J Sign Process Syst (2014) 74:423–435
DOI 10.1007/s11265-013-0862-z

Pitch-Scaled Spectrum Based Excitation Model for HMM-based Speech Synthesis

Zhengqi Wen · Jianhua Tao · Shifeng Pan · Yang Wang

Received: 9 October 2012 / Revised: 11 October 2013 / Accepted: 25 October 2013 / Published online: 19 December 2013

Abstract The speech generated by hidden Markov model (HMM)-based speech synthesis systems (HTS) suffers from a 'buzzing' sound, which is due to an over-simplified vocoding technique. This paper proposes a new excitation model that uses a pitch-scaled spectrum for the parametric representation of speech in HTS. A residual signal produced using inverse filtering retains the detailed harmonic structure of speech that is not part of the linear prediction (LP) spectrum. By using pitch-scaled spectrums, we can compensate the LP spectrum using the detailed harmonic structure of the residual signal. This spectrum can be compressed using a periodic excitation parameter so that it can used to train HTS. We define an aperiodic measure as the harmonics-to-noise ratio, and calculate a voicing-cut off frequency to fit the aperiodic measure to a sigmoid function. We combine the LP coefficient, pitch-scaled spectrum, and sigmoid function to create a new parametric representation of speech. Listening tests were carried out to evaluate the effectiveness of the proposed technique. This vocoder received a mean opinion score of 4.0 in analysis-synthesis experiments, before dimensionality reduction. By integrating this vocoder into HTS, we improved the sound of the synthesized speech compared with the pulse train excitation model, and demonstrated an even better result than STRAIGHT-HTS.

Z. Wen · J. Tao (✉) · S. Pan · Y. Wang
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Science, Beijing, China
e-mail: jhtao@nlpr.ia.ac.cn

Z. Wen
e-mail: zqwen@nlpr.ia.ac.cn

S. Pan
e-mail: sfpan@nlpr.ia.ac.cn

Y. Wang
e-mail: yangwang@nlpr.ia.ac.cn

Keywords Speech synthesis · HMM-based speech synthesis · Parametric representation of speech · Excitation model · Pitch-scaled spectrum

1 Introduction

An appropriate parametric representation of speech is a very important aspect of statistical parametric speech synthesis methods [1] such as hidden Markov model (HMM)-based speech synthesis (HTS) [2]. The main problem with speech generated using HTS is a 'buzzing' sound. It is caused by over-simplified vocoding techniques, such as the simple pulse train excitation model used in Linear Prediction Coding (LPC)-based vocoders [1].

Several high-quality parametric representations of speech exist. One is the harmonic plus noise model (HNM) proposed by Yannis [3]. In the HNM model, the speech spectrum is split into a low-frequency harmonic region and a high-frequency noisy region by the voicing cut-off frequency (VCO) [4]. Kawahara et al. proposed STRAIGHT, which is a speech transformation and representation method based on adaptive interpolation of a weighted spectrogram [5]. This vocoding technique extracts a spectral envelope without a periodic structure, in the time domain and the frequency domain. These two methods are effective for speech reconstruction and have been successfully integrated into concatenative speech synthesis [6] and HTS systems [7, 8]. STRAIGHT-HTS is a state-of-the art HTS synthesizer. However, HTS can still be improved. In [9], Cabral et al. incorporated the Liljencrants-Fant model (LF) [10] into STRAIGHT-HTS. They used the LF parameters to control the glottal source signal in HTS. This technique can produce flexible speech and better sound than STRAIGHT-HTS. In [11], Raitio et al. extracted the vocal-tract and glottal parameters separately, using iterative adaptive inverse filtering (IAIF) [12]. They used the original glottal

source pulses as the source signal for synthesizing. This technique has been shown to generate better male voice sounds than STRAIGHT-HTS. These two methods indicated that improving the excitation technique in the parametric representation could improve the quality of synthesized speech produced by HTS. In this paper, we propose an excitation model that improves the synthesized speech by keeping the detailed harmonic structure of the residual signal.

Some successful excitation models for HTS have been proposed. In [13], Yoshimura et al. integrated the mixed excitation linear prediction coder (MELP) [14] into HTS. In [15], Drugman et al. used principal component analysis (PCA) to de-correlate the residual frames of two-pitch periods. They took the PCA coefficient as the excitation parameter for the HTS training. In [16], Maia et al. proposed a trainable excitation model that directly minimized the weighted distortion between the generated excitation and residual signal. All of these methods have been shown to produce high-quality synthesized speech. However, little attention has been given to the detailed harmonic structures of speech or residual signal. For example, Maia et al. [16] only extracted an all-pole filter for the excitation signal, which focused on the formant structure but not the detailed harmonic structure. Yoshimura et al. [13] considered the harmonic structure in the mixed excitation model, but only the Fourier amplitudes of first ten pitch harmonics. Drugman et al. [15] kept a detailed harmonic structure in the time domain, but interpolation in the time domain introduces some energy holes in the frequency domain when synthesizing. Actually, keeping a detailed harmonic structure is very important in speech reconstruction, especially for female voices. In [17], Skoglund et al. noted that female voices sound much better than male voices in sinusoidal coders [18]. This is because they are generally very good at reconstructing the harmonic structure of speech, but do not model the pitch-cycle phase very accurately. This is consistent with Kawahara's findings (http://www.wakayama-u.ac.jp/~kawahara/phaseEffect/) that the phase is less audible in the high fundamental frequency than the low fundamental frequency. These results inspired us to develop a new parametric representation of speech that keeps the detailed harmonic structure of the residual signal.

Our proposed excitation model uses a pitch-scaled spectrum [19] of the residual signal, which retains its detailed harmonic structure. This excitation model is divided into two parts: periodic spectrum and aperiodic measure. The first part is motivated by the STRAIGHT-based vocoding technique and pitch-scaled analysis, and the second part uses the two-band idea proposed in HNM with the sigmoid function used in [20, 21].

We extract the periodic spectrum from the even line of the pitch-scaled spectrum which is calculated from the residual signal. This line retains the Fourier amplitude of the pitch harmonics. When synthesizing, the pitch-cycle periodic excitation is generated using zero-phase criterion [18]. For integration into HTS, these periodic spectrums are normalized to have constant length, and compressed for dimensionality reduction.

We define an aperiodic measure based on the harmonics-to-noise ratio (HNR) extracted from the pitch-scaled spectrum. We use the sigmoid function introduced in [21] to match the aperiodic measure, and carry out two modifications. One is to introduce the VCO used in HNM to smooth the trajectory of the aperiodic measure. The other is to replace the low section of the sigmoid function with a parabolic function to reduce the abrupt noise in the low-frequency region.

The remainder of this paper is organized as follows. In Section 2, we will introduce the fundamental techniques used in this paper such as the source-filter model and pitch-scaled analysis. In Section 3, we calculate the periodic spectrum and aperiodic measure using the pitch-scaled spectrum, which we use to propose a new parametric representation. In Section 4, we integrate the proposed excitation model (including periodic spectrum and aperiodic measure) into the HTS system. We describe our experiments in Section 5, which evaluated the effectiveness of the proposed excitation model in the direct analysis-synthesis and the integrated HTS system. Finally, we present our conclusions and future work in Section 6.

