Chiang *EURASIP Journal on Audio, Speech, and Music Processing* (2018) 2018:5
https://doi.org/10.1186/s13636-018-0129-5

EURASIP Journal on Audio,
Speech, and Music Processing

RESEARCH Open Access

A parametric prosody coding approach for Mandarin speech using a hierarchical prosodic model

Chen-Yu Chiang

Abstract

In this paper, a novel parametric prosody coding approach for Mandarin speech is proposed. It employs a hierarchical prosodic model (HPM) as a prosody-generating model in the encoder to analyze the speech prosody of the input utterance to obtain a parametric representation of four prosodic-acoustic features of syllable pitch contour, syllable duration, syllable energy level, and syllable-juncture pause duration for encoding. In the decoder, the four prosodic-acoustic features are reconstructed by a synthesis operation using the decoded HPM parameters. The reconstructed prosodic features are lastly used in an HMM-based speech synthesizer to generate the reconstructed speech. Objective and subjective evaluations showed that the proposed prosody coding approach encoded speech with better quality and lower data rate than the conventional segment-based coding scheme with vector or scalar quantization approach did. The reconstructed speech encoded by the proposed approach has good quality at low data rates of 81.4 and 72.7 bps for speaker-dependent and speaker-independent tasks, respectively. An application of the proposed prosody coding approach to speaking rate conversion by directly changing the HPM parameters to those of a different speaking rate is also illustrated. An informal listening test confirmed that both converted speeches of high and low speaking rate sounded very smooth.

Keywords: Parametric prosody coding, Hierarchical prosodic model, Speaking rate conversion

1 Introduction

Speech coding is a process to transform a digitized speech signal into a bit-efficient representation that keeps reasonable speech quality so as to facilitate speech transmission over a band-limited channel or speech storage in a memory-limited media. In general, speech coding techniques can be classified into three categories, including waveform coding, parametric coding, and hybrid coding. The waveform coding technique attempts to maintain the waveform shape of the original speech signal in sample level without any knowledge about the speech generation process. Famous standard speech coders of this category are G.711 A-law and μ-law Pulse Code Modulation (PCM) coders [1], and G.726 and G.727 Adaptive Differential PCM coders [2]. Generally, a waveform coder works well at a high bit rate of 32 kbps or above. The parametric coding technique represents a speech signal by parameters of a speech-generating model. Among various speech-generating models, the most successful one is the linear predictive coding (LPC) model that assumes speech signal is the output of an all-pole model (autoregressive model) fed with an excitation input signal. The parameters of the all-pole filter conceptually represent the vocal tract shape that is highly correlated with the spectral envelope of the speech, while the excitation signal uses a quasiperiodic impulse train to represent information of fundamental frequency (or F0) for voiced speech, pseudorandom noise for unvoiced speech, or a combination of the two (i.e., mixed excitation). Coders of this type encode speech signal in a frame-based processing manner and could operate at low bit rates ranging from 2 to 5 kbps. Differing from waveform coders, parametric coders make no attempt to preserve the original waveform shape but to keep the perceptual quality of the reconstructed speech. Famous standard LPC-based speech coders are FS 1015 LPC of LPC-10e algorithm [3, 4] and

Correspondence: cychiang@mail.ntpu.edu.tw
Department of Communication Engineering, National Taipei University, New Taipei City, Taiwan

MELP (mixed excitation linear prediction) [5]. The hybrid coding technique tries to combine the advantages of both waveform coding and parametric coding. Coders of this type are similar to parametric coders in utilizing speech-generating models, but also similar to waveform coders in keeping encoded speech waveforms close to the original ones by more detailed modeling of the excitation signal. They generally adopt the code-excited linear prediction (CELP) algorithm [6] to minimize perceptually weighted error. Representative standard hybrid coders are FS1016 CELP [7, 8], ITU-T G.728 LD-CELP [9], ETSI AMR-ACELP [10], etc. A hybrid coder generally operates at a medium bit rate of 5 to 15 kbps.

To further encode speech at a very low bit rate of less than 2 kbps, the abovementioned existing sampled-based or frame-based speech coders are unable to obtain reconstructed speech with good intelligibility and naturalness due to the loss of modeling accuracy at such a low data rate. Therefore, segment-based speech coders, such as segment vocoders [11–23], phonetic vocoders [24–33], and text-to-speech (TTS)-based speech coders [34–36], were proposed to overcome this limitation. Those coders generally process speech at segment level instead of sample or frame level. Generally, a segment vocoder [11–23] firstly divides the speech signal into a sequence of fixed- or variable-length segments by a speech segmenter or a speech recognizer and then quantizes each segment by a codebook of pre-stored speech segments. A segment vocoder of fixed length simply quantizes a sequence of speech segments of l frames ($l > 1$) by matrix quantization (MQ) [14, 22]. In a variable-length segment vocoder [11–13, 15–23], segmental unit is usually pre-determined in the design of speech segmenter or speech recognizer. General segmental units can be phones, di-phones, syllables, or automatically derived acoustic units. Most segment vocoders reported in literatures operated at very low bit rates for speaker-dependent speech coding. However, to apply segment vocoders to speaker-independent speech coding, higher bit rate may be required due to the use of a larger codebook to capture the speaker variability. A phonetic vocoder [24–33] adopts a recognition-synthesis scheme in which speech is delimited into a sequence of phonetic segments (usually phone or phone-like units) by a speech recognizer, and reconstructed by a TTS synthesizer given with the recognized phone identities and their corresponding quantized prosodic features. Generally, the speech quality of this type coder is subject to the performance of the speech recognizer and the TTS synthesizer used. Due to the fact that phonetic vocoders do not encode the information of speaker, the speaker identity is missing in the reconstructed speech. Therefore, phonetic vocoders are suitable for speaker-dependent speech coding. A TTS-based speech coder [34–36] adopts a TTS synthesizer to generate speech from a given text by concatenating speech units properly selected from a large speech inventory and modifying the prosody of the selected speech units to the quantized prosodic parameters. A speech coder of this type can be viewed as a phonetic vocoder which operates in an oracle status that the correct text is given, and all speech segments are well segmented with correct phonetic transcriptions, i.e., speech is segmented by forced alignment. Although segment-based speech coders generally operate with longer coding delay and higher computational complexity than the conventional sample- and frame-based coders, they are potentially very useful in some applications that require a large amount of pre-recorded speech with limited memory space, such as the speech coding of story readings in an electronic book, computer-assisted language learning systems and electronic dictionaries, and saving speech in a matrix bar code, i.e., quick response (QR) code.

Concluding from above discussions, we find that those previous studies mainly focused on the modeling or encoding of spectral information. For frame-based speech coding, one milestone was the use of vector quantization (VQ) in encoding LPCs [37] or line spectral frequencies (LSFs) [38, 39] to greatly reduce the bit amount for spectral information via taking advantage of high intra-frame correlation among LPC/LSF coefficients. Predictive VQ [40] was proposed to further reduce the bit-rate by using the property of inter-frame spectral redundancy or correlation. For segment vocoders [11–23], the main study issues focused on the choice of segmental units, the realization of segmentation and segment quantization, and the design of segment codebook. For phonetic vocoders [24–33], the studies mainly focused on the choice of acoustic unit for speech recognition/synthesis [24–27] and speaker adaptation of spectral information [28, 30, 32, 33]. For TTS-based vocoders [34–36], the main concern lay in the methods of unit selection for speech synthesis. On the other hand, encoding of the prosodic information of speech signal was rarely addressed. Prosody refers to certain inherent suprasegmental properties of speech signal that carry melodic, timing, rhythmic, and pragmatic information of continuous speech. Prosodic features are physically represented by any domain's (generally phone, syllable, word, phrase, sentence, etc.) variations on pitch contour, energy level, duration, and silence of spoken utterances. In conventional speech coding methods, prosodic features are generally ignored, or simply scalar- or vector-quantized. For sample-based waveform coding [1, 2], no prosodic features are needed to be encoded owing to the fact that a waveform coder attempts to maintain the waveform shape of the original signal. For frame-based coders [3–10], information of pitch contour and gain is embedded in the framed excitation signal which can be efficiently represented by positions and amplitudes of important residual samples [3, 4], encoded by an excitation codebook [6–10], or represented by a mixed excitation model in terms of pitch period, bandpass voicing strengths and Fourier magnitudes [5]. For segment-based speech coding, the prosodic information associated with each segmental unit is usually encoded directly after quantization without considering underlying prosodic models. Methods proposed for encoding segmental pitch contour include scalar-quantization of segmental mean value [11, 13] or values at segmental end points [17], vector-quantization [25, 31], scalar-quantization after being parameterized by piecewise linear approximation (PLA) [12, 18, 19, 24, 32, 33, 36], the frame-by-frame scheme as used in the frame-based coder [14, 15, 21, 23, 27, 28], and quantization by using stored pitch contour patterns [34]. For segment duration, it is usually directly encoded/scalar-quantized [11, 12, 15, 17–19, 23, 24, 26–28], or vector-quantized [29–33]. Aside from properly encoding prosodic information for bit saving, post-modification or manipulation on the prosody of the encoded speech is also an interesting topic for a segment-based speech coder to realize some attractive and fancy functions such as changing speaking rate, changing speaker identity, and changing speaking style or emotion. Therefore, a parametric prosody coding approach basing on a prosody-generating model, which well describes the suprasegmental variations of the prosodic features, is highly desirable to have more potential than the conventional prosody coding approaches to not only save bit amount for efficiently encoding prosodic features but also be easier to realize a useful function of post-modification on the encoded prosody.

In this paper, a novel parametric prosody coding approach to efficiently encoding prosodic-acoustic features for segment-based Mandarin speech coding is proposed. It differs from the conventional prosody coding approaches using simple scalar- or vector-quantization mainly on adopting an analysis-synthesis scheme to obtain a parametric representation of the prosodic features of the input speech for encoding by an analysis operation in the encoder, and to reconstruct the prosodic features from the decoded parameters by a synthesis operation in the decoder. A hierarchical prosodic model (HPM) proposed previously [41] is employed to serve as the prosody-generating model in the analysis-synthesis scheme. The HPM is a sophisticated speech prosody model to well describe the various relations among prosodic-acoustic features, prosodic structure, and linguistic features so that it can be used to produce a compact and accurate representation of the prosodic features of the input speech for high-performance prosody coding. Besides, the HPM also provides us a platform to easily realize some post-modifications on the decoded prosody via manipulating its parameters. An example of modifying the speaking rate of the reconstructed speech via directly replacing the HPM parameters will be demonstrated in this study.

