-HTS, a sequence of pulses positioned according to the generated pitch is used as the excitation signal. STRAIGHT-HTS [47] is one of the most widely used methods for high-quality speech synthesis and uses mixed excitation parametric approach for source modeling. STRAIGHT-HTS uses TANDEM STRAIGHT method [17] as the source model. In this source model, the excitation signal consists of a sequence of impulses and noise components weighted by band-pass filtered aperiodicity parameters. The source codes of this model are obtained from the authors of STRAIGHT method. DSM-HTS [11] is one of the recently proposed popular approaches which models the pitch-synchronous residual frames based on the deterministic plus stochastic model. In this work, two versions of DSM-HTS are used for evaluation. In the first version, only the first eigen-

Table 2 Scores for the CMOS test

Score	Subjective perception
3	Much better
2	Better
1	Slightly better
0	About the same
−1	Slightly worse
−2	Worse
−3	Much worse

vector is used as the deterministic component. The excitation signal is constructed by modifying the deterministic component and the average energy envelope of the noise component according to the generated pitch. In the second version, the deterministic component of every residual frame is represented using 20 PCA coefficients. At the time of training, PCA coefficients are modeled using HMMs. During synthesis, the deterministic component is obtained from the PCA coefficients generated from HMMs, and the noise component is obtained by resampling the average noise energy envelope according to the generated pitch. Two versions of DSM-HTS, namely (1) single eigenvector and (2) 20 eigenvectors are compared with the proposed source model. Before evaluation, the energies of synthesized speech signals 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 other from the existing methods. Two versions were randomly shuffled to avoid the bias toward any specific method. Subjects were asked to grade the overall preference on a 7-point scale. The 7-point scale used to rank the preference between pairs of synthesized speech samples is shown in Table 2. A positive score indicates that the proposed method is preferred over other method, and negative score implies the opposite. In preference tests, subjects were asked to give the preference between a pair of synthesized speech utterances. The subjects had the option either to prefer one of the synthesized speech utterances or to prefer both as equal.

CMOS scores with 95% confidence intervals and preference scores are provided in Figs. 11 and 12, respectively. On comparing the proposed method with the pulse-HTS, it can be observed that the CMOS scores are varying between 1.1–1.5 and more than 60% of the subjects preferred the proposed method for both male and female speakers. This indicates that the quality of speech synthesized by the proposed method is clearly better than the speech synthesized by the pulse-HTS. The subjects noticed that the speech synthesized from the pulse-HTS is artificial and unnatural. The excitation signals generated using the combination of deterministic and noise components are much better than the sequence of pulses.

On comparison of the proposed method with the STRAIGHT-HTS, it can be observed that the CMOS scores vary between 0.4–0.7, and the subjects preferred the proposed method for about 40% of cases and preferred the STRAIGHT-HTS for about 28% of cases. Both measures indicate that the proposed method is better than

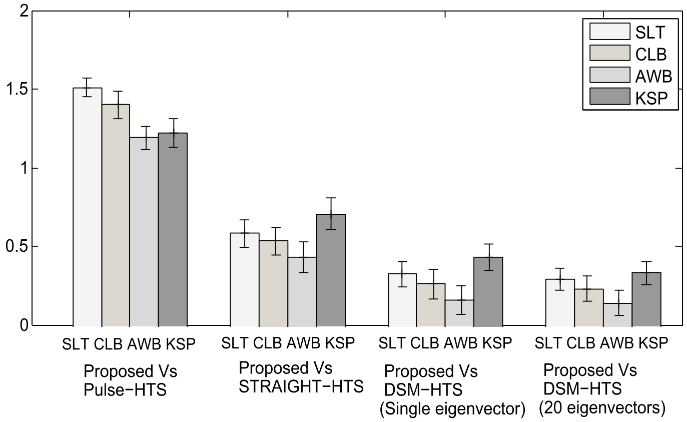


Fig. 11 CMOS scores with 95% confidence intervals obtained by comparing the proposed method with the existing methods