2 Fundamental Techniques

2.1 Source-Filter Model

The acoustic characteristics of speech are usually modeled as a sequence of source, vocal tract filter and radiation characteristics. In the source-filter theory of speech production models [22], the source spectrum represents the spectrum of typical glottal air flow using a fundamental frequency, and the filter models the effects of the vocal tract and lip radiation. In the frequency domain, this model can be represented using Eq. (1)

$$S(\omega) = P(\omega)D(\omega)G(\omega)V(\omega)L(\omega), \tag{1}$$

where $P(\omega)$ is a pre-emphasis filter, $D(\omega)$ is the Fourier transform (FT) of an impulse train, $G(\omega)$ is the FT of a glottal pulse, $V(\omega)$ is the vocal tract transfer function, and $L(\omega)$ are the lip radiation characteristics.

For the LPC-based vocoder shown in Fig. 1, the source-filter model can be simplified to

$$S(\omega) = D(\omega)H(\omega). \tag{2}$$

The filter $H(\omega)$ combines the effect of the glottis, vocal tract, and lip radiation with the excitation $D(\omega)$. Either a pulse train or white Gaussian noise is used.

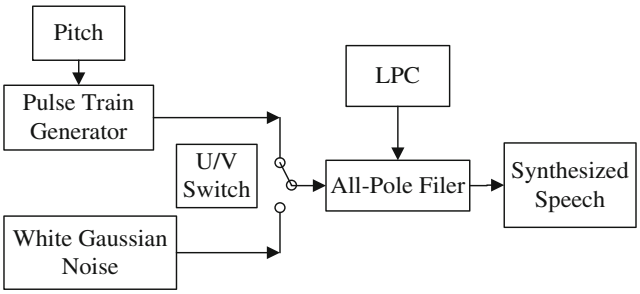


Figure 1 The LPC-based vocoder.

This simplified version of the source-filter model suffers from a 'buzzing' sound that is caused by the strong harmonic structure of voiced segments [1]. This is because the pulse train used as the excitation signal for the voiced segments is not consistent with the residual signal.

$$R(\omega) = S(\omega)/H(\omega) \tag{3}$$

Ideally, the residual signal derived from the inverse filtering in Eq. (3) should have a flat spectrum like the pulse train's spectrum. This is because the linear prediction (LP) spectrum, $H(\omega)$, combines the effects of pre-emphasis, glottis, vocal tract and lip radiation. However, the amplitude spectrum of the residual signal (shown at the bottom of Fig. 2) retains a complicated harmonic structure, and has a noisy structure in the high-frequency region. This is mainly because the LP spectrum only represents the auto-regressive components of natural speech, and the pulse train excitation model does not include the detailed harmonic structure of the residual signal. Thus, one way to improve the synthesized speech quality of the LPC-based vocoder is to keep the detailed harmonic structure of the residual signal and add noise to the voiced segments, especially in the high-frequency region.

2.2 Pitch-Scaled Analysis

A speech signal can be decomposed into a harmonic-related periodic component and a noise-related aperiodic component. In [23], Yegnanarayana et al. proposed an iterative algorithm in the frequency and time domains that decomposed speech signals into periodic and aperiodic components. In their algorithm, the harmonic region and the noisy region must be accurately identified. In [19], Jackson et al. proposed a harmonic filter (PSHF). It was based on a pitch-scaled analysis that separated the voiced and turbulence-noise components of speech signals, using an interpolation of the aperiodic spectrum. Here, pitch-scaled analysis is used to identify the harmonic region and noisy region.

Pitch-scaled analysis is an analysis framework that contains a small multiple of the pitch period. This technique is different from pitch-synchronous analysis, which is based on glottal closure instant (GCI) detection [24].

Let $s(k), k=1\ldots N$ be a residual frame with the length of two-pitch periods, and $S(n), n=1\ldots N$ be the corresponding discrete Fourier transform (DFT). The transformation uses the Hanning window [25] that has a largest side-lobe of −31.47 dB and an asymptotic decay of 18 dB/octave. Let

$$\begin{aligned} N &= 2 \times f_s/f_0, \\ f_k &= f_s \times k/N = f_s \times k/(2 \times f_s/f_0) = f_0 \times k/2, \end{aligned} \tag{4}$$

where f_0, f_s, and f_k are the fundamental, sampling, and *kth* point in $S(n)$ frequencies.

Equation 4 suggests that the frequency components, each indicated by a line in the spectrum shown in the middle of Fig. 2, can be divided into two groups. The first group consists of even-numbered frequency components, each corresponding to the integral multiple of fundamental frequency; the second group consists of odd-numbered frequency components, which can be considered to be aperiodic frequency components. Figure 2 shows an example of a residual frame

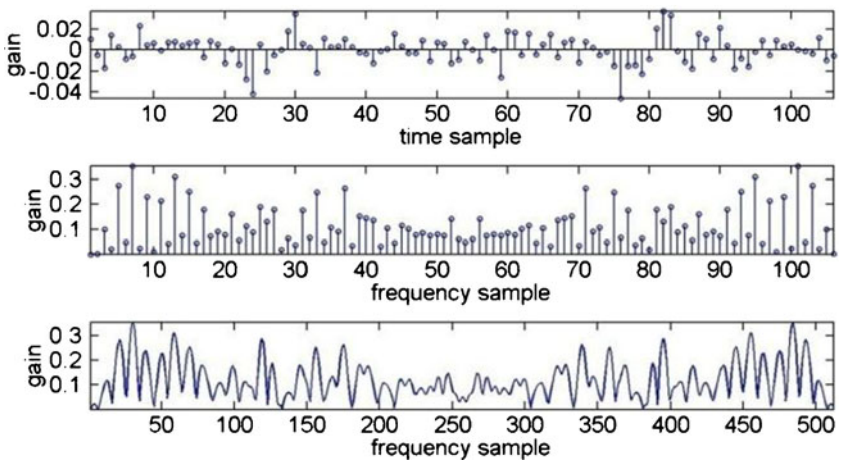


Figure 2 *Top*: a residual frame of two-pitch periods; *Middle*: Power spectrum of the residual frame of two-pitch periods; *Bottom*: Power spectrum of the residual frame with 512 points.

(top) and its corresponding DFT spectrum (the length of two-pitch periods in the middle, and 512 points in the bottom). From this figure, we can see that the harmonic energy is almost identical.

3 Proposed Excitation Model

The proposed excitation model consists of two parts: one represents the periodic spectrum and the other measures aperiodicity. Both are derived from the pitch-scaled spectrum. The periodic spectrum is compressed to reduce the dimensionality, and the aperiodic measure is fitted to a sigmoid function for integration into HTS.

3.1 Periodic Spectrum

3.1.1 Periodic Measure

The importance of phase in speech reconstruction was analyzed in [17] and (http://www.wakayama-u.ac.jp/~kawahara/phaseEffect/). It was found that harmonic reconstruction is much more important than pitch-cycle phase for voices with a high fundamental frequency, especially for female voices. Therefore, we first considered the detailed harmonic structure of the residual signal and ignored the phase information.

To keep the detailed harmonic structure of the residual signal, the periodic measure is typically defined by concatenating the peak points in the harmonic frequencies of the spectrum. These peak points (see the bottom of Fig. 2) can be identified using a peak-searching algorithm. However, the comparison in subsection 2.2 indicates that there is not a large difference in the harmonic frequencies of the normal and pitch-scaled spectrums. So an easy way of defining periodic measure is to directly concatenate the even-numbered frequency components, each corresponding to the integral multiple of fundamental frequency, of pitch-scaled spectrum.

3.1.2 Normalization

The goal of the proposed excitation model is to synthesize speech smoothing and so the extracted parameters for the parametric representation of speech should have an interpolative characteristic. However, the periodic measure extracted in subsection 3.1.1 requires different lengths for different pitch period values. The measure should first be interpolated so that it has constant length in the frequency domain, then interpolated in the time domain.