The paper is organized as follows. Section 2 presents the proposed Mandarin-speech prosody coding approach in detail. Section 3 discusses the experimental results of evaluating the proposed prosody coding approach on two continuous-speech databases. In Section 4, an application of the parametric prosody coding to speaking rate conversion is demonstrated. Some conclusions are given in the last section.

2 The proposed method

Figure 1 shows a schematic diagram of the proposed parametric prosody coding approach. In the encoder, the input utterance is firstly segmented into syllable segments interleaving with pauses by a forced aligner using the linguistic information of the associated text. The prosodic-acoustic features associated with each syllable segment are then extracted. Then, a parametric representation of the prosodic-acoustic features of a syllable segment is estimated by a prosody analysis operation based on the HPM. Lastly, the HPM parameters and some low-level linguistic features are encoded and transmitted to the decoder. In the decoder, the prosodic-acoustic features of each syllable segment are firstly reconstructed by a prosody synthesis operation which feeds the decoded low-level linguistic features and HPM parameters into the prosody-generating model, i.e., the HPM. The output speech is finally generated by an HMM-based speech synthesizer using the reconstructed prosodic-acoustic features and the decoded low-level linguistic features. Some primary parts of the proposed approach are discussed in detail in the following subsections. The HPM serving as the prosody-generating model is firstly introduced in 2.1. Then, the analysis-synthesis operations and prosody-parameter coding are described in Section 2.2. Lastly, the reconstruction of speech signal is discussed in Section 2.3.

2.1 The prosody-generating model HPM

The HPM used in this study is the statistical prosodic model proposed previously [41, 42]. Although the detail of the HPM has been included in [41], we briefly reintroduce it here to make the presentation of the proposed parametric prosody coding approach more complete and easier to understand. The HPM is a model designed to describe the various relationships of prosodic-acoustic features, prosodic structure, and linguistic features. Three types of prosodic-acoustic features are modeled in the HPM: syllable prosodic-acoustic features, syllable-juncture prosodic-acoustic features, and inter-syllable differential prosodic-acoustic features. The syllable prosodic-acoustic features include syllable pitch contour sp_n, syllable duration sd_n, and syllable energy level se_n of the n-th syllable. Here, the pitch contour of each syllable is represented by a 3-rd order orthogonal polynomial expansion [43]. The

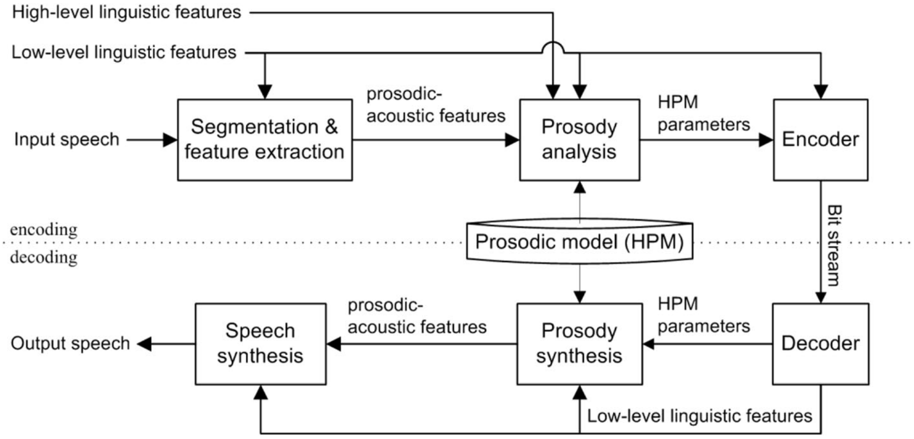

High-level linguistic features
Low-level linguistic features
Input speech
Segmentation & feature extraction
prosodic-acoustic features
Prosody analysis
HPM parameters
Encoder
Bit stream
encoding
decoding
Prosodic model (HPM)
Output speech
Speech synthesis
prosodic-acoustic features
Prosody synthesis
HPM parameters
Decoder
Low-level linguistic features

Fig. 1 A schematic diagram of the proposed method

basis polynomials used are normalized, in length, to [0,1] and can be expressed as:

$$\begin{aligned}
\phi_0\left(\frac{i}{M}\right) &= 1\\
\phi_1\left(\frac{i}{M}\right) &= \left[\frac{12\cdot M}{M+2}\right]^{1/2}\cdot\left[\frac{i}{M}-\frac{1}{2}\right]\\
\phi_2\left(\frac{i}{M}\right) &= \left[\frac{180\cdot M^3}{(M-1)(M+2)(M+3)}\right]^{1/2}\cdot\left[\left(\frac{i}{M}\right)^2-\frac{i}{M}+\frac{M-1}{6\cdot M}\right]\\
\phi_3\left(\frac{i}{M}\right) &= \left[\frac{2800\cdot M^5}{(M-1)(M-2)(M+2)(M+3)(M+4)}\right]^{1/2}\\
&\quad\cdot\left[\left(\frac{i}{M}\right)^3-\frac{3}{2}\left(\frac{i}{M}\right)^2+\frac{6M^2-3M+2}{10\cdot M^2}\left(\frac{i}{M}\right)-\frac{(M-1)(M-2)}{20\cdot M^2}\right]
\end{aligned}\tag{1}$$

for $0\le i\le M$, where $M+1$ is the length of the current syllable log-pitch contour and $M\ge 3$ in frame. They are, in fact, discrete Legendre polynomials. The pitch contour $F_n(i)$ of syllable n can then be approximated by:

$$F_n(i)\approx\sum_{j=0}^{3}\alpha_{j,n}\cdot\phi_j\left(\frac{i}{M_n}\right)\quad i=0\sim M_n,\tag{2}$$

where

$$\alpha_{j,n}=\frac{1}{M_n+1}\sum_{i=0}^{M_n}F_n(i)\cdot\phi_j\left(\frac{i}{M_n}\right)\qquad j=0\sim 3\tag{3}$$

Then, the four coefficients of syllable n form a vector $sp_n=[\alpha_{0,\ n},\alpha_{1,\ n},\alpha_{2,\ n},\alpha_{3,\ n}]^T$ to represent its pitch contour. The syllable-juncture prosodic-acoustic features include pause duration pd_n and energy-dip level ed_n of the syllable juncture between the n-th and $(n+1)$-th syllables (referred to as syllable juncture n thereafter). The inter-syllable differential prosodic-acoustic features include the normalized pitch-level jump pj_n, and the two normalized duration lengthening factors dl_n and df_n of syllable juncture n. Note that these differential features are obtained after eliminating the effects of low-level linguistic features, i.e., tone and base-syllable type. Specifically, the normalized pitch-level jump is defined by:

$$pj_n=\left(sp_{n+1}(1)-\chi_{t_{n+1}}\right)-\left(sp_n(1)-\chi_{t_n}\right)\tag{4}$$

where $sp_n(1)$ is the first dimension of syllable pitch contour sp_n (i.e., syllable pitch level); $t_n\in\{1,2,3,4,5\}$ is the tone of syllable n; and χ_t is the average pitch-level of tone t. The two normalized duration lengthening factors are defined by:

$$dl_n=(sd_n-\pi_{t_n}-\pi_{s_n})-(sd_{n-1}-\pi_{t_{n-1}}-\pi_{s_{n-1}})\tag{5}$$

$$df_n=(sd_n-\pi_{t_n}-\pi_{s_n})-(sd_{n+1}-\pi_{t_{n+1}}-\pi_{s_{n+1}})\tag{6}$$

where π_t and π_s represent respectively the average syllable durations of tone t and of base-syllable type s. So, the complete prosodic-acoustic feature sequence is $\mathbf{A}=\{\mathbf{X},\mathbf{Y},\mathbf{Z}\}=\{\mathbf{sp},\mathbf{sd},\mathbf{se},\mathbf{pd},\mathbf{ed},\mathbf{pj},\mathbf{dl},\mathbf{df}\}$, where $\mathbf{X}=\{\mathbf{sp},\mathbf{sd},\mathbf{se}\}$, $\mathbf{Y}=\{\mathbf{pd},\mathbf{ed}\}$, and $\mathbf{Z}=\{\mathbf{pj},\mathbf{dl},\mathbf{df}\}$ represent sequences of the syllable prosodic-acoustic features, the syllable-juncture prosodic-acoustic features, and the inter-syllable differential prosodic-acoustic features, respectively.

The prosodic structure considered in the HPM is a four-layer prosody hierarchy shown in Fig. 2. It is a modified version of the hierarchical prosodic phrase grouping (HPG) model proposed by Tseng [44]. It is composed of four types of layered prosodic constituents, from bottom to top, syllable (SYL), prosodic word (PW), prosodic phrase (PPh), and breathe/prosodic phrase group (BG/PG). In the HPM, the prosody hierarchy is represented in terms of two types of prosody tags $\mathbf{T}=\{\mathbf{B},\mathbf{P}\}$: the break type $\mathbf{B}$ of syllable juncture and the prosodic state $\mathbf{P}$ of syllable. The break type $\mathbf{B}$ is used to specify the boundaries of the prosodic constituents while the prosodic state $\mathbf{P}$ is used to specify the patterns of the higher-level prosodic