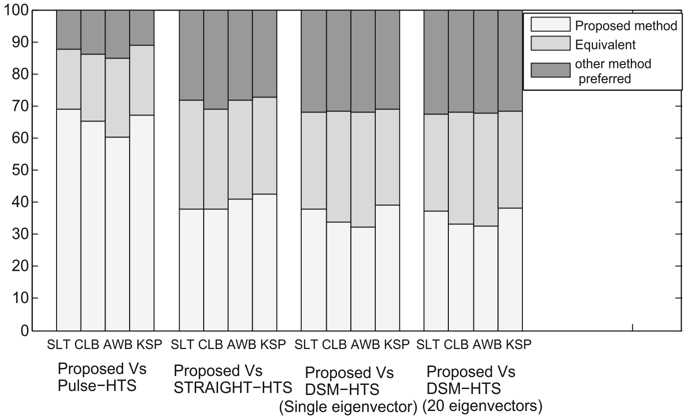


Fig. 12 Preference scores obtained by comparing the proposed method with the existing methods

STRAIGHT-HTS. The STRAIGHT vocoder uses the mixed excitation parameters to model and generate the voice source signal. In the proposed method, the deterministic component or the segment of residual signal around GCI, which is important for the perception of speech is accurately represented. Hence, the generated excitation signal is close to the natural source signal.

Regarding the comparison of the proposed method with the two versions of DSM-HTS, it can be observed that the CMOS scores are varying in the range of 0.1–0.4 and the subjects preferred the proposed method for about 35% of cases and preferred the DSM-HTS for about 30% of cases. Both measures show that the proposed method is slightly better than the DSM-HTS. In the DSM-HTS, the overall residual frame is generated using either single or 20 eigenvectors. Among two versions of the DSM-HTS, the speech synthesized using 20 eigenvectors has higher quality. The main reason for this due to the utilization of 20 PCA coefficients for the generation of the deterministic

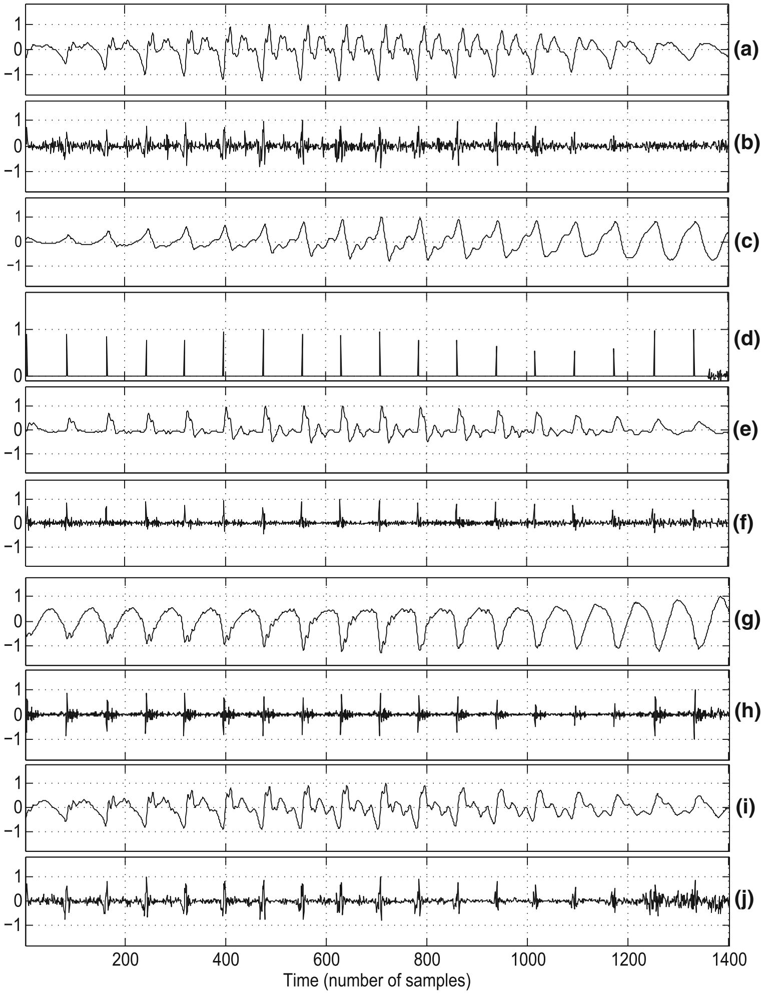


Fig. 13 **a** Natural speech, **b** excitation signal, **c** speech synthesized by pulse-based source model, **d** excitation signal constructed by pulse-based source model, **e** speech synthesized by STRAIGHT source model, **f** excitation signal constructed by STRAIGHT source model, **g** speech synthesized by DSM-based source model, **h** excitation signal constructed by DSM-based source model, **i** speech synthesized by proposed source model and **j** excitation signal constructed by proposed source model

component in every cycle of the residual frame. Utilization of 20 PCA coefficients can result in the reconstruction of about 60% (from Fig. 2) of the residual frame. But in the proposed method, the segment of the residual frame around GCI which is perceptually important is accurately (about 95%) generated for every residual frame. Accurate generation of the segment of the residual frame around GCI results in the incorporation of characteristics of real voice source in the excitation signal. Synthesized speech samples of the proposed and existing source models are made available online at http://www.sit.iitkgp.ernet.in/~ksrao/parametric-hts/pcd-hts.html.