In [15], Drugman et al. used the following equation to choose the number of points for length-normalization.

$$F_0^* \le \frac{F_N}{F_m} \cdot F_{0,\min}, \tag{5}$$

where F_0^* is the normalized fundamental frequency, F_N is the useful band of the deterministic part, F_m is the Nyquist frequency, and $F_{0,\min}$ is the minimum pitch value of the considered speaker. Equation (5) ensures that energy holes will not appear during synthesis.

In order to avoid energy holes in the frequency domain, we use Eq. (5) in the frequency domain. Figure 3 shows the distribution of pitch period for 1,000 sentences from a female database and the corresponding accumulated ratio. The constant half pitch period length used in this paper is 128 points, which covers 99.66 % of the pitch period values in this database.

3.1.3 Dimensionality Reduction

In subsection 3.1.2, we normalized the periodic spectrum so that it had a constant length of 128. But this is still too large to incorporate into the speech synthesis system, so we must reduce the dimensionality to compress the normalized periodic spectrum. Two types of compression methods are commonly used: codebook and dimensionality reduction methods. In [26], Drugman et al. proposed a pitch-synchronous residual signal codebook for hybrid HMM/frame selection speech synthesis. In [27], Raitio et al. proposed a glottal source library for improving the excitation signal for HMM-based speech synthesis. Fodor listed a number of dimensionality reduction techniques in [28]. For example, principle component analysis (PCA) [29] is the best linear dimensionality reduction technique in terms of mean-square error. In [15], Drugman et al. used PCA to compress the pitch-synchronous residual frame, and proposed a deterministic plus stochastic model of the residual signal for improved parametric speech synthesis.

In this paper, we have used both methods for the proposed excitation model in our integrated HTS system, and conducted a detailed comparison. In [30], the authors used a codebook-based method. They extracted normalized periodic spectrums and used the Linde-Buzo-Gray (LBG) algorithm [31] to construct codebooks for every Mandarin final[1]. In the synthesis stage, they used a Viterbi-based searching algorithm to smooth the trajectory of the generated periodic spectrum. In the dimensionality reduction based method [32], the authors used PCA to de-correlate the periodic spectrums and reduce their dimension. We define the relative error as the Euclidean distance between the original periodic and the reconstructed periodic spectrums. The relative error for different numbers of eigenvectors is shown in Fig. 4. Using 16 eigenvectors can decrease the error by 30.79 %.

[1] Simple of compound vowel of Chinese syllable.

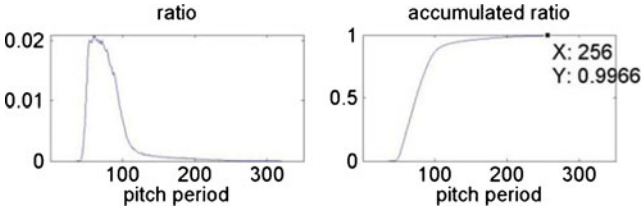


Figure 3 *Left*: pitch period distribution of 1,000 sentences; *Right*: the accumulated ratio of the pitch period distribution.

3.2 Aperiodic Measure

The aperiodic measure has a very important influence on the naturalness of the synthesized speech. We define it as the harmonics-to-noise ratio (HNR) in the frequency domain.

3.2.1 Aperiodicity Definition

In [4], Hermus et al. proposed an aperiodic measure based on the ratio of the aperiodic and periodic energies of the two-pitch scaled spectrum shown in Fig. 2. However, directly estimating the ratio as the solid line shown in Fig. 5 will introduce some vibration. Therefore, we expand the window and DFT lengths, and use two measures to define the aperiodicity.

Let M and N be the window and DFT lengths, which are both multiples of the pitch period. In this paper we have used 10 times the pitch period. So the harmonic region, P_i, and noise region, D_i, can be defined

$$\begin{aligned} P_i &= \{k | k_i - 2N/M \le k \le k_i + 2N/M\}, \\ D_i &= \{k | k_{i-1} + 2N/M \le k \le k_i - 2N/M\}, \end{aligned} \tag{6}$$

where k_i is i^{th} multiple fundamental frequency (which can be easily determined using pitch-scaled analysis), and $2N/M$ is the bandwidth of the window.

A triangle can be constructed from the peak value in one harmonic region and the peak values in the left and right neighboring noisy regions. The following equations define an area value and a symmetric score for this triangle.

$$\begin{aligned} Symmetry &= \left(P_{left} - P_{right}\right)/P_{harmonic}, \\ Area &= 2 \times \left(P_{harmonic} - P_{right}\right) - 0.5 \times 2 \times \left(P_{left} - P_{right}\right) \\ &\quad -0.5 \times \left(P_{harmonic} - P_{right}\right) - 0.5 \times \left(P_{harmonic} - P_{left}\right), \\ Aperiodicity &= Symmetry/Area, \end{aligned} \tag{7}$$

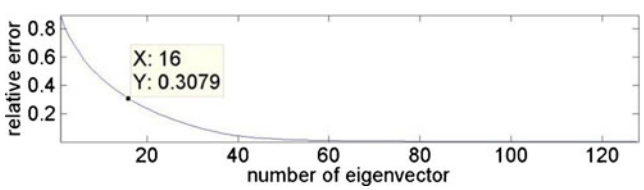


Figure 4 The error between the original periodic and reconstructed periodic spectrums for different numbers of eigenvectors.

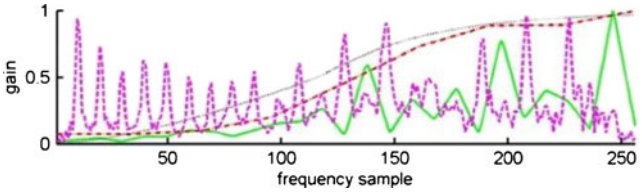


Figure 5 *Dotted line*: residual spectrum; *Solid line*: aperiod-to-period ratio in [4]; *Dash-dot line*: proposed aperiodic measure; *Dashed line*: sigmoid fitting result.

where $P_{harmonic}$ is the peak values in one harmonic region, and P_{left} and P_{right} are the peak values in the left and right neighboring noise regions. We define the aperiodic measure as the ratio of the symmetric score and the area value. This measure increases, shown as the dash-dot line in Fig. 5.

3.2.2 VCO Calculation

The aperiodic measure is normalized to the region [0,1] using the assumptions that a strong harmonic structure exists in the low-frequency region and a pure noisy structure exists in the high-frequency region. VCO is defined as the frequency where the aperiodic measure has the maximum slope, and can be estimated as the size of the smaller shaded area in the left of Fig. 6. The following equation defines the dash area, with an example shown in Fig. 6.

$$Dash(k) = sum(abs([0(1,k); 1(1, length - k + 1)] - Ap)), \tag{8}$$

where Ap is the aperiodic measure.

After a rough calculation of the VCO contour, we smoothed it using the Viterbi algorithm for time smoothing. The target score, T_Cost, and concatenate score, C_Cost, used in the Viterbi algorithm are defined as

$$\begin{aligned} T\_Cost(i,j) &= Dash(i,j), \\ C\_Cost(j,k) &= \exp(abs(j-k)) \times \alpha, \\ Score(i,j) &= \arg\min \begin{pmatrix} T\_Cost(i,j) + \\ C\_Cost(j,k) + \\ Score(i-1,k) \end{pmatrix}, \end{aligned} \tag{9}$$

where i is the frame index, j and k are the candidate indexes, and α is used to control the smoothness of the VCO contour.