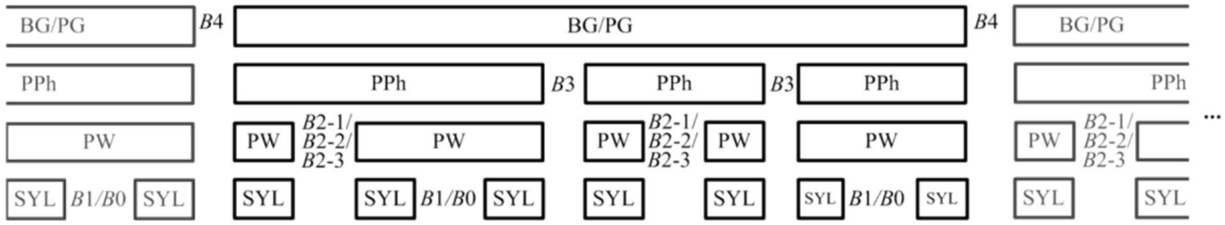


Fig. 2 The prosody hierarchical structure of Mandarin speech used in this study [42]

constituents. As shown in Fig. 2, the four prosodic constituents are delimited by seven break types denoted as *B*0, *B*1, *B*2–1, *B*2–2, *B*2–3, *B*3, and *B*4 [41, 42]. First, *B*0 and *B*1 represent respectively non-breaks of reduced syllable boundary (or tightly-coupling syllable juncture) and normal syllable boundary, within a PW, which have no identifiable pauses between SYLs. Second, PW boundary *B*2 = {*B*2–1, *B*2–2, *B*2–3} is perceived as a minor-break boundary where a slight tone of voice change usually follows. Here, *B*2–1, *B*2–2, and *B*2–3 represent PW boundaries with F0 reset, short pause, and pre-boundary syllable duration lengthening, respectively. Third, PPh boundary *B*3 is perceived as a clear pause. Fourth, *B*4 is defined for a breathing pause or a complete speech paragraph end characterized by final lengthening coupled with weakening of speech sounds. The prosodic state **P** of syllable is conceptually defined as the state in a prosodic phrase to account for the prosodic-acoustic feature variations imposed on higher-level prosodic constituents (i.e., PW, PPh, and BG/PG). In the HPM, three types of prosodic states are used, i.e., pitch prosodic state **p**, duration prosodic state **q**, and energy prosodic state **r**. So, the complete prosodic tag sequence is **T** = {**B**, **P**}, where **B** = {B_n} is a break type sequence with B_n∈{*B*0, *B*1, *B*2–1, *B*2–2, *B*2–3, *B*3, *B*4} being the break type of syllable juncture *n*, and **P** = {**p**, **q**, **r**} is the prosodic-state tag sequence with **p** = {p_n}, **q** = {q_n} and **r** = {r_n}.

The linguistic features involved in the HPM can be classified into two classes: the low-level linguistic features and the high-level linguistic features. The low-level linguistic features are those accounting for the prosodic-acoustic feature variation resulting from the prosodic constituent of the lowest level, i.e., SYL, while the high-level linguistic features account for the syllable prosodic-acoustic feature variations imposed on higher-level prosodic constituents (i.e., PW, PPh, and BG/PG) through the prosodic state. The low-level linguistic features are syllable-level features including lexical tone sequence **t**, base-syllable sequence **s**, and final type sequence **f**. The high-level linguistic features are word-level features or above. For simplicity, only the word-level linguistic features are used in the HPM. They include word length sequence **WL**, part-of-speech sequence **POS**, and punctuation mark sequence **PM**. In summary, the linguistic feature sequence used is **L** = {**t**, **s**, **f**, **WL**, **POS**, **PM**}.

To give a clearer picture of notations for the features and prosodic tags used in this study, we summarize them in Table 1.

The HPM is a model *P*(**T**, **A**| **L**) designed to describe the various relationships of prosodic-acoustic features, prosodic structure, and linguistic features. The model is formulated as

$$\begin{aligned} P(\mathbf{T},\mathbf{A}|\mathbf{L}) &= P(\mathbf{A}|\mathbf{T},\mathbf{L})P(\mathbf{T}|\mathbf{L}) = P(\mathbf{X},\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{P},\mathbf{L})P(\mathbf{B},\mathbf{P}|\mathbf{L}) \\ &\approx P(\mathbf{X}|\mathbf{B},\mathbf{P},\mathbf{L})P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L})P(\mathbf{P}|\mathbf{B})P(\mathbf{B}|\mathbf{L}) \end{aligned} \tag{7}$$

where *P*(**X**| **B**, **P**, **L**) is the syllable prosodic-acoustic model which describes the influences of the two types of prosodic tags and the contextual linguistic features on the variations of syllable F0 contour, duration, and energy level; *P*(**Y**, **Z**| **B**, **L**) is the syllable-juncture prosodic-acoustic model describing the inter-syllable acoustic characteristics specified for different break type and surrounding linguistic features; *P*(**P**| **B**) is the prosodic state model describing the variation of prosodic state conditioned on the neighboring break type; and *P*(**B**| **L**) is a break-syntax model describing the dependence of break occurrence on the surrounding linguistic features. The four models are further elaborated as follows.

The syllable prosodic-acoustic model *P*(**X**| **B**, **P**, **L**) is further divided into three sub-models by:

$$\begin{aligned} P(\mathbf{X}|\mathbf{B},\mathbf{P},\mathbf{L}) &\approx P(\mathbf{sp}|\mathbf{B},\mathbf{p},\mathbf{t})P(\mathbf{sd}|\mathbf{B},\mathbf{q},\mathbf{t},\mathbf{s})P(\mathbf{se}|\mathbf{B},\mathbf{r},\mathbf{t},\mathbf{f}) \\ &\approx \prod_{n=1}^{N} P\left(sp_n|B_{n-1}^{n},p_n,t_{n-1}^{n+1}\right)P(sd_n|q_n,s_n,t_n)P(se_n|r_n,f_n,t_n) \end{aligned} \tag{8}$$

where $P\left(sp_n|B_{n-1}^{n},p_n,t_{n-1}^{n+1}\right)$, $P(sd_n|\,q_n, s_n, t_n)$, and $P(se_n|\,r_n, f_n, t_n)$ are sub-models for the pitch contour, duration and energy level of syllable *n*, respectively; t_n, s_n and f_n denote the tone, base-syllable type and final type of syllable *n*; $B_{n-1}^{n} = (B_{n-1}, B_n)$; and $t_{n-1}^{n+1} = (t_{n-1}, t_n, t_{n+1})$. $P\left(sp_n|B_{n-1}^{n},p_n,t_{n-1}^{n+1}\right)$ is further elaborated to consider four major affecting factors. With an assumption that all affecting factors are combined additively, we have

$$sp_n = sp_n^r + \beta_{t_n} + \beta_{p_n} + \beta^f_{B_{n-1},tp_{n-1}} + \beta^b_{B_n,tp_n} + \mu_{sp} \tag{9}$$

where sp_n is the 4-dimensional vector representing the

Table 1 Notations of prosodic tags, prosodic-acoustic features and linguistic features

T: prosodic tags	**B**: break types	
	P: prosodic states	**p**: pitch prosodic states
		q: duration prosodic states
		r: energy prosodic states
A: prosodic-acoustic Features	**X**: syllable prosodic-acoustic features	**sp**: syllable pitch contours
		sd: syllable durations
		se: syllable energy levels
	Y: syllable-juncture prosodic-acoustic features	**pd**: pause durations
		ed: energy-dip levels
	Z: inter-syllable differential prosodic-acoustic features	**pj**: normalized pitch-level jumps
		dl: normalized duration lengthening factor 1
		df: normalized duration lengthening factor 2
L: linguistic features	**POS**: part-of-speeches **PM**: punctuation marks **WL**: word lengths	
	t: tones	
	s: base-syllable types	
	f: final types	

observed log-F0 contour of syllable n; sp_n^r is the modeling residue; β_{t_n} and β_{p_n} are the affecting patterns (APs) for t_n and p_n, respectively; the subscript tp_n represents the tone pair t_n^{n+1}; $\beta^f_{B_{n-1},tp_{n-1}}$ and $\beta^b_{B_n,tp_n}$ are the forward and backward coarticulation APs contributed from syllable $n - 1$ and syllable $n + 1$, respectively; and μ_{sp} is the global mean of pitch vector. In this study, β_{p_n} is set to have nonzero value only in its first dimension in order to restrict the influence of prosodic state merely on the log-F0 level of the current syllable. Besides, μ_{sp} is also assumed to have nonzero value only in its first dimension for simplicity. By assuming that sp_n^r is zero-mean and normally distributed, i.e., $N(sp_n^r; 0, R_{sp})$, we have

$$P\left(sp_n|B_{n-1}^n, p_n, t_{n-1}^{n+1}\right) = N\left(sp_n; \beta_{t_n} + \beta_{p_n} + \beta^f_{B_{n-1},tp_{n-1}} + \beta^b_{B_n,tp_n} + \mu_{sp}, R_{sp}\right) \tag{10}$$

It is noted that sp_n^r is a noise-like residual signal of very small deviation so that we model it by a normal distribution.

Similar to the design of the syllable pitch contour model, the syllable duration model $P(sd_n|\, q_n, s_n, t_n)$ and the syllable energy level model $P(se_n|\, r_n, f_n, t_n)$ are formulated by

$$P(sd_n|q_n, s_n, t_n) = N\left(sd_n; \gamma_{t_n} + \gamma_{s_n} + \gamma_{q_n} + \mu_{sd}, R_{sd}\right) \tag{11}$$

$$P(se_n|r_n, f_n, t_n) = N\left(se_n; \omega_{t_n} + \omega_{f_n} + \omega_{r_n} + \mu_{se}, R_{se}\right) \tag{12}$$

where sd_n and se_n are the observed duration and energy level of syllable n, respectively; γ's and ω's represent APs for syllable duration and syllable energy level; μ_{sd} and μ_{se} are their global means; and R_{sd} and R_{se} are variances of modeling residues.

The syllable-juncture prosodic-acoustic model, $P(\mathbf{Y}, \mathbf{Z}|\, \mathbf{B}, \mathbf{L})$, is further divided into five sub-models by

$$\begin{aligned} P(\mathbf{Y}, \mathbf{Z}|\mathbf{B}, \mathbf{L}) &\approx P(\mathbf{pd}, \mathbf{ed}, \mathbf{pj}, \mathbf{dl}, \mathbf{df}|\mathbf{B}, \mathbf{L}) \\ &\approx \prod_{n=1}^{N-1} P\left(pd_n, ed_n, pj_n, dl_n, df_n|\mathbf{B}, \mathbf{L}\right) \\ &\approx \prod_{n=1}^{N-1} \Big\{ g\left(pd_n; \alpha_{B_n,L_n}, \eta_{B_n,L_n}\right) N\left(ed_n; \mu_{ed,B_n,L_n}, \sigma^2_{ed,B_n,L_n}\right) \end{aligned} \tag{13}$$

where $g(pd_n; \alpha_{B_n,L_n}, \eta_{B_n,L_n})$ is a Gamma distribution for pause duration pd_n of syllable juncture n; and the other four features, ed_n, pj_n, dl_n, and df_n, are all modeled as normal distributions. Since the space of L_n is large, the CART algorithm [45] with the node splitting criterion of maximum likelihood (ML) gain is adopted to concurrently classify the five features of pd_n, ed_n, pj_n, dl_n, and df_n for each break type according to a question set. The question set consists of 216 questions considering the following linguistic features around the current juncture: (1) the initial type of the following syllable; (2) interword/intraword indicator; (3) lengths; and (4) POSs of the words before and after the juncture if it is an interword; and (5) PM type for an interword juncture. Each leaf node represents the product of the five sub-models. So, seven decision trees are constructed for the syllable-juncture prosodic-acoustic model.