In order to analyze the effectiveness of excitation models without any influence from statistical models, natural excitation signal is modeled using four source models, namely (1) pulse, (2) STRAIGHT, (3) DSM and (4) proposed method. Using natural spectrum, F0 and excitation signals constructed from four source models, the speech signals are synthesized. Figure 13 shows the natural speech, excitation signal, synthesized speech and corresponding excitation signals constructed from four source modeling methods. In every source model, after constructing the excitation signal, the energy contour is modified according to the target energy envelope. As energy is modified, the peak amplitude values in the excitation signal are varying in Fig. 13. In pulse-based source model, the sequence of pulses positioned according to pitch period is used as the excitation signal. As the energy of the excitation signal is modified according to the target energy envelope, the amplitudes of pulse excitation are varying in Fig. 13d. In addition to this, the excitation signal generated from the pulse-based source model (Fig. 13d) is having nonzero amplitude values only at GCIs. This kind of excitation signal is an imprecise approximation of natural excitation signal (Fig. 13b). From the excitation signal generated from STRAIGHT source model (Fig. 13f), it can be observed that in addition to nonzero amplitude values at GCIs, small amount noise is also present around GCIs. On comparing this signal with the natural excitation signal, significant differences can be observed in the waveform shapes of each pitch cycle. The excitation signal constructed from DSM-based source model (Fig. 13h) is closer to natural excitation compared to pulse and STRAIGHT source models. In DSM-based source model, the excitation signal is constructed by using single instance of deterministic component and noise component. The single instance of deterministic component and noise component is repeated for all cycles of excitation signal. The noise component is obtained by imposing average spectral and amplitude envelopes on the white Gaussian noise. As noise component is obtained from average spectral and amplitude envelopes, the excitation signal appears to be smoothly varying. Since single instance of deterministic component and noise component is used, cycle-to-cycle variations occurring in the natural excitation signal are not present in the excitation signal of the DSM-based source model. But in the proposed method, the excitation signal is uniquely constructed for every cycle of pitch-synchronous residual frame. This results in the incorporation of natural variation of the excitation signal. The excitation signal constructed from the proposed method (Fig. 13j) is very much closer to the natural excitation signal (Fig. 13b) compared to three other source models. The speech waveform produced by the proposed source model (Fig. 13i) is also very close to the natural speech waveform (Fig. 13a) compared to three other source models.

To understand the significance of segregating the regions of the residual frames based on their perceptual importance, instead of considering only the region around GCI, the entire length of the residual frame is considered as the deterministic component. The deterministic component is represented using 20 PCA coefficients. The noise component is obtained by subtracting the deterministic component from the residual frame. The deterministic and noise components are parameterized and modeled using the steps described in Sect. 2. The proposed method which considers the entire residual frame as the deterministic component (proposed method 2) is compared with proposed method which considers the region around GCI as the deterministic component (proposed method 1) and DSM-HTS (20 eigenvectors). Proposed method 1 is the original

Fig. 14 CMOS scores with 95% confidence intervals obtained by comparing the proposed method 2 with the proposed method 1 and DSM-HTS (20 eigenvectors)

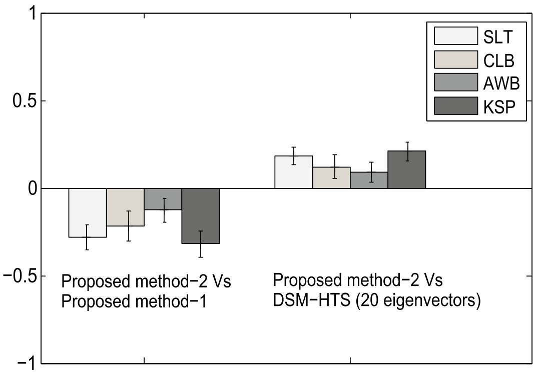
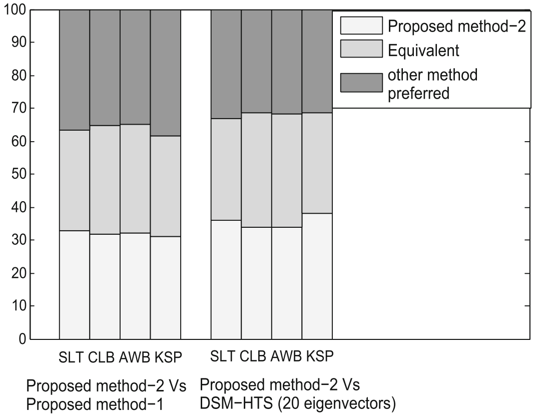