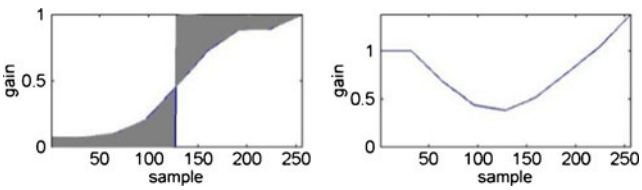


Figure 6 *Left*: Aperiodic measure and the corresponding dash area. *Right*: Dash area measure moving from left to right.

We calculate the smoothed VCO contour by minimizing the score.

3.2.3 Sigmoid Function Fitting

We used the sigmoid function introduced in [21] to fit the aperiodic measure in the following equation.

$$r(f) = \frac{(f/f_c)^{\alpha}}{1+(f/f_c)^{\alpha}}, \tag{10}$$

where α is a transition slope parameter, and f_c is a boundary frequency parameter.

We have made two modifications to Eq. (10). One is to replace f_c by the VCO calculated in subsection 3.2.2, which ensures that the aperiodic measure is smooth. The other is to replace the low part of the sigmoid function with a parabolic function to reduce the noise in the low-frequency region, as defined in Eq. (11).

$$r(f) = 0.5 \times (f/f_c)^{\beta} \qquad 0 \le f \le f_c, \tag{11}$$

where β is a slope parameter (1.5 in our experiments). An example is shown as a dashed line in Fig. 5.

3.3 Vocoder

The vocoder based on our proposed excitation model is an extension of the simplified version of the LPC-based vocoder. In the proposed excitation model, we have considered the harmonic structure and aperiodic measure to improve the naturalness of synthesized speech.

3.3.1 Analysis

In the analysis stage shown in Fig. 7, we segment the input speech into signal frames using a 5 ms long Hanning window. We extracted a STRAIGHT-based LPC [8] instead of directly estimating it from the auto-regressive coefficients [33]. We constructed an inverse filter using the LPC to generate the residual signal. We conducted pitch-scaled analysis on the two pitch-period long residual signal to extract the periodic spectrum. We used these periodic spectrums to construct codebooks for every Mandarin final, or compressed them using PCA to reduce the dimensionality. We extracted the aperiodic measure from the pitch-scaled spectrum using ten pitch-periods. We extracted VCO from this aperiodic measure, and fitted a sigmoid function.

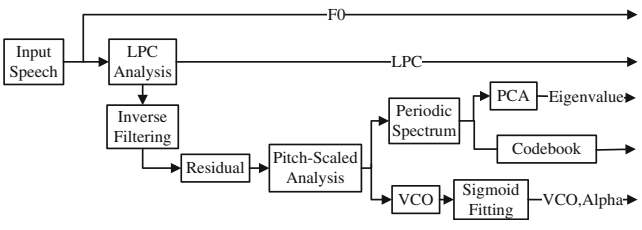


Figure 7 The analysis stage of the vocoder based on the proposed excitation model.

3.3.2 Synthesis

In the synthesis stage shown in Fig. 8, there are two ways to get the one pitch period spectrum. One is to search the codebooks using a Viterbi-based algorithm. Euclidean distance was used to measure the similarity between the synthesized periodic spectrum and the spectrum from the codebook. To ensure smoothness, we kept the 10 candidate spectrums with the maximum similarity scores and used the Viterbi algorithm to smooth the trajectory of the periodic spectrum by minimizing a cumulative score. The Viterbi algorithm uses the following equations:

$$\begin{aligned} T\_Cost\ (i,j) &= Score_{similarity}(i,j), \\ C\_Cost\ (j,k) &= \exp(abs(j-k)) \times \alpha, \\ Score(i,j) &= \arg \min_{1\le k\le 10} \begin{pmatrix} T\_Cost(i,j) + \\ C\_Cost(j,k) + \\ Score(i-1,k) \end{pmatrix}, \\ h_i &= \min_{1\le j\le 10} (Score(i,j)), \end{aligned} \tag{12}$$

where T_Cost is the target cost, C_Cost is the concatenation cost, $Score$ is the cumulative score, i is the frame index, j and k are the candidate indices, and h_i is the minimum cost from the first to the i^{th} frame.

This cumulative score includes a target cost that ensures that the algorithm is most likely to choose a periodic spectrum with the desired properties, and a concatenation cost that ensures the trajectory of periodic spectrums has no abrupt changes between adjacent frames.

The other way to generate the one pitch period spectrum is to generate the PCA coefficients. The eigenvector that was trained and kept as a codebook (as described in subsection 3.1.3) was used to synthesize the new periodic spectrum together with the PCA coefficient.

After we constructed the one pitch period spectrum, we used the inverse discrete Fourier transform (IDFT) with zero-phase criterion to synthesize the one pitch-cycle excitation signal. Then, the periodic excitation was generated from these one pitch-cycle excitation signals using the on overlap add (OLA) method.

The aperiodic measure was generated from the VCO and α using the sigmoid function. We constructed an IIR filter using this measure, which we used to filter the white Gaussian noise and generate the aperiodic excitation.

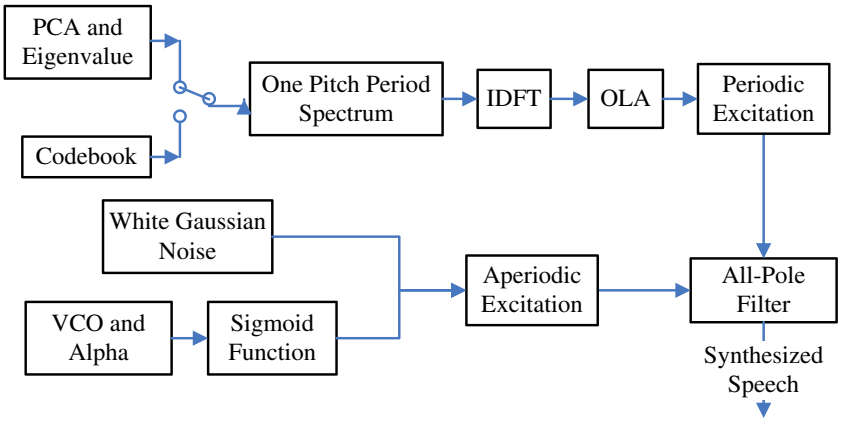


Figure 8 The synthesis stage of the vocoder based on the proposed excitation model.

Finally, the periodic and aperiodic excitations were added together to create the excitation signal. This excitation signal passes through an all-pole filter constructed using LPC to generate the speech.

4 Integration into HTS

4.1 Parameter Preparation

Before integrating the proposed excitation model into HTS, we must extract the speech parameters. The parameters used in HTS training are shown in the "PROPOSED" column of Table 1. The speech is represented by a fundamental frequency (F0), energy, vocal tract parameter, periodic spectrum parameter and aperiodic measure.

4.2 HTS Training

Our speech synthesis system (Fig. 9) uses the HTS toolkit available in [2]. Some modifications were needed to integrate it with the proposed excitation model.

Along with the speech features shown in Table 1, we also assumed that the corresponding contextual labels were available. These labels include contextual factors that would affect the spectral and excitation parameters. These factors include phone identity, stress and location-related factors [1]. The HTS was trained using four streams of the speech features. The vocal tract and aperiodic streams were treated as a single Gaussian distribution with diagonal covariance, similarly to the baseline HTS system. The F0 and periodic streams have zero values in unvoiced regions and should be treated differently in voiced and unvoiced regions. Multi-space-probability distributions (MSD) [34] were used to train these two streams. Finally, context dependent HMMs were trained and clustered using a tree-based context clustering technique with the minimum description length (MDL) principle [35].