The prosodic state model $P(\mathbf{P}|\,\mathbf{B})$ is further divided into three sub-models:

$$\begin{aligned} P(\mathbf{P}|\mathbf{B}) &\approx P(\mathbf{p}|\mathbf{B})P(\mathbf{q}|\mathbf{B})P(\mathbf{r}|\mathbf{B}) \\ &\approx P(p_1)P(q_1)P(r_1) \\ &\left[\prod_{n=2}^{N} P(p_n|p_{n-1},B_{n-1})P(q_n|q_{n-1},B_{n-1})P(r_n|r_{n-1},B_{n-1})\right] \end{aligned} \tag{14}$$

where $P(p_n|\,p_{n-1}, B_{n-1})$,$P(q_n|\,q_{n-1}, B_{n-1})$, and $P(r_n|\,r_{n-1}, B_{n-1})$ are prosodic state transition models for syllable pitch level, duration and energy level, respectively. Notice that, in above formulation, the dependency on the break type of the preceding syllable juncture makes these models be able to properly model significant pitch/energy resets and the pre-boundary lengthening effect across major breaks. We also note that the three prosodic states are independently modeled for simplicity.

Lastly, the break-syntax model $P(\mathbf{B}|\,\mathbf{L})$ is approximated by

$$P(\mathbf{B}|\mathbf{L}) \approx \prod_{n=1}^{N-1} P(B_n|L_n) \tag{15}$$

where $P(B_n|\,L_n)$ is the break type model for juncture n, and L_n is the contextual linguistic features surrounding juncture n. Since the space of linguistic features L_n is large, we partition it into several classes $C(L_n)$ by the CART decision tree algorithm [45] using the maximum likelihood gain criterion and the same question set used in the training of the syllable-juncture prosodic-acoustic model.

The HPM can be trained automatically from a prosody-unlabeled speech database by a joint prosody labeling and modeling (PLM) algorithm [41]. The PLM algorithm is a sequential optimization procedure based on the ML criterion to jointly label the prosodic tags for all utterances of the training corpus and estimate the parameters of all 12 prosodic sub-models.

2.2 The parametric prosody coding approach

The proposed parametric prosody coding approach considers the coding of four prosodic-acoustic features including syllable pitch contour, syllable duration, syllable energy level, and syllable-juncture pause duration. It takes four sub-models of the HPM as the generating models for these four prosodic features. The four sub-models are the syllable pitch contour sub-model $P(sp_n|B_{n-1}^{n}, p_n, t_{n-1}^{n+1})$, the syllable duration sub-model $P(sd_n|\,q_n, s_n, t_n)$, the syllable energy level sub-model $P(se_n|\,r_n, f_n, t_n)$, and the syllable-juncture pause duration sub-model $g(pd_n; \alpha_{B_n,L_n}, \eta_{B_n,L_n})$. The first three sub-models are controlled directly by low-level linguistic features and prosodic tags through their APs, while the last one is controlled implicitly by high-level linguistic features and prosodic tags through the break type-dependent decision trees. The low-level and high-level linguistic features can be simply obtained from a linguistic processor while the prosodic tags, p_n, q_n, r_n, and B_n, are obtained by the prosody analysis operation to be discussed below. We discuss the prosody analysis operation and the parameter coding for these four prosodic-acoustic features in detail as follows.

2.2.1 Prosody analysis operation

The task of the prosody analysis operation is to find the best prosodic state and break type sequences for the encoding utterance with given prosodic-acoustic features and linguistic features. Based on the HPM, the task is formulated by

$$\mathbf{T}^* = \{\mathbf{B}^*, \mathbf{P}^*\} = \arg\max_{\mathbf{B},\mathbf{P}} Q \tag{16}$$

where

$$\begin{aligned} Q &= P(\mathbf{B}|\mathbf{L})P(\mathbf{P}|\mathbf{B})P(\mathbf{X}|\mathbf{B},\mathbf{P},\mathbf{L})P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L}) \\ &= \left(\prod_{n=1}^{N-1} P(B_n|L_n)\right) \\ &\quad \left(P(p_1)P(q_1)P(r_1)\left[\prod_{n=2}^{N} P(p_n|p_{n-1},B_{n-1})P(q_n|q_{n-1},B_{n-1})P(r_n|r_{n-1},B_{n-1})\right]\right) \\ &\quad \left(\prod_{n=1}^{N} P\left(sp_n|B_{n-1}^{n},p_n,t_{n-1}^{n+1}\right)P(sd_n|q_n,s_n,t_n)P(se_n|r_n,f_n,t_n)\right) \\ &\quad \Bigg(\prod_{n=1}^{N-1} g\left(pd_n;\alpha_{B_n,L_n},\eta_{B_n,L_n}\right) \\ &\quad \cdot N\left(ed_n;\mu_{ed,B_n,L_n},\sigma^2_{ed,B_n,L_n}\right)N\left(pj_n;\mu_{pj,B_n,L_n},\sigma^2_{pj,B_n,L_n}\right) \\ &\quad N\left(dl_n;\mu_{dl,B_n,L_n},\sigma^2_{dl,B_n,L_n}\right)N\left(df_n;\mu_{df,B_n,L_n},\sigma^2_{df,B_n,L_n}\right)\Bigg) \end{aligned} \tag{17}$$

The task is realized by the following iterative procedure:

1) *Initialization*
For $i = 0$, find the initial break type sequence by

$$\mathbf{B}^i = \arg\max_{\mathbf{B}} P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L})P(\mathbf{B}|\mathbf{L}) \tag{18}$$

2) *Iteration*
Starting from $i = 1$, estimate the prosodic state sequence and the break type sequence iteratively by the following three steps:

Step 1: Given $\mathbf{B}^{i-1}$, re-label the prosodic state sequence of each utterance by the Viterbi algorithm so as to maximize Q defined in (17), i.e.,

$$\mathbf{P}^i = \arg\max_{\mathbf{P}} P\left(\mathbf{X}|\mathbf{B}^{i-1},\mathbf{P},\mathbf{L}\right)P\left(\mathbf{Y},\mathbf{Z}|\mathbf{B}^{i-1},\mathbf{L}\right)P\left(\mathbf{P}|\mathbf{B}^{i-1}\right)P\left(\mathbf{B}^{i-1}|\mathbf{L}\right) \tag{19}$$

Step 2: Given with $\mathbf{P}^i$, re-label the break type sequence of each utterance by the Viterbi algorithm so as to maximize Q, i.e.,

$$\mathbf{B}^i = \arg\max_{\mathbf{B}} P\left(\mathbf{X}|\mathbf{B},\mathbf{P}^i,\mathbf{L}\right)P\left(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L}\right)P\left(\mathbf{P}^i|\mathbf{B}\right)P\left(\mathbf{B}|\mathbf{L}\right) \tag{20}$$

Step 3: If a convergence of the value Q is reached, exit the iteration; otherwise, increase i by 1 and go to step 1.

3) *Termination*

$$\mathbf{B}^* = \mathbf{B}^i \text{ and } \mathbf{P}^* = \mathbf{P}^i \tag{21}$$

2.2.2 Coding of prosody parameters

In the HPM, the syllable pitch contour sp_n, the syllable duration sd_n, and the syllable energy level se_n are linearly modeled in Eqs. (10), (11), and (12) to consider several major affecting factors that influence their variations. The affecting factors involved in these three sub-models are some low-level linguistic features and prosodic tags, including tone t_n, base-syllable type s_n, final type f_n, break-type tag B_n, and the three prosodic-sate tags of p_n, q_n, and r_n. These affecting factors are the only parameters required to represent the three syllable prosodic-acoustic features by the HPM. So, in the encoder, we only need to consider the encoding of these seven affecting factors. In the decoder, the decoded versions of these seven affecting factors can be used to reconstruct the three syllable prosodic-acoustic features. Specifically, the syllable pitch contour, the syllable duration, and the syllable energy level are simply reconstructed by superimposing the APs associated with these affecting factors, i.e.,

$$sp'_n = \beta_{t_n} + \beta_{p_n} + \beta^f_{B_{n-1},tp_{n-1}} + \beta^b_{B_n,tp_n} + \mu_{sp} \tag{22}$$

$$sd'_n = \gamma_{t_n} + \gamma_{s_n} + \gamma_{q_n} + \mu_{sd} \tag{23}$$

$$se'_n = \omega_{t_n} + \omega_{f_n} + \omega_{r_n} + \mu_{se} \tag{24}$$

We note that the three means, μ_{sp}, μ_{sd}, and μ_{se}, are sent in advance to the decoder as side information. We also note that the three modeling residuals, sp^r_n, sd^r_n, and se^r_n, are neglected in the above three equations because their variances are all small.

In the HPM, the pause duration is modeled by the syllable-juncture pause duration sub-model, $g\left(pd_n; \alpha_{B_n,L_n}, \eta_{B_n,L_n}\right)$. The sub-model describes the variation of syllable-juncture pause duration influenced by some contextual linguistic features and break type and is organized into 7 break type-dependent decision trees (BDTs). For each break type, a decision tree is used to determine the probability density function (*pdf*) of syllable-juncture pause duration according to the contextual linguistic features. Here, all *pdf*s are assumed to be Gamma distributed. By an analysis on these 7 decision trees, it is found that the means of *pdfs* in the leaf nodes for the break types with very short pause duration (< 0.03 s), i.e., *B*0, *B*1, *B*2–1, and *B*2–3, are very close to the *pdfs* of the root nodes. For the break types with pause durations, i.e., *B*2–2, *B*3, and *B*4, the *pdfs* of leaf nodes have more sophisticated pause duration distributions. However, by an informal listening test, we found that the synthesized speeches with pause durations encoded by the *pdfs* of the root nodes sound almost the same as the ones encoded by the *pdfs* of the leaf nodes. The pause duration information, therefore, can be encoded solely by the *pdfs* of the root nodes. In other words, only the symbols of the break types need to be encoded and sent to the decoder. The means of the *pdfs* for the 7 break types are sent to the decoder as the side information. The decoder reconstructs the syllable-juncture pause duration as the means of the *pdfs* of the root nodes of the 7 break types.