Fig. 15 Preference scores obtained by comparing the proposed method 2 with the proposed method 1 and DSM-HTS (20 eigenvectors)



proposed work, and the proposed method 2 is a variation of the proposed method 1. Here, proposed method 2 is compared with the DSM-HTS (20 eigenvectors) as both methods are closely related. CMOS and preference scores are used for comparison. In CMOS score, a positive score indicates proposed method 2 is better than other methods, and negative score indicated the opposite. Figures 14 and 15 provide the CMOS and preference scores, respectively. On comparison of proposed method 2 with DSM-HTS (20 eigenvectors), it can be observed that the proposed method 2 is slightly better. From the CMOS and preference scores, it can be concluded that the proposed method 1 is better than the proposed method 2. Segregating the regions of residual frame based on the perceptual importance has resulted in the improvement of the quality of synthesis.

In addition to subjective evaluation, objective evaluation of the quality of the synthesized speech is performed. The speech is synthesized using analysis–synthesis framework, where each of the speech utterances is parameterized using the proposed (both proposed method 1 and proposed method 2) and existing source modeling methods and using the parameters, the speech signal is reconstructed. Ten speech utterances synthesized from each of the source models are used for objective evaluation. The objective evaluation is carried out using the perceptual measure, ITU-T Rec. P.862

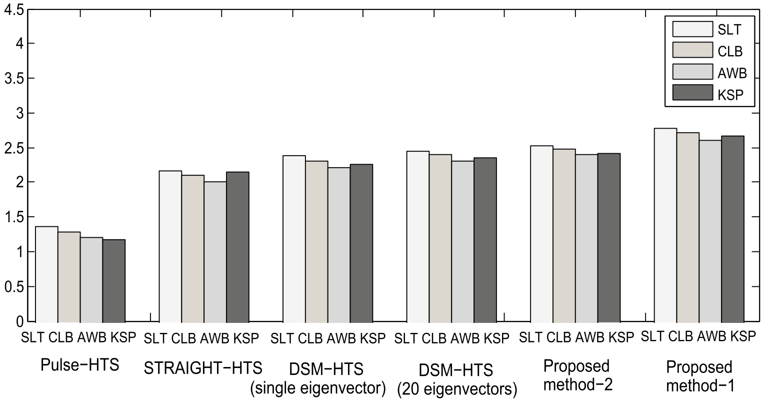


Fig. 16 PESQ values obtained for the proposed and existing source modeling methods

Perceptual Evaluation of Speech Quality (PESQ) [15,19]. The PESQ is specifically designed for the assessment of speech quality of the narrow-band telephone networks and speech codecs. In PESQ, audible difference between the reference and test signals is computed. In the context of speech synthesis, the reference signal is the natural speech waveform, and the test signal is the synthesized speech utterance from one of the source models. For a pair of speech utterances, the PESQ is a single value in the range -1 to 4.5. If PESQ value is closer to 4.5, then the synthesized speech is perceptually closer to the corresponding natural speech waveform. If the PESQ value is closer-1, then the synthesized speech is perceptually degraded compared to the corresponding natural speech waveform. The PESQ values obtained by comparing the natural speech utterances with the synthesized speech utterances of the proposed and existing source modeling methods are shown in Fig. 16. From Fig. 16, it can observed that the PESQ scores of the proposed method 2 and DSM-HTS (20 eigenvectors) are very close. The main reason for this is that both the proposed method 2 and DSM-HTS (20 eigenvectors) are conceptually very close. In both proposed method 2 and DSM-HTS (20 eigenvectors), the entire length of the residual frame is represented using 20 eigenvectors. From the figure, it can be observed that the proposed method 1 which is the original proposed work has the highest PESQ values compared to all other source models. This objectively proves that the proposed method 1 is perceptually better than other existing source models.