Table 1 Speech features and the number of parameters in HTS training.

Feature	PROPOSED	PULSE_40	STRAIGHT	GLOTTAL
F0	1	1	1	1
Energy	1	1	1	1
Vocal Tract	24	40	40	30
Period	16	0	0	10
Aperiod	2	2	5	5
Total	44	44	47	47

4.3 HTS Synthesizing

In the synthesis stage (Fig. 9), the input text was analyzed and transformed into a context dependent label sequence. Then, we constructed context-dependent HMMs from the label sequence. After that, the parameter generation algorithm [36] was used to generate the vocal tract and excitation parameters from the sentence HMM. Finally, these parameters were converted into a speech signal using the vocoder described in Section 3.

5 Experiments

We evaluated the effectiveness of the proposed excitation model in the LPC-based vocoder in two ways. Firstly, we evaluated it using direct analysis-synthesis experiments, comparing it with the simple pulse train excitation model in the LPC-based vocoder and the STRAIGHT vocoding technique. Secondly, we assessed it as part of the integrated HTS system by comparing it with HTS systems based on the pulse train excitation model, the STRAIGHT vocoding technique, and the glottal inverse filtering technique [11].

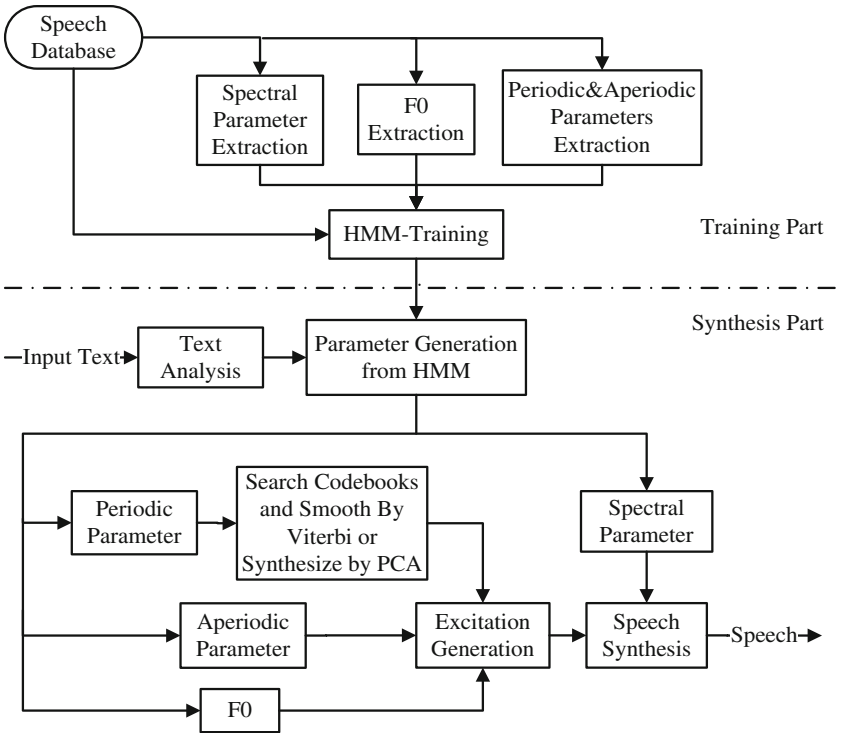


Figure 9 The workflow of HTS integrated with the proposed excitation model.

The corpora used in these experiments were two Mandarin voices recorded by professional announcers. The first was a female speaker who read approximately 5,000 sentences over 4 h, and the second was a male speaker who read approximately 5,000 sentences over 3 h. We sampled the corpora at 16 k Hz. The frame shift and length used for the analysis of speech parameters were 5 ms and 25 ms, respectively. We extracted F0 using the TANDEM algorithm [37].

We used the AB preference test and the mean opinion score (MOS) to compare the methods. Ten native speakers, who had previously undergone listening tests in our lab, took part in our experiments. We randomly selected 15 sentences (that were approximately 2 min long) as the test material. In the AB preference test, they were asked to listen to two versions of synthesized speech and choose the one that sounded best. In the MOS scoring test, they were asked to listen to the sentences and give a MOS according to Table 2.

Table 2 Mean opinion score.

MOS	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly Annoying
2	Poor	Annoying
1	Bad	Very annoying

5.1 Analysis-Synthesis Experiments

The excitation model used in the traditional LPC-based vocoder was a pulse train for the voiced segments, or white Gaussian noise for the unvoiced segments. An example of the LPC-based vocoder is shown in Fig. 1. The speech generated using this vocoder suffers from a 'buzzing' sound, which is due to the strong harmonic structure in the voiced segments. Our proposed excitation model makes two modifications. One is the addition of detailed harmonic structures from the residual signal to improve the naturalness of the synthesized speech. The other is the addition of noise in the high-frequency regions of the voiced segments.

The STRAIGHT vocoding technique is one of the most successful tools for speech reconstruction and modification. Therefore, a direct comparison with STRAIGHT gives a clear impression about the effectiveness of our proposed method. In the analysis stage of STRAIGHT, F0 is first analyzed for each speech frame, and then the spectral envelope is extracted by a pitch-adaptive time-frequency analysis of the FFT speech spectrum. The aperiodic measure is extracted from the ratio between the lower and upper spectral envelopes. In the synthesis stage, F0 is used to find the pulse position. For every pulse position, a spectral envelope is interpolated and a minimum-phase response is generated from this spectral envelope. These minimum-phase responses are added together to create a periodic signal using the pitch-synchronous overlap add (PSOLA) algorithm [38]. Then, the aperiodic signal is constructed from

the aperiodic measure and minimum phase filter. The speech signal is generated by adding these two parts together.

In this experiment, three versions of synthesized speech (PULSE, STRAIGHT and PROPOSED) were scored according to Table 2. The mean MOSs are shown in Fig. 10. The female voice synthesized by the LPC-based vocoder with the proposed excitation model achieved an MOS of 4.15, which is very close to the results for STRAIGHT (4.18). The pulse train excitation model only acheived an MOS of 3.56. However, our method performed worse for the male voice, and only achieved an MOS of 3.88. However, the proposed model is a clear improvement over the pulse excitation model for both voices, and a "buzzing sounding" was not audible. Our method does not compare well to the STRAIGHT vocoding technique, especially for the male voice. This is because STRAIGHT not only extracts detailed spectral information, but also uses an all-pass filter to control the fine pitch and temporal structure of the source signal [5]. The speech synthesized by the STRAIGHT vocoding technique is of a high quality for both male and female voices. Our proposed excitation model only considers the detailed harmonic structure of the residual signal and ignores the phase information. Phase information is a very important factor in the quality of synthesized speech, especially for the male voice [17]. Therefore, the degradation of our model for the male voice was obvious. We can improve the method by adding an all-pass filter used in mixed excitation linear prediction (MELP, [14]) coders. The coefficients used by the all-pass filter are listed in [14]. Listening tests showed that the all-pass filter improved the synthesized speech so that it had a MOS of 3.91. This demonstrates that the filter does improve the synthesized speech of the male voice, but it is still not comparable to the female voice.