In summary, the symbols needed to be encoded for each syllable segment and its following pause duration in the proposed parametric prosody coding approach include tone, base-syllable type, prosodic-state tag, and break-type tag. Table 2 lists the bit assignments for these symbols based on two experimental settings conducted in this study. The two experiments are prosody coding for a speaker-independent (SI) case and a speaker-dependent (SD) case. The numbers of prosodic states for p_n, q_n, and r_n are all empirically set to be 16, while the number of break types is 7 determined based on the hierarchical prosody structure used in designing the HPM. There are five lexical tones and 411 base-syllable types in Mandarin Chinese. As shown in Table 2, the total number of bits per syllable is 27 for the SI and SD cases.

Table 2 Bit assignment for symbols used in the parametric prosody coding

Symbol	No. of symbol	No. of bit
Lexical tone t_n	5	3
Base-syllable type s_n	411	9
Pitch prosodic state p_n	16	4
Duration prosodic state q_n	16	4
Energy prosodic state r_n	16	4
Break type B_n	7	3
Total bit number per syllable		27

Aside from the above regular bitstream, some HPM parameters are also needed in the decoder to help to reconstruct the prosodic-acoustic features. They are sent to the decoder in advance as side information. They include the affecting patterns (APs) $\{\beta_t, \beta_p, \beta^f_{B,tp}, \beta^b_{B,tp}, \mu_{sp}\}$ of the syllable pitch-contour sub-model, the APs $\{\gamma_t, \gamma_s, \gamma_q, \mu_{sd}\}$ of the syllable duration sub-model, the APs $\{\omega_t, \omega_f, \omega_r, \mu_{se}\}$ of the syllable energy level sub-model, and the means $\{\mu^{pd}_T\}$ of the root-node *pdf*s of the syllable-juncture pause duration sub-model. It is noted that the subscript n which represents syllable index is eliminated from the APs listed above, i.e., β's, γ's, ω's, to simplify the representation of the types of the APs associated with the affecting factors. Specifically, the tone of syllable n, i.e., t_n, could be one of the set of tone $t \in \{1, 2, 3, 4, 5\}$ and each of the APs $\{\beta_t, \gamma_t, \omega_t\}$, therefore, has five patterns. Similarly, the base syllable type of syllable n, i.e., s_n, could be one of the set of base syllable type $s \in \{411 \text{ base syllable types}\}$, and hence, γ_s has 411 patterns. Therefore, the total number of parameters in the syllable pitch-contour sub-model is 1477, including 20 (=5 × 4) for β_t, 16 (=16 × 1) for β_p, 720 (=(7 × 5 × 5 + 5)× 4) for $\beta^f_{B,tp}$, 720 (=(7 × 5 × 5 + 5)× 4) for $\beta^b_{B,tp}$, and one for μ_{sp}; 433 parameters for the syllable duration sub-model, including 5 for γ_t, 16 for γ_q, 411 for γ_s, and 1 for μ_{sd}; 62 parameters for the syllable energy level sub-model, including 5 for ω_t, 40 for ω_f, 16 for ω_r, and 1 for μ_s; and the means of the pdfs for the syllable-juncture pause duration corresponding to 7 break types, i.e., 7 BDT root node means μ^{pd}_T . Table 3 summarizes the side information of the coding system.

2.3 Speech synthesis

In this study, an HMM-based speech synthesizer [46–49] is used to generate the synthetic voice. The standard context-dependent HMM training for speech synthesis [46–48] is adopted here to simultaneously construct spectral, voiced/unvoiced and state duration models using the labels containing the information of contextual influential factors.

Five-state left-to-right HMMs are used to model synthesis units of 21 syllable *Initials* and 39 syllable *Finals*. The observations in each HMM state consist of two streams. One is a 75-dimensional spectral feature vector composed of 24-dimensional mel-generalized cepstral coefficients (MGC) [50], delta MGCs, delta-delta MGCs, energy, delta energy, and delta-delta energy. Another is a discrete symbol to indicate the voiced/unvoiced status of a frame. The spectral features in each HMM state are modeled by a multi-variate single Gaussian, while voiced/unvoiced symbols are modeled by a discrete probability distribution of the two events. The state durations of each HMM form a 5-dimentional vector which is modeled by a multi-variate single Gaussian. Since spectrum, voiced/unvoiced status and state duration have their own contextual influential factors, the distributions for MGC, voiced/unvoiced indicator and state duration are clustered independently. The question set used for the decision tree-based context clustering of HMMs is formed by using the context labels. To achieve a better tree clustering result, we merge some individual linguistic features to form several complex questions according to their effects on producing spectrum, state duration and voiced/unvoiced status. For examples, the initial/final types are classified by the manner or place of articulation; the prosodic states of duration are tied according to their values of APs; the break types are merged into broader classes according to their corresponding prosodic-acoustic features, etc. There are in total 399 questions formed for the decision tree-based context clustering of HMMs in this study. It is noted that these context-dependent HMMs (CD-HMMs) are embedded-trained [48] with feature vectors consisting of the MGCs along with the voiced/unvoiced indicators so as to avoid the discrepancy between spectrum and voiced/unvoiced status in each HMM state. In this study, the training data for the SD case is large enough to conduct speaker-dependent HMM training. The trained HMM can be directly used to generate synthesized speeches given the encoded symbols illustrated in Section 2.2. On the other hand, for the SI case, the speakers in the training set do not overlap the speakers in the test set and the number of the utterances for each speaker of the test set is very small. The HMM model of each speaker in the test set is therefore adapted from the HMM of the SD case by the CMAPLR approach [51] given solely the test utterances.

Figure 3 shows the schematic diagram of the HMM-based speech synthesizer used in this study. To synthesize speech by the HMM-based synthesizer, we first generate the state durations for each syllable segment. The state duration is assumed to be normally distributed and affected by the contextual information of *Initial*, *Final*, and prosodic tags, i.e.,

$$P\left(d_{n,c}|I\left(s^{n+1}_{n-1}\right), F\left(s^{n+1}_{n-1}\right), p_n, q_n, r_n, B^n_{n-1}\right) = N\left(d_{n,c}; \mu_{n,c}, \sigma^2_{n,c}\right) \quad \text{for } c = 1 \sim C \tag{25}$$

where $d_{n,\ c}$ denotes the duration of the c-th state of

Table 3 Side information of the proposed coding system

Type	Parameter no.
Lexical tone APs:$\beta_t/\gamma_t/\omega_t$	20/5/5
Coarticulation APs: $\beta^f_{B,tp}/\beta^b_{B,tp}$	720/720
Prosodic state APs: $\beta_p/\gamma_q/\omega_r$	16/16/16
Global mean APs: $\mu_{sp}/\mu_{sd}/\mu_{se}$	1/1/1
Base-syllable type and final type APs: γ_s/ω_f	411/40
BDT root node mean: μ^{pd}_T	7
Total	1979

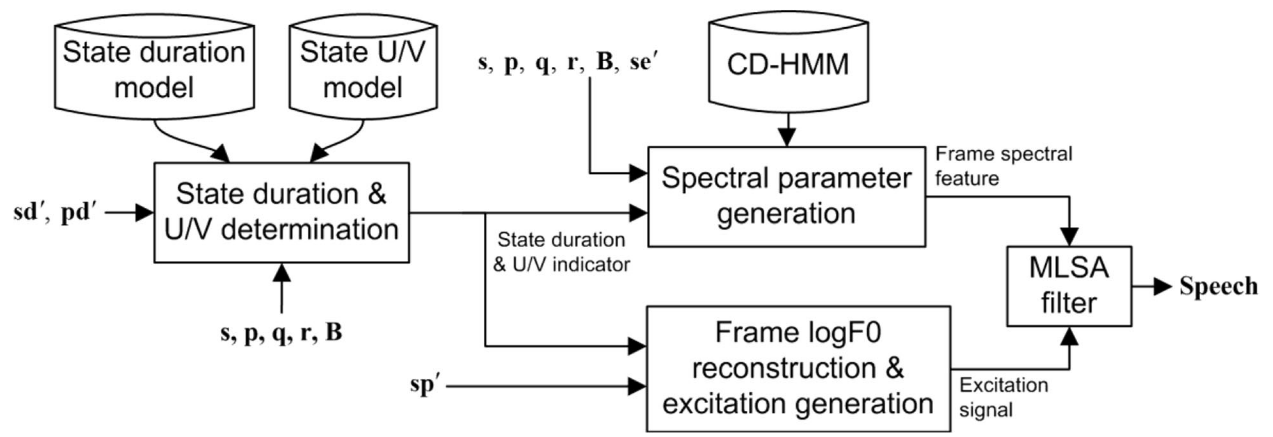

State duration model
State U/V model
s, p, q, r, B, se′
CD-HMM
sd′, pd′
State duration & U/V determination
Spectral parameter generation
Frame spectral feature
State duration & U/V indicator
MLSA filter
Speech
s, p, q, r, B
Frame logF0 reconstruction & excitation generation
sp′
Excitation signal

Fig. 3 A schematic diagram of the HMM-based speech synthesizer used in this study

syllable n; C is the total number of state; $I(x)$ and $F(x)$ denote respectively the *Initials* and *Finals* of the base-syllable sequence x; $\mu_{n,\ c}$ and $\sigma_{n.\ c}$ are respectively the mean and standard deviation of the state duration model obtained by finding the leaf node on the decision tree using the contextual information, i.e., $I(s_{n-1}^{n+1})$, $F(s_{n-1}^{n+1}), p_n$, q_n, r_n, and B_{n-1}^{n}. Given the reconstructed syllable duration sd'_n, the state durations of the syllable segment can be estimated based on maximizing the summed log likelihood [48], i.e.,

$$d^*_{n,1}\cdots d^*_{n,C} = \arg\max_{d_{n,1}\cdots d_{n,C}} \sum_{c=1}^{C} \log N\left(d_{n,c};\mu_{n,c},\sigma^2_{n,c}\right) \tag{26}$$

under the constraint

$$sd'_n = \sum_{c=1}^{C} d_{n,c} \tag{27}$$

The resulting state durations are expressed by

$$d_{n,c} = \mu_{n,c} + \rho\cdot\sigma^2_{n,c} \quad \text{for } c = 1\sim C \tag{28}$$

where

$$\rho = \left(sd'_n - \sum_{c=1}^{C}\mu_{n,c}\right) \Big/ \left(\sum_{c=1}^{C}\sigma^2_{n,c}\right) \tag{29}$$