6 Conclusion

This paper proposes a parametric approach of modeling the excitation signal for improving the quality of HTS. Analysis of characteristics of the residual frames around GCI is performed using PCA. Based on the analysis, the segment of the residual frame around GCI is considered as deterministic component and the remaining part of the residual frame is considered as the noise component. The deterministic components are accurately modeled using PCA coefficients. 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. Both subjective and objective evaluation results indicated that the quality of the proposed method is considerably better compared to existing excitation modeling methods. In this work, PCA is performed by considering the residual frames of all phones. Instead, PCA can be performed on the residual frames of every phone and improvement in the quality of synthesized speech can be analyzed. The relation between time-domain and frequency-domain decomposition of excitation signal can also be analyzed.

References

1. N. Adiga, S.R.M. Prasanna, Significance of instants of significant excitation for source modeling, in *Proceedings of Interspeech* (2013), pp. 1677–1681
2. P. Alku, Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. Speech Commun. **11**(2–3), 109–118 (1992)
3. J.P. Cabral, S. Renals, J. Yamagishi, K. Richmond, HMM-based speech synthesiser using the LF-model of the glottal source, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2011), pp. 4704–4707
4. J.P. Cabral, Uniform concatenative excitation model for synthesising speech without voiced/unvoiced classification, in *Proceedings of Interspeech* (2013) pp. 1082–1086
5. CMU ARCTIC speech synthesis databases (**online**). http://festvox.org/cmu_arctic/
6. T.G. Csapó, G. Németh, A novel irregular voice model for HMM-based speech synthesis. in *Proceedings of ISCA Speech Synthesis Workshop* (2013), pp. 229–234
7. T.G. Csapó, G. Németh, Modeling irregular voice in statistical parametric speech synthesis with residual codebook based excitation. IEEE J. Sel. Top. Signal Process. **8**(2), 209–220 (2014)
8. T. Drugman, A. Moinet, T. Dutoit, G. Wilfart, Using a pitch-synchrounous residual codebook for hybrid HMM/frame selection speech synthesis, in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2009), pp. 3793–3796
9. T. Drugman, G. Wilfart, T. Dutoit, A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis, in *Proceeding of Interspeech* (2009), pp. 1779–1782
10. T. Drugman, G. Wilfart, T. Dutoit, Eigenresiduals for improved parametric speech synthesis, in *Proceedings of European Signal Processing Conference (EUSIPCO)* (2009), pp. 2177–2180
11. T. Drugman, T. Dutoit, The deterministic plus stochastic model of the residual signal and its applications. IEEE Trans. Audio Speech Lang. Process. **20**(3), 968–981 (2012)
12. T. Drugman, T. Raitio, Excitation modeling for HMM-based speech synthesis: breaking down the impact of periodic and aperiodic components, in *Proceedings of International Conference on Audio, Speech and Signal Processing (ICASSP)* (2014), pp. 260–264
13. HMM-based speech synthesis system (HTS) (**online**). http://hts.sp.nitech.ac.jp/
14. X. Huang, A. Acero, H.W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development* (Prentice Hall, Upper Saddle River, 2001)
15. ITU-T Draft Recommendation P.862, *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs* (2000)
16. H. Kawahara, I. Masuda-Katsuse, 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 Commun. **27**, 187–207 (1998)
17. H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, H. Banno, Tandem-STRAIGHT: a temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation, in *Proceeding of International Conference on Audio, Speech and Signal Processing (ICASSP)* (2008), pp. 3933–3936
18. S. Kim, J. Kim, M. Hahn, HMM-based Korean speech synthesis system for hand-held devices. IEEE Trans. Consum. Electron. **52**, 1384–1390 (2006)
19. P. Loizou, *Speech Enhancement: Theory and Practice* (CRC Press, Boca Raton, 2007)
20. S.L. Maguer, N. Barbot, O. Boeffard, Evaluation of contextual descriptors for HMM-based speech synthesis in French, in *Proceedings of ISCA Speech Synthesis Workshop* (2013), pp. 153–158