5.2 HTS Integration Experiments

The integrating experiments were divided into four parts. First, we evaluated two compressing methods (PCA-based and codebook-based) for the periodic spectrum of the HTS system. Then we compared HTS combined with the proposed excitation model to the traditional HTS, which used the simple pulse train excitation model. Finally, we compared the proposed technique to the STRAIGHT vocoding and glottal inverse filtering techniques. We used AB preference tests, and passed the synthesized speech of the male voice through the all-pole filter mentioned in subsection 5.1.

5.2.1 Codebook-based vs. PCA-based

We evaluated the HTS systems based on the two different compressing methods, using only the female voice. The results of the AB preference test are shown in Fig. 11.

Figure 11 shows that there is no obvious difference between the codebook and PCA based methods. In theory, the codebook-based method should be more effective at keeping the dynamic structure of the periodic spectrum, because the periodic spectrum generated from the codebook is closer to the original than that constructed using PCA coefficients. However, this method suffers from a serious problem. There are abrupt changes in the constructed periodic spectrum even though we have used the Viterbi-based time smoothing method. This problem also arises when using the unit-selection based speech synthesis system [1]. The PCA-based method can handle this problem much better, and the reconstructed periodic spectrum is smooth. However, the accuracy of the PCA-based method (Fig. 4) is not as high as we expected. Furthermore, the PCA parameter does not have the interpolative property, which is very important in HTS training. Therefore, it is still possible to improve the method and more effective compression methods should be investigated.

5.2.2 Comparison with Traditional HTS

In this experiment, two HTS systems (PROPOSED and PULSE_40 in Table 1) were trained on the male and female voices. To keep the total number of speech parameters consistent, the LSF order for the proposed excitation model is 24 and the LSF order for the pulse train excitation model is 40.

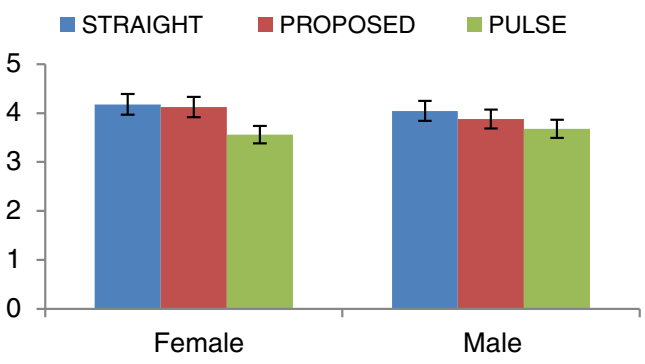


Figure 10 Mean opinion score obtained from three versions of vocoding techniques, for a female and male corpus.

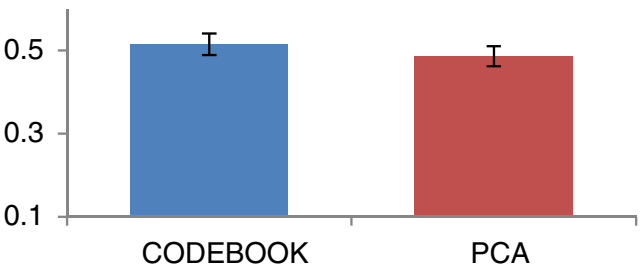


Figure 11 The mean preference scores between the codebook-based method and the PCA-based method for HTS.

We conducted AB preference tests for the female and male voices. The mean preference scores are shown in Fig. 12. The proposed excitation model shows a clear improvement over the pulse train excitation model for both the male and female voices. Keeping the detailed harmonic structure of the residual signal is an effective way to improve the synthesized speech quality of the HTS system, and is much better than directly increasing the LPC order.

5.2.3 Comparison with STRAIGHT-HTS

The STRAIGHT vocoding technique has been successfully integrated into HTS and is one of the most successful speech synthesizing engines [8]. In the analysis stage, the smooth STRAIGHT spectral envelope is converted into mel-generalized cepstrum (MGC) for every speech frame. The aperiodic measure is split into five frequency bands: [0–1], [1, 2], [2–4], [4–6], and [6–8]. Together with F0, the MGC, and the band-pass filtered aperiodic measure is used to train the HTS in three parametric streams. In the synthesis stage, the five band-pass filtered aperiodic measures are interpolated to be 513 points. The smooth STRAIGHT spectral envelope is synthesized from the generated MGC. Together with F0, these three parts are used by the STRAIGHT synthesizing engine to generate the speech signal.

In this comparison, we conducted AB preference tests for the female and male voices. The proposed excitation model was clearly better than the STRAIGHT vocoding technique, as shown in Fig. 13. This result is consistent with our comment in subsection 5.2.1 that improving the excitation technique is better for improving the synthesizing speech quality of HTS than directly increasing the number of spectral parameters. This is also true for the STRAIGHT vocoding technique, because the spectral parameter is included in STRAIGHT-HTS but not the detailed harmonic structure. Our proposed excitation model based on the spectrum of the residual signal considered the detailed harmonic structure and retained more spectral information.

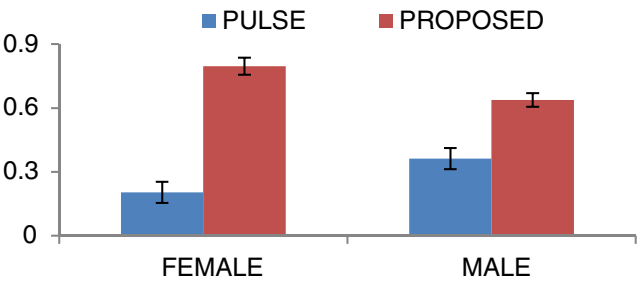


Figure 12 The mean preference scores between the proposed excitation model and pulse train excitation model in HTS.

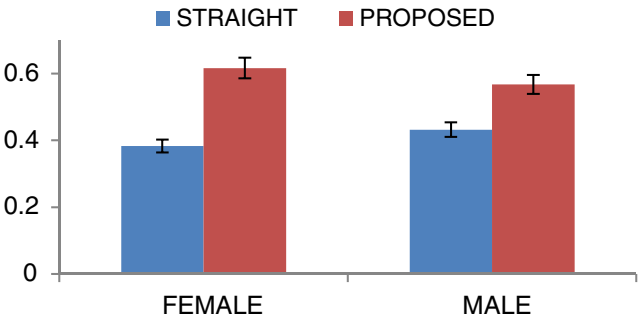


Figure 13 The mean preference scores between the proposed excitation model and STRAIGHT vocoding technique in HTS.

5.2.4 Comparison with GLOTTAL-HTS

In [11], the authors reported a new excitation model for HTS based on glottal inverse filtering. In the analysis stage, input speech is split into frames and windowed. They use the iterative adaptive inverse filtering (IAIF) [12] algorithm. This method can automatically decompose voiced speech into the vocal tract transfer function and the glottal source. The outputs of this algorithm are a p^{th} order all-pole model (the vocal tract filter), and a g^{th} order all-pole model (a parametric model of the spectral envelope of the estimated glottal flow). These models are converted into line spectral pair frequencies (LSF) [39], and used to train the HTS as the vocal tract and excitation parameters. They made two modifications to the synthesis stage of the conventional synthesis methods. Firstly, they used natural glottal flow pulses extracted from the glottal source to create the voiced excitation signal. Secondly, they modified the spectral properties of the excitation signal using an adaptive infinite impulse response (IIR) filter to reproduce time-varying changes in the real voice source and preserve the original quality. The experimental results reported in [11] showed a clear improvement over STRAIGHT-HTS for the male voice.