An HMM state is set to be voiced if the probability of being the voiced status is larger than that of being the unvoiced status. Therefore, the length and place of syllable pitch contour can be simply determined using the information of the estimated state durations and the voiced/unvoiced indicators of the HMM states. Using the reconstructed syllable pitch contour parameter $sp'_n = [\alpha'_{0,n}, \alpha'_{1,n}, \alpha'_{2,n}, \alpha'_{3,n}]$, we can reconstruct the pitch contour of syllable n by orthogonal expansion [43], i.e.,

$$F'_n(i) = \sum_{j=0}^{3} \alpha'_{j,n}\cdot\phi_j\left(\frac{i}{M'_n}\right) \quad \text{for } i = 0\sim M'_n \tag{30}$$

where $M'_n + 1$ is the estimated length of the pitch contour of syllable n. Then, the excitation signal can be generated using the reconstructed syllable pitch contours. On the other hands, the frame spectral feature (i.e., MGC) vector sequence is generated by an HMM parameter generation algorithm [52] given with the CD-HMMs, the estimated state durations, and the contextual information (i.e., $I(s_{n-1}^{n+1})$, $F(s_{n-1}^{n+1}), p_n, q_n, r_n$, and B_{n-1}^{n}). It is noted that the energy level of each syllable CD-HMM (i.e., an *Initial* CD-HMM connecting with a *Final* CD-HMM) is scaled to se'_n before executing the parameter generation algorithm so as to make the generated energy contour smooth and approximate the desired syllable energy levels. Lastly, the output speech is synthesized directly from the generated MGC coefficients and the excitation signal by the MLSA filter [53].

3 Experimental results

3.1 Database & Experiment Setting

The proposed parametric prosody coding approach was evaluated on two large Mandarin read speech databases, the Treebank speech corpus and the TCC300 [54]. The Treebank speech corpus was designed for constructing a TTS and consisted of 420 utterances with 55,766 syllables uttered by a female professional announcer in a quiet room. Its associated texts were all short paragraphs composed of several sentences selected from the Sinica Treebank Version 3.0 [55] text corpus. The TCC300 database was collected for Mandarin automatic speech recognition (ASR). It consisted of two sets: 103-speaker short sentential utterances (set A) designed for considering phonetic balance and 200-speaker paragraphic utterances (set B) designed for the use of prosody in ASR study. The texts of set B were selected from the Academia Sinica Balanced Corpus of Modern Chinese (ASBC) [56]. Each database was further divided into training and testing subsets. Table 4 displays the usages of the subsets for each speech corpus and their statistics.

The Treebank speech corpus and the TCC300 were digitally recorded in forms of 20 kHz sampling rate/16-bit resolution and 16 kHz sampling rate/16-bit resolution, respectively. The associated texts were automatically word-segmented and POS-tagged and then manually checked. The tone and base-syllable type of each syllable were transcribed by a linguistic processor with a 130,000-word lexicon and then manually error-corrected. Syllable segmentations of the two test sets, TestTB and TestTC, were accomplished via performing forced alignment by the Hidden Markov Model Toolkit (HTK) [57] using respectively a speaker-dependent (SD) acoustic model (AM) trained from TrainTB and a speaker-independent (SI) AM trained from TrainTC1. F0 detection was firstly done by WaveSurfer [58], then error corrected automatically by the method proposed in [59], and lastly corrected manually.

3.2 Training of the HPMs

Two HPMs were trained for the SD and SI prosody coding tasks from the subsets of TrainTB and TrainTC2, respectively. For the HPM training, the eight prosodic-acoustic features, including syllable pitch contour vector, syllable duration, syllable energy level, syllable-juncture pause duration and energy-dip level, inter-syllable normalized pitch-level jump, and two inter-syllable normalized duration lengthening factors, were extracted after obtaining the time-alignment information of syllable segments. It is noted that to compensate the speaker variability in the case of the SI prosody coding, syllable pitch contour vectors were extracted from the frame-based F0 values normalized by speaker-level mean and variance, while both syllable duration and syllable energy level were normalized by their corresponding speaker-level means and variances. In the case of the SD prosody coding, only syllable duration and syllable energy level were normalized by their corresponding utterance-level means and variances to compensate utterances' variability. The associated texts were processed by the linguistic processor mentioned previously to extract all linguistic features needed in the HPM training. The PLM algorithm [41] was then applied to automatically generate two sets of 12 prosodic sub-models from the two training subsets of TrainTB and TrainTC2, respectively. In realizing the PLM algorithm, the numbers of pitch, duration, and energy prosodic states were all set to be 16. For avoiding over-fitting the decision trees of the break-syntax model and the syllable-juncture prosodic-acoustic model, the following two stop criteria were empirically set: (1) The size of a leaf node must be larger than 700/250 syllables for the SI/SD case, and (2) the relative improvement of likelihood must be larger than 0.0065 in a node splitting for both SI and SD cases. Table 5 shows the total numbers of nodes and leaf nodes for the break-syntax model and the syllable-juncture prosodic-acoustic model in the SI and SD HPMs. As shown in the table, the SD HPM uses much larger decision trees for the two models. This mainly results from the better quality of prosody pronunciation for the Treebank speech corpus uttered by a professional announcer.

3.3 Performance evaluations

The performance of the proposed approach is evaluated by the objective and subjective measures. The objective measures are the root-mean-square errors (RMSEs) of the four reconstructed prosodic-acoustic features and bit rates. The subjective tests are MUSHRA-style [60] listening tests which ask listeners to rate synthesized speeches from different systems on a scale from 0 to 100 using the sliders on the screen. To give a meaningful reference of the performance, a baseline prosody coding system (without using knowledge of prosodic characteristics and prosodic structures for Mandarin Chinese) that uses generic approaches, i.e. vector or scalar quantization techniques, is constructed for comparison with the proposed approach. The syllable pitch contour sp_n which is represented in a four-dimensional vector is vector-quantized by the k-Means clustering algorithm with the squared Euclidean distance metric. The syllable

Table 4 The usages of the subsets for each speech corpus and their statistics

Corpus	Subsets	Usages	Spk#	Utt#	Syl#	Hours	Remark
Treebank	TrainTB	Training of the HPM, the AM for forced-alignment and the models for HMM-based speech synthesizer	1	376	51,868	3.9	
	TestTB	Evaluation of prosody coding	1	44	3898	0.3	
TCC300	TrainTC1	Training of the AM for forced-alignment	274	8036	300,728	23.9	Include all set A and 90% of set B
	TrainTC2	Training of the HPM	164	962	106,955	8.3	Subset of TrainTC1
	TestTC	Evaluation of prosody coding and adaptation of HMM model for speech sythesis	19	226	26,357	2.4	Selected from Set B of TCC300

Table 5 Numbers of nodes (leaf nodes) for the break-syntax model and the syllable-juncture prosodic-acoustic models in SI and SD HPMs

Task	Subsets	Break-syntax model	Syllable-juncture prosodic-acoustic model
SD	TrainTB	139(70)	91(49)
SI	TrainTC2	63(31)	43(25)

duration sd_n, the syllable energy level se_n, and the syllable-juncture pause duration pd_n are independently scalar-quantized by the *k*-Means clustering algorithm with the squared Euclidean distance metric. Therefore, each of the prosodic-acoustic features, i.e. sp_n, sd_n, se_n, and pd_n, has its own codebook and corresponding codewords. The speech synthesizer for the baseline system is also implemented by the HMM-based speech synthesizer. Initial or final is taken as modeling the HMM unit with 5 state (the same as the proposed approach). Each modeling HMM unit is described by the features of initial/final type and the codewords for sp_n, sd_n, se_n, and pd_n. The question set used for the decision tree-based context clustering of HMMs is formed by contextual initial or final type and prosodic properties of the codewords for sp_n, sd_n, se_n, and pd_n.

3.3.1 Comparison between the equal-RMSE baselines and the proposed approach

Because the coding of sp_n, sd_n, se_n, and pd_n by the proposed coding approach share the symbols of tone, base-syllable type and break types, rate-distortion comparison between the baseline and the proposed prosody coding approach for each of the prosodic-acoustic feature is not feasible. The root-mean-square errors (RMSEs) of the four reconstructed prosodic-acoustic features by the proposed approach are hence taken as references for comparison. Table 6 shows the RMSEs of the four reconstructed prosodic-acoustic features for the inside and outside datum. For the SD task (SD-HPM), the RMSE was 0.070/0.064 logHz (inside/outside) in F0 coding, 4.8/4.7 ms in syllable duration coding, 0.68/0.70 dB in syllable energy-level coding, and 41.4/34.3 ms in syllable-juncture pause duration coding. The corresponding performances were 0.065/0.056 logHz, 9.3/7.5 ms, 0.80/0.66 dB, and 44.8/44.9 ms for the SI task (SI-HPM). Except the RMSEs of the pause-duration coding, all other values are quite low. Table 7 shows the RMSEs of the reconstructed pause duration for different break types. It can be seen from the table that they were high only for *B*2–2, *B*3 and *B*4. Since these three break types are minor and major breaks and are more tolerant to large coding errors, the performance was reasonably good.