21. R. Maia, T. Toda, H. Zen, Y. Nankaku, K. Tokuda, An excitation model for HMM-based speech synthesis based on residual modeling, in *Proceeding of International Speech Communication Association Speech Synthesis Workshop 6 (ISCA SW6)* (2007), pp. 131–136
22. J.D. Markel, A.H. Gray, *Linear Prediction of Speech* (Springer, Berlin, 1976)
23. A. McCree, K. Truong, E. George, T. Barnwell, V. Viswanathan, A 2.4 kbit/s MELP coder candidate for the new U.S. Federal Standard, in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (1996), pp. 200–203
24. A. McCree, A 14 kb/s wideband speech coder with a parametric highband model, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2000), pp. 1153–1156
25. K.S.R. Murty, B. Yegnanarayana, Epoch extraction from speech signals. IEEE Trans. Audio Speech Lang. Process. **16**(8), 1602–1613 (2008)
26. N.P. Narendra, K.S. Rao, K. Ghosh, R.R. Vempada, S. Maity, Development of syllable-based text to speech synthesis system in Bengali. Int. J. Speech Technol. **14**(3), 167–181 (2011)
27. N.P. Narendra, K.S. Rao, K. Ghosh, V.R. Reddy, S. Maity, Development of Bengali screen reader using Festival speech synthesizer, in *Proceedings of IEEE India Conference (INDICON)* (2011), pp. 1–4
28. N.P. Narendra, K.S. Rao, Robust voicing detection and F0 estimation for HMM-based speech synthesis. Circuits Syst. Signal Process. **34**(8), 2597–2619 (2015)
29. N.P. Narendra, K.S. Rao, A deterministic plus noise model of excitation signal using principal component analysis for parametric speech synthesis, in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2016), pp. 5635–5639
30. J.J. Odella, *The Use of Context in Large Vocabulary Speech Recognition*. Ph.D. thesis, Cambridge University, Cambridge (1995)
31. K. Paliwal, W. Kleijn, Quantization of LPC parameters, in *Speech Coding and Synthesis*, ed. by W. Kleijn, E.K. Paliwal (Elsevier, Amsterdam, 1995)
32. Y. Pantazis, Y. Stylianou, Improving the modeling of the noise part in the harmonic plus noise model of speech, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4609–4612 (2008)
33. B. Picart, T. Drugman, T. Dutoit, HMM-based speech synthesis with various degrees of articulation: a perceptual study. J. Neurocomput. **132**, 142–147 (2014)
34. T.F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice* (Prentice Hall, Upper Saddle River, 2002)
35. E.V. Raghavendra, K. Prahallad, A multilingual screen reader in Indian languages, in *Proceedings of National Conference on Communications (NCC)* (2010), pp. 1–5
36. T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, P. Alku, HMM-based speech synthesis utilizing glottal inverse filtering. IEEE Trans. Audio Speech Lang. Process. **19**(1), 153–165 (2011)
37. T. Raitio, A. Suni, H. Pulakka, M. Vainio, P. Alku, Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis, in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2011), pp. 4564–4567
38. K. Shinoda, T. Watanabe, MDL-based context-dependent subword modeling for speech recognition. J. Acoust. Soc. Jpn. (E) **21**(2), 79–86 (2000)
39. F. Soong, B. Juang, Line spectrum pair (LSP) and speech data compression, in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (1984) pp. 37–40
40. Y. Stylianou, *Harmonic Plus Noise Models for Speech, Combined with Statistical Methods, for Speech and Speaker Modification*. Ph.D. thesis, Ecole Nationale Supérieure des Télécommunications (1996)
41. T. Toda, K. Tokuda, A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. IEICE Trans. Inform. Syst. **90**(5), 816–824 (2007)
42. K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura, Speech parameter generation algorithms for HMM-based speech synthesis, in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)* (2000), pp. 1315–1318
43. K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, K. Oura, Speech synthesis based on hidden Markov models. Proc. IEEE **101**(5), 1234–1252 (2013)
44. Z. Wen, J. Tao, S. Pan, Y. Wang, Pitch-scaled spectrum based excitation model for HMM-based speech synthesis. J. Signal Process. Syst. **74**(3), 423–435 (2013)
45. T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, Mixed-excitation for HMM-based speech synthesis, in *Proceedings of Eurospeech* (2001), pp. 2259–2262

46. E. Yumoto, W. Gould, T. Baer, Harmonics-to-noise ratio as an index of the degree of hoarseness. J. Acoust. Soc. Am. **71**(6), 1544–1550 (1982)
47. H. Zen, T. Toda, M. Nakamura, K. Tokuda, Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005. IEICE Trans. Inform. Syst. **E90-D**, 325–333 (2007)
48. H. Zen, T. Toda, K. Tokuda, The Nitech-NAIST HMM-based speech synthesis system for the Blizzard Challenge 2006. IEICE Trans. Inform. Syst. **E91-D**(6), 1764–1773 (2008)