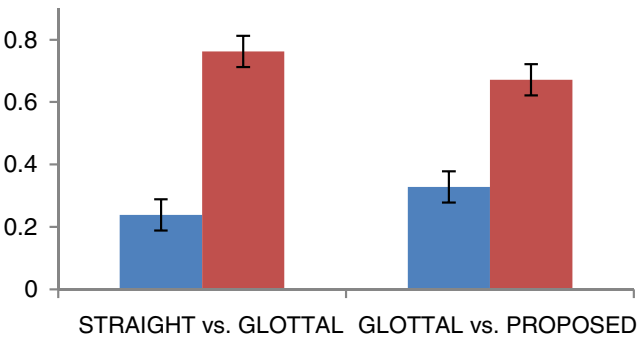


Figure 14 The mean preference scores between the STRAIGHT vocoding technique and the glottal inverse filtering technique, and between the glottal inverse filtering technique and proposed excitation model for the male voice in HTS.

In this experiment, we trained an HTS system based on glottal inverse filtering using only the male voice. We conducted AB preference listening tests to compare the proposed excitation model with the glottal inverse filtering technique. One test confirmed the effectiveness of the glottal inverse filtering technique in HTS by comparing it with STRAIGHT-HTS. Then, we compared the glottal inverse filtering technique with our method. The mean preference scores are shown in Fig. 14.

The results shown in Fig. 14 are consistent with the experimental results reported in [11]. The glottal inverse filtering technique generated better sounds than STRAIGHT-HTS. However, our technique generated even better sounds than the glottal inverse filtering technique. The excitation information in the glottal inverse filtering technique is the LPC parameter extracted from the glottal flow and pulse signals. The phase information retained in the glottal pulse signal has vastly improved the quality of the synthesized speech. But it has not taken the detailed harmonic structure of the glottal flow signal into consideration. Our proposed technique considers both, and so the result is even better.

6 Conclusion

Speech generated using the simple pulse-train excitation model in LPC-based vocoders and typical HTS systems suffers from a "buzzing" sound. This is mostly due to the strong harmonic structure that exists in the voiced segments. In this paper, we proposed a new excitation model based on the pitch-scaled spectrum of the residual signal for LPC-based vocoders. Our excitation model makes two modifications to the simple pulse train model in LPC-based vocoders. We added detailed harmonic structures to the excitation signal to improve the naturalness of the synthesized speech. We also added noise to the voiced segments, especially the high-frequency regions.

The residual signal derived from inverse filtering keeps some of the detailed harmonic structure of speech signals that is not included in the linear prediction (LP) spectrum and pitch-scaled spectrum. This can extract the harmonic structure of a signal to complement the LP spectrum. Therefore, we normalized the pitch-scaled spectrum, which we called the periodic spectrum. We also proposed an aperiodic measure based on a pitch-scaled spectrum, which we calculated as the energy ratio between the periodic and aperiodic regions. Together with LPC and F0, the pitch-scaled spectrum and aperiodic measure compose a new parametric representation of speech. The effectiveness of the new vocoding technique was tested in two groups of listening tests: one in direct analysis-synthesis experiments and the other in integrated HTS systems.

In the analysis-synthesis experiments, we compared the proposed vocoding technique with the simple pulse-train excitation model in the LPC-based vocoder and the STRAIGHT vocoding technique. The results of listening tests indicated that the proposed excitation improved the simple pulse-train model in the LPC-based vocoder, and that it generated high-quality speech equal to the STRAIGHT vocoding technique. These results indicated the effectiveness of the pitch-scaled spectrum in the LPC-based vocoder, and the effectiveness of the proposed parametric representation of speech.

In the integrated HTS systems, this vocoding technique was compared with the pulse train excitation model, STRAIGHT vocoding technique, and the glottal inverse filtering technique. The results of listening tests demonstrated a clear improvement over the pulse train excitation model and the STRAIGHT vocoding technique. This is because the proposed excitation model considers the detailed harmonic structure of the residual signal, but the other HTS systems only consider the vocal tract parameter. The HTS with the glottal inverse filtering technique took the natural glottal flow as a source signal in the synthesis stage, and the results were very promising. But our proposed vocoding technique could generate even better sound than the glottal inverse filtering technique. This is because sampling in the time domain introduces some energy holes, which does not occur when directly sampling the frequency domain.

Listening results confirmed the effectiveness of the proposed vocoding technique in the analysis-synthesis process and the integrated HTS system. In addition, this technique provides a new parametric representation of speech. However, the proposed method has some shortcomings. It does not accurately reconstruct the periodic spectrum. We evaluated PCA and codebook based methods. PCA parameters do not have an interpolative characteristic, which is very important when training HTS. In addition, there were some abrupt changes in the codebook-based method. Another disadvantage of our method is that there is no effective phase model included in the proposed vocoding technique, which is very important for the reconstruction of male voices. In addition, our method does not take into account changes from the unvoiced segment to the voiced segment.

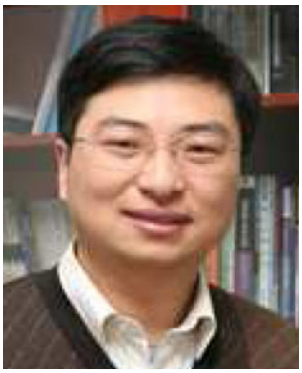
In future work, we will consider these three shortcomings. We will investigate more effective methods for reconstructing the periodic spectrum such as nonlinear dimensionality reduction methods. We will include a phase model into the proposed vocoding technique, and consider the special segment different from the unvoiced and voiced segments.

Acknowledgments The work was supported by NSFC-JSPS joint project (No.61011140075), China-Singapore Institute of Digital Media (CSIDM) and National Science Foundation of China (No.61273288, No.61233009, No.60873160, No.90820303 and No.61203258). The authors would like to thank Hideki Kawahara. I learned a lot during my three month stay in Prof. Kawahara's lab and his direction has significantly helped me in preparing this research.