Figures 4 and 5 show the numbers of the codewords versus RMSEs of the four prosodic-acoustic features respectively for the SD and SI prosody codings of the training sets. The baseline prosody coder yields the similar RMSEs to the proposed coder at around 24, 19, 16, and 3 codewords for sp_n, sd_n, se_n, and pd_n, respectively,

Table 6 RMSE of syllable logF0 contour (sp_n), syllable duration (sd_n), syllable energy level (se_n), and syllable-juncture pause duration (pd_n) for the proposed approach and the baseline systems with various logF0 codebook

(a) SD case

	Inside (TrainTB)				Outside (TestTB)			
	sp_n(logHz)	sd_n(ms)	se_n(dB)	pd_n(ms)	sp_n(logHz)	sd_n(ms)	se_n(dB)	pd_n(ms)
SD-HPM	.070	4.8	.68	41.4	.064	4.7	.70	34.3
SD-BSL-24	.069	5.0	.64	33.5	.066	4.7	.59	32.9
SD-BSL-32	.065	5.0	.64	33.5	.061	4.7	.59	32.9
SD-BSL-64	.053	5.0	.64	33.5	.050	4.7	.59	32.9
SD-BSL-128	.044	5.0	.64	33.5	.042	4.7	.59	32.9
SD-BSL-256	.037	5.0	.64	33.5	.042	4.7	.59	32.9

(b) SI case

	Inside (TrainTC2)				Outside (TestTC)			
	sp_n(logHz)	sd_n(ms)	se_n(dB)	pd_n(ms)	sp_n(logHz)	sd_n(ms)	se_n(dB)	pd_n(ms)
SI-HPM	.065	9.3	.80	44.8	.056	7.5	.66	44.9
SI-BSL-10	.063	9.1	.78	42.0	.060	10.9	.88	39.4
SI-BSL-16	.056	9.1	.78	42.0	.054	10.9	.88	39.4
SI-BSL-32	.047	9.1	.78	42.0	.046	10.9	.88	39.4
SI-BSL-64	.040	9.1	.78	42.0	.039	10.9	.88	39.4
SI-BSL-128	.034	9.1	.78	42.0	.033	10.9	.88	39.4
SI-BSL-256	.029	9.1	.78	42.0	.029	10.9	.88	39.4

Table 7 The RMSE (ms) performance of the reconstructed pause duration with respect to different break types

		B0	B1	B2–1	B2–2	B2–3	B3	B4
Treebank	Inside (TrainTB)	1.2	13.7	20.7	54.1	21.5	68.1	127.7
	Outside (TestTB)	1.5	14.9	19.1	41.2	21.4	60.0	103.4
TCC300	Inside(TrainTC2)	0.9	11.1	10.9	24.6	9.1	112.7	199.8
	Outside (TestTC)	1.2	22.4	20.7	39.5	7.5	164.5	249.8

in the SD case, and at around 10, 11, 10, and 3 codewords for sp_n, sd_n, se_n, and pd_n, respectively, in the SI case. Besides of the encoding of the four prosodic-acoustic features, the base-syllable type information must be encoded for speech synthesis. Table 8 lists the bit assignments of each type of the codewords for the SD and SI cases. The total number of bits per syllable, as a result, are 25 and 23 for the SD and SI cases. This result shows that the baseline approach can achieve the same RMSE performance as the proposed approach with fewer bits.

It should be noted that a lower or equal RMSE does not necessarily always indicate a higher subjective quality. We therefore conduct two independent MUSHRA-style listening tests for the SD and SI cases. Each of the MUSHRA-style listening test (SD or SI case) present a sequence of pages to listeners. Listeners were asked to rate qualities of speech prosody produced from the following four systems or methods: the proposed approach (HPM), the baseline system (BSL), vocoded natural speech (NAT), and the proposed approach with correct prosodic-acoustic features (CPRO). The vocoded natural speeches were made by MLSA filter [53] with 25-dimensional MGC parameters [50] and are provided to listeners to serve as a reference (100 point) for their assessment. The speeches synthesized by the proposed approach with correct prosodic-acoustic feature are taken as oracle performance of the proposed prosody coding which is bounded by the speech quality of the HMM-based speech synthesizer.

For the SD case, we randomly select 10 utterances from the test set (TestTB) for the MUSHRA test. For the SI case, we randomly select three female and three male speakers in the test set (TestTC). Two utterances from each of the selected speakers were then randomly chosen as utterance samples for the MUSHRA test. Therefore, there are 10 and 12 (2 genders * 3 speakers * 2 utterances) pages for the SD and SI cases and each of the pages has four parallel synthesized speech instances generated for each of the four systems or methods. We recruited 15 native Mandarin Chinese speakers for the SD and SI MUSHRA tests.

Table 9 shows the means and standard deviations of the scores rated in the two MUSHRA tests for the SD and SI cases. It can be seen from the table that the proposed approach (HPM) gets the higher scores than the baseline (BSL) gets for the SD and SI cases. The HPM in the SD case gets very close score to the CPRO, indicating the proposed approach can almost reach the performance of the oracle performance (correct prosodic-acoustic feature) bounded by the speech quality of the HMM-based speech synthesizer. Interestingly, we found that the HPM in the SI case even gets slightly higher scores than the CPRO. This unexpected result is due to the poor F0 extraction for the SI case, and this poor F0 extraction is less harmful for the F0 reconstruction by the HPM parameter with the information of tone and the prosodic structure. The scores for NAT are not 100 are due to the glitch caused by human error. Figure 6 shows the boxplots for the scores in the SD and SI cases. On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually

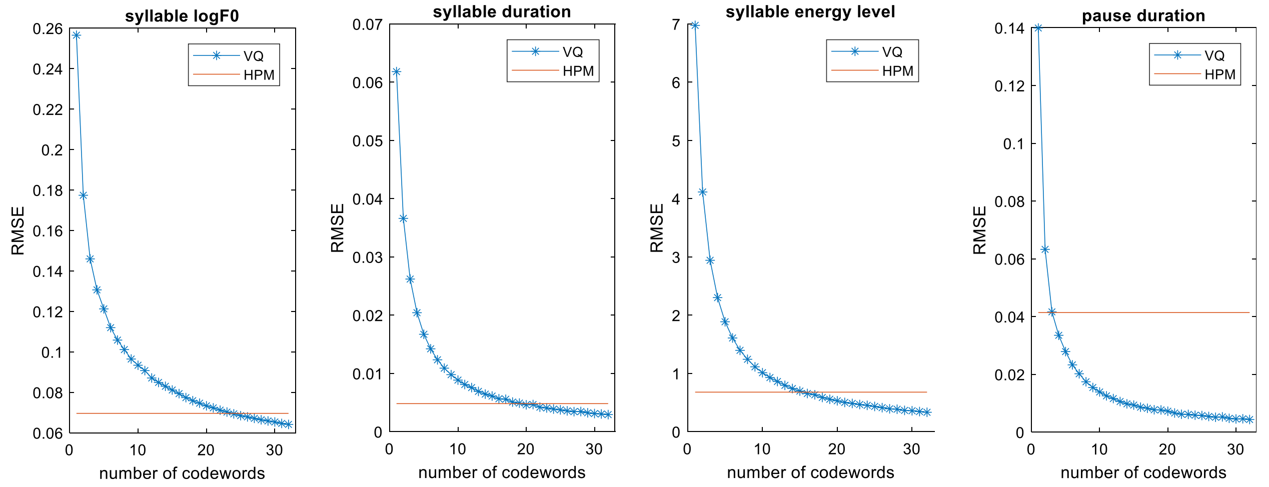


Fig. 4 The numbers of the codewords versus RMSEs of the four prosodic-acoustic features for the SD prosody coding of the inside data (TrainTB)

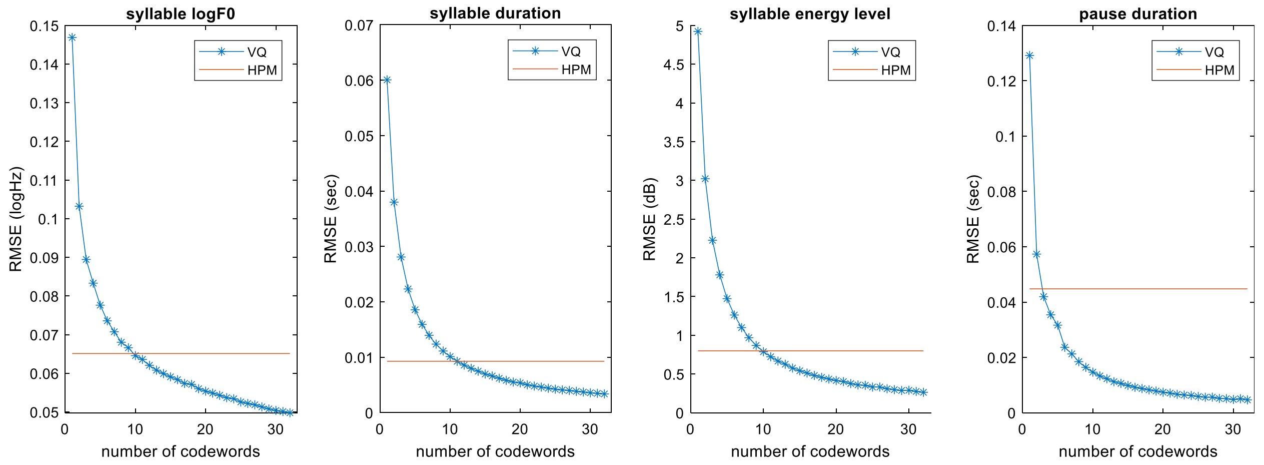


Fig. 5 The numbers of the codewords versus RMSEs of the four prosodic-acoustic features for the SI prosody coding of the training sets (TrainTC2)

using the "+" symbol. It is found that the score distributions of BSL are different from the ones of HPM and CPRO. We then examine the rank order of the different approach in terms of their MUSHRA scores and check if differences in the rank order were significant by the Kruskal-Wallis test at a p value of 0.01. It is found that in both of the SD and SI cases HPM, CPRO and NAT have mean ranks significantly different from BSL, and HPM has mean ranks not significantly different from CPRO.

3.3.2 Comparison between the baselines with various numbers of logF0 Codewords and the proposed approach

The results of the MUSHRA tests show that the baseline coder cannot generate as natural as the proposed coder does. By a detail analysis, we find that many syllables generated by the baseline approach sound incorrect in tone perception. This indicates that the baseline approach does not have enough codewords to represent crucial logF0 contours for Mandarin Chinese. We therefore increases number of the codewords for the logF0 contour and keep the numbers of codewords for syllable duration, syllable energy level and pause duration the same when implementing the baseline system. For the SD case, we conduct a MUSHRA test in which each presented page contains synthesized speeches from the baseline system with 24, 32, 64, 128, and 256 logF0 codewords, the proposed approach (SD-HPM), and the proposed approach with correct prosodic-acoustic features. The baseline systems with 24, 32, 64, 128, and 256 codewords are denoted as SD-BSL-24, SD-BSL-32, SD-BSL-64, SD-BSL-128, and SD-BSL-256, respectively. Because this MUSHRA experiment focused on the logF0 contour feature, the speech from the proposed approach with correct prosodic-acoustic features (denoted by SD-CPRO) instead of the vocoded natural speech is taken as a reference (100 point) for the assessment. The utterances chosen for this MUSHRA test were the same as the ones used in the previous SD MUSHRA test. This SD MUSHRA test therefore has ten pages, and each of the pages has 7 speech instances from the different systems. For the SI case, we also conduct a MUSHRA test with the same utterances chosen in the previous SI MUSHRA test. In this SI MUSHRA test, each presented page contains synthesized speeches from the baseline system with 10, 16, 32, 64, 128, and 256 codewords, the proposed approach, and the proposed approach with correct prosodic-acoustic features. The speech from the proposed approach with correct prosodic-acoustic features (SI-CPRO) is also taken as a reference (100 point) for the assessment. Accordingly, this SI MUSHRA test

Table 8 Bit assignment for Codewords used in the baseline prosody coding

Symbol	No. of symbol (SD/SI)	No. of bit (SD/SI)
Base-syllable type s_n	411/411	9/9
Syllable pitch contoursp_n	24/10	5/4
Syllable duration sd_n	19/11	5/4
Syllable energy level se_n	16/10	4/4
Pause duration pd_n	3/3	2/2
Total number per syllable	473/445	25/23

Table 9 The statistics of the scores (mean ± 1 standard deviation) rated in the SI and SD MUSHRA tests with the vocode natural speech (NAT), the proposed approach with correct prosodic-acoustic features (CPRO), the proposed approach (HPM), and the baseline system (BSL) with similar RMSE to the proposed approach

	BSL	HPM	CPRO	NAT
SD	50.1 ± 22.0	74.5 ± 18.8	78.7 ± 16.5	94.9 ± 9.9
SI	48.3 ± 21.5	74.1 ± 17.4	72.2 ± 18.3	95.8 ± 10.4

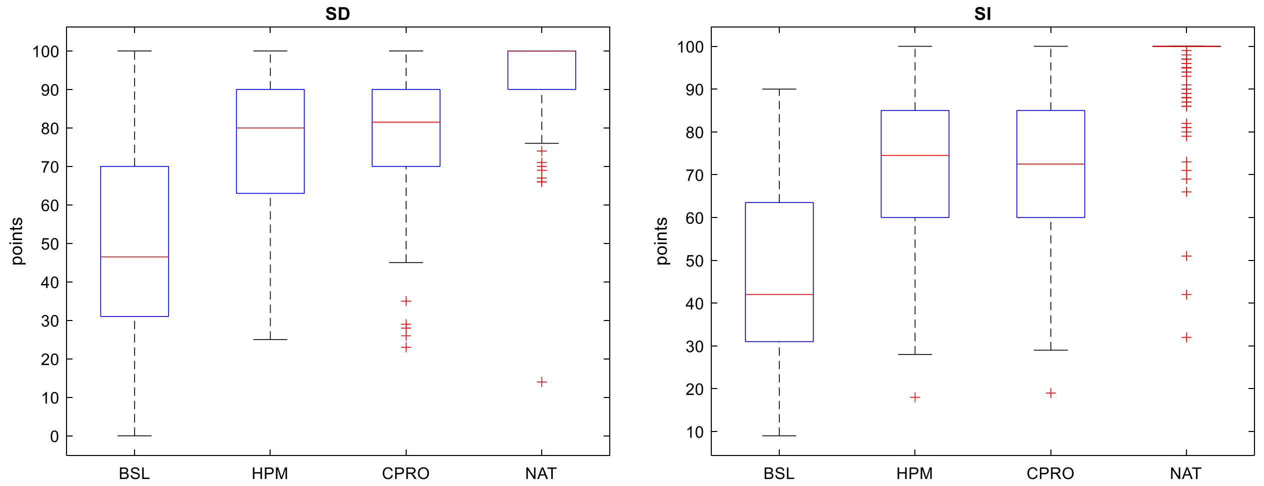


Fig. 6 Boxplots for the approaches of BSL, HPM, CPRO, and NAT in the SD case (left) and the SI case (right)

has 8 speech instances for each of the test pages. The baseline system with 10, 16, 32, 64, 128, and 256 codewords are denoted as SI-BSL-10, SI-BSL-16, SI-BSL-32, SI-BSL-64, SI-BSL-128, and SI-BSL-256, respectively. Table 6 also shows the RMSEs of logF0 encoding with various codeword sizes for the SD and SI cases. The decreasing trends of RMSRs for logF0 can be observed for the both cases when the numbers of the codewords increase. Table 10 shows the means and standard deviations of the rated scores in the MUSHRA tests with various codeword sizes for encoding logF0. Figure 7 shows the results of the MUSHRA tests. It can be found from the tables and figures that the subjective scores increase as the numbers of codewords increase. The subjective scores of the 256 codewords (SD-BSL-256 and SI-BSL-256) are close to but still lower than the scores of the proposed coding approach (SD-HPM and SI-HPM). We also examine the rank order of the different approach in terms of their MUSHRA scores and check if differences in the rank order were significant by the Kruskal-Wallis test at a p value of 0.01. In the SD case, besides of the pairs of {SD-BSL-24, SD-BSL-32} and {SD-BSL-128, SD-BSL-64}, all the other pairs are significantly different in the rank orders. The result proved that the proposed coding scheme outperforms the baseline in the subjective tests for the SD case. In the SI case, SI-BSL-16 is not significantly different from SI-BSL-10 and SI-BSL-32; SI-BSL-64 is not significantly different from SI-BSL-32, SI-BSL-128, and SI-BSL-256; and SI-HPM is not significant different from SI-CPRO and SI-BSL-256. The result indicated that the performance of the proposed approach is close to the ones of the baseline system with 256 codewords for logF0 (SI-BSL-256) and the oracle prosody encoder with correct prosodic-acoustic features (SI-CPRO).

To find the reason why the best baseline systems with 256 codewords for the SD and SI case still cannot reach the performance of the proposed approach (HPM), we did a very careful check for the synthesized utterances by the baseline systems. It was found that some tone-2 and tone-4 syllables in the first half of a sentence or a PPh are spoken with high pitch but encoded by some improper logF0 contours, and therefore sounded unnatural or incorrect in tone perception. This result indicated that the baseline approach that works without knowledge of tone and prosodic structure of Mandarin cannot learn meaningful logF0 contours as the proposed approach can. The proposed approach, however, can encode perceptually meaningful logF0 contours with the additive APs of tone, coarticulation, and pitch prosodic state, where the tone APs represent typical local patterns for tone; the coarticulation

Table 10 The statistics of the scores (mean ± 1 standard deviation) rated the MUSHRA tests

(a) the SD case							
SD-CPRO	SD-HPM	SD-BSL-24	SD-BSL-32	SD-BSL-64	SD-BSL-128	SD-BSL-256	
93.7±13.2	91.1 ± 9.6	64.9 ± 19.8	67.2 ± 18.1	77.0 ± 16.6	80.3 ± 15.7	86.5 ± 12.1	
(b) The SI case							
SI-CPRO	SI-HPM	SI-BSL-10	SI-BSL-16	SI-BSL-32	SI-BSL-64	SI-BSL-128	SI-BSL-256
90.3±12.1	91.0 ± 12.6	64.3 ± 23.4	67.6 ± 23.5	75.2 ± 20.6	78.8 ± 19.1	81.1 ± 19.1	84.8 ± 16.3

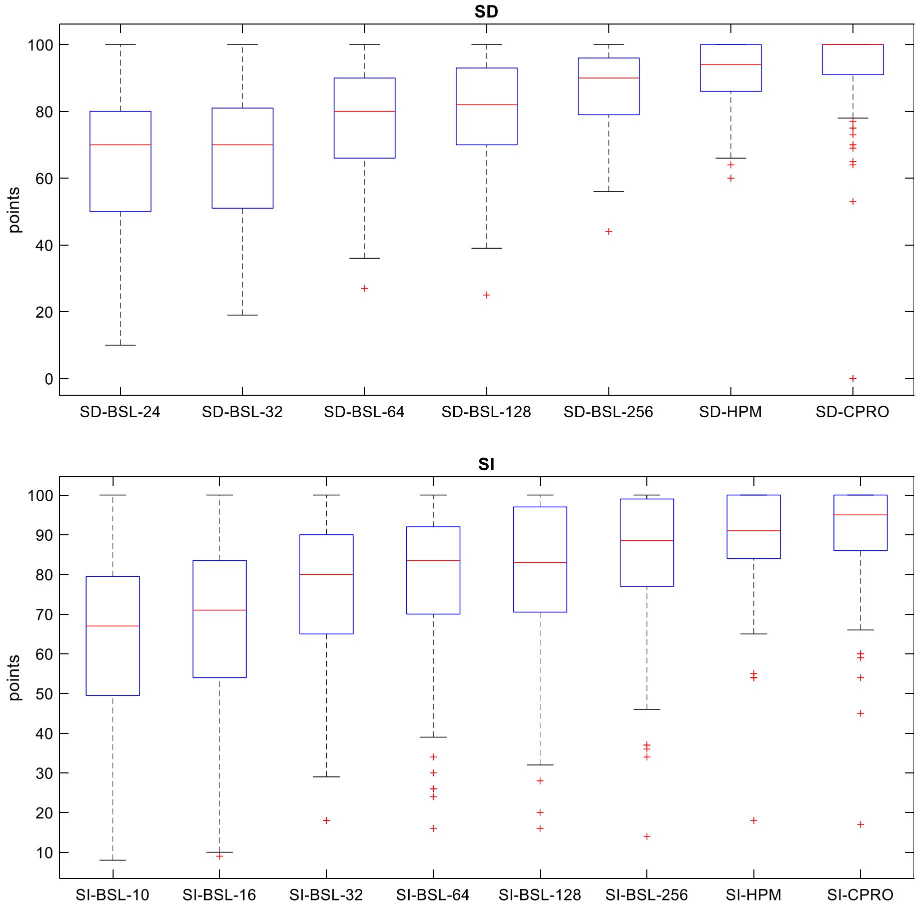