References

1. Zen, H., Tokuda, K., & Black, A. (2009). Statistical parametric speech synthesis. *Speech Communication, 51*(11), 1039–1064.
2. [online] HMM-based Speech Synthesis System (HTS). http://hts.sp.nitech.ac.jp/.
3. Stylianou, Y. (1996). Harmonic plus Noise Model for Speech, combined with Statistical Methods, for Speech and Speaker Modification. *Ph.D. thesis, Ecole Nationale Supèrieure des Télécommunications*. Paris, France.
4. Hermus, K., Van Hamme, H., & Irhimeh, S. (2007). Estimation of the voicing cut-off frequency contour based on a cumulative harmonicity score. *IEEE Signal Processing Letters, 14*(11), 820–823.
5. Kawahara, H., Masuda-Katsuse, I., & de Cheveigné, A. (1999). 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, 27*(5), 187–207.
6. Stylianou, Y. (2001). Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Transactions on Speech Audio Processing, 9*(1), 21–29.
7. Hemptinne, C. (2006). Integration of the harmonic plus noise model (HNM) into the hidden markov model-based speech synthesis, system (HTS). *Master thesis*. IDIAP Research Institute, IDIAP-RR 69, Switzerland.
8. Zen, H., Toda, T., Nakamura, M., & Tokuda, K. (2007). Details of the Nitech HMM-based speech synthesis for Blizzard Challenge 2005. *IEICE Transactions on Information and Systems, E90*(D), 325–333.
9. Cabral, J. P., Renals, S., Yamagishi, J., & Richmond, K. (2011). HMM-based Speech Synthesizer Using the LF-model of the Glottal Source. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4704–4707.
10. Fant, G., Liljencrants, J., & Lin, Q. (1985). *A four-parameter model of glottal flow*. Stockholm: STL-QPSR, KTH.
11. Raitio, T., Suni, J., Yamagishi, H., Pulakka, A., Nurminen, J., Vainio, M., & Alku, P. (2010). HMM-based speech synthesis utilizing glottal inverse filtering. *IEEE Transactions on Speech Audio Processing, 19*(1), 153–165.
12. Plumpe, M. D., Quatieri, T. F., & Reynolds, D. A. (1999). Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Transactions on Speech Audio Processing, 7*(5), 569–585.
13. Yoshimura, T., Tokuda, K., Masuko, T., & Kitamura, T. (2001). Mixed excitation for HMM-based speech synthesis. *9th European Conference on Speech Communication and Technology*, 2263–2266.
14. Macree, A. V., Truong, K., George, E. B., Barnwell, T. P., & Viswanathan, V. (1996). A 2.4 kbitsfs MELP Coder Candidate for the New US. Federal Standard. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 200–203.
15. Drugman, T., Wilfart, G., & Dutoit, T. (2009). A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis. *Proceedings of Interspeech, 2009*, 1779–1782.
16. Maia, R., Toda, T., Zen, H., Nankaku, Y., & Tokuda, K. (2007). An excitation model for HMM-based speech synthesis based on residual modeling. *6the ISCA Workshop on Speech Synthesis*, 131–136.
17. Skoglund, J., & Bastiaan, W. K. (2000). On time-frequency masking in voiced speech. *IEEE Transactions on Speech Audio Processing, 8*(4), 361–369.
18. Robert, J. M., & Thomas, F. Q. (1986). Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Speech Audio Processing, 4*(34), 744–754.
19. Jackson, P. J. B., & Shadle, C. H. (2001). Pitch-scaled estimation of simultaneous voiced and trubulence-noise components in speech. *IEEE Transactions on Speech Audio Processing, 9*(7), 713–726.
20. Wen, Z. Q., & Tao, J. H. (2011). Inverse filtering based harmonic plus noise excitation model for HMM-based speech synthesis. *Proceedings of Interspeech, 2011*, 1805–1808.
21. Kawahara, H., Morise, M., Takahashi, T., Banno, H., Nisimura, R., & Irino, T. (2010). Simplification and extension of non-periodic excitation source representations for high-quality speech manipulation systems. *Proceedings of Interspeech, 2010*, 38–41.
22. Fant, G. (1960). *Acoustic Theory of Speech Production*. The Hague: Mouton.
23. Yegnanarayana, B., d'Alessandro, C., & Darsinos, V. (1998). An iterative algorithm for decomposition of speech signals into periodic and aperiodic components. *IEEE Transactions on Speech Audio Processing, 6*(1), 1–11.
24. Naylor, P., Kounoudes, A., Gudnason, J., & Brookes, M. (2007). Estimation of glottal closure instants in voiced speech using the DYPSA algorithm. *IEEE Transactions on Speech Audio Processing, 15*(1), 34–43.
25. Nuttal, A. H. (1981). Some windows with very good sidelobe behavior. *IEEE Transactions on Acoustics, Speech and Audio Processing, 29*(1), 84–91.
26. Drugman, T., Moinet, A., Dutoit, T., & Wilfart, G. (2009). Using a pitch-synchronous redisual codebook for hybrid HMM/frame selection speech synthesis. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3793–3796.
27. Raitio, T., Suni, A., Pulakka, H., Vainio, M., & Alku, P. (2011). Utilizing glottal source pulse library for generation improved excitation signal for HMM-based speech synthesis. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4564–4567.
28. Fodor, I. K. (2002). A survey of dimension reduction techniques. *Technical Report UCRL-ID-148494, Lawrence Livermore National Laboratory*. Center for Applied Scientific Computing, USA.
29. Jackson, J. E. (1991). *A User's Guide to Principal components*. New York: John Wiley and Sons.
30. Wen, Z. Q., & Tao, J. H. (2011). An excitation model based on inverse filtering for speech analysis and synthesis. *IEEE International Workshop on Machine Learning for Signal Processing*.
31. Linde, Y., Buzo, A., & Gray, R. M. (1980). An algorithm for vector quantizer design. *IEEE Transaction on Communications, 28*(1), 84–85.
32. Wen, Z. Q., Tao, J. H., & Hain, H. U. (2012). Pitch-scaled spectrum based excitation model for HMM-based speech synthesis. *IEEE 11th International Conference on Signal Processing*.
33. John, M. (1975). Linear prediction: a tutorial review. *Proceedings of the IEEE, 63*(4), 561–580.
34. Tokuda, K., Masuko, T., Miyazaki, N., & Kobayashi, T. (2002). Multi-space probability distribution HMM. *IEICE Transactions on Information and Systems, E85-D*(3), 455–464.
35. Shinoda, K., & Watanabe, T. (2000). MDL-based context-dependent subword modeling for speech recognition. *The Journal of the Acoustical Society of Japan (e), 21*(2), 79–86.
36. Tokuda, K., Kobayashi, T., & Imai, S. (1995). Speech parameter generation from HMM using dynamic features. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 660–663
37. Kawahra, H., Morise, M., Takahashi, T., Nisimura, R., & Irino, T. (2006). Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. *Proceedings of ICASSP*, 3933–3936.
38. Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication, 9*(5), 453–467.
39. Itakura, F. (1975). Line spectrum representation of linear predictor coefficients of speech signals. *Journal of the Acoustical Society of America, 57*(S1), S35–S35.

Zhengqi Wen received the B.S. degree from the Department of Automation, University of Science and Technology of China (USTC), Hefei, in 2008. He is currently pursuing the Ph.D. degree in the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing. From March 2009 to June 2009, he was an intern student in Nokia Research Center, China. From December 2011 to March 2012, he was an intern student in the Faculty of Systems Engineering, Wakayama University, Japan. In 2013, he was an Assistant Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests include speech analysis and speech synthesis.



Jianhua Tao (M'98) received the M.S. degree from Nanjing University, Nanjing, China, in 1996 and the Ph.D. degree from Tsinghua University, Beijing, China, in 2001. He is currently a Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing. His current research interests include speech synthesis and recognition, human–computer interaction, and emotional information processing. He has published more than 60 papers in major journals and proceedings, such as the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, ICASSP, Interspeech, ICME, ICPR, ICCV, ICIP, etc. In 2006, he was elected as Vice-Chairperson of the ISCA Special Interest Group of Chinese Spoken Language Processing (SIG-CSLP), and Executive Committee member of the HUMAINE association. He is the subject editor for the Speech Communication (SPEECH COMMUN), the Editorial Board Member for the Journal on Multimodal User Interfaces (JMUI), the International Journal of Synthetic Emotions (IJSE), and the Steering Committee Member for the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING.

Shifeng Pan received the B.S. and M.S. in electrical engineering from Southeast University, Nanjing, China, in 2002 and 2005, respectively, and Ph.D. degree in Pattern Recognition and Intelligent System from Institute of Automation, Chinese Academy of Sciences, Beijing China, in 2011. From 2011 to 2013, he was an Assistant Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. He joined Microsoft China in 2013, where he is currently a Speech Scientist. His research interest is in the area of speech synthesis, speech signal processing.

Yang Wang received the B.E. and M.E. degrees in computer science from Xi'an Jiaotong University, Xi'an, in 2006 and 2009, respectively. He is currently pursuing the Ph.D. degree in the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include statistical speech synthesis and statistical machine learning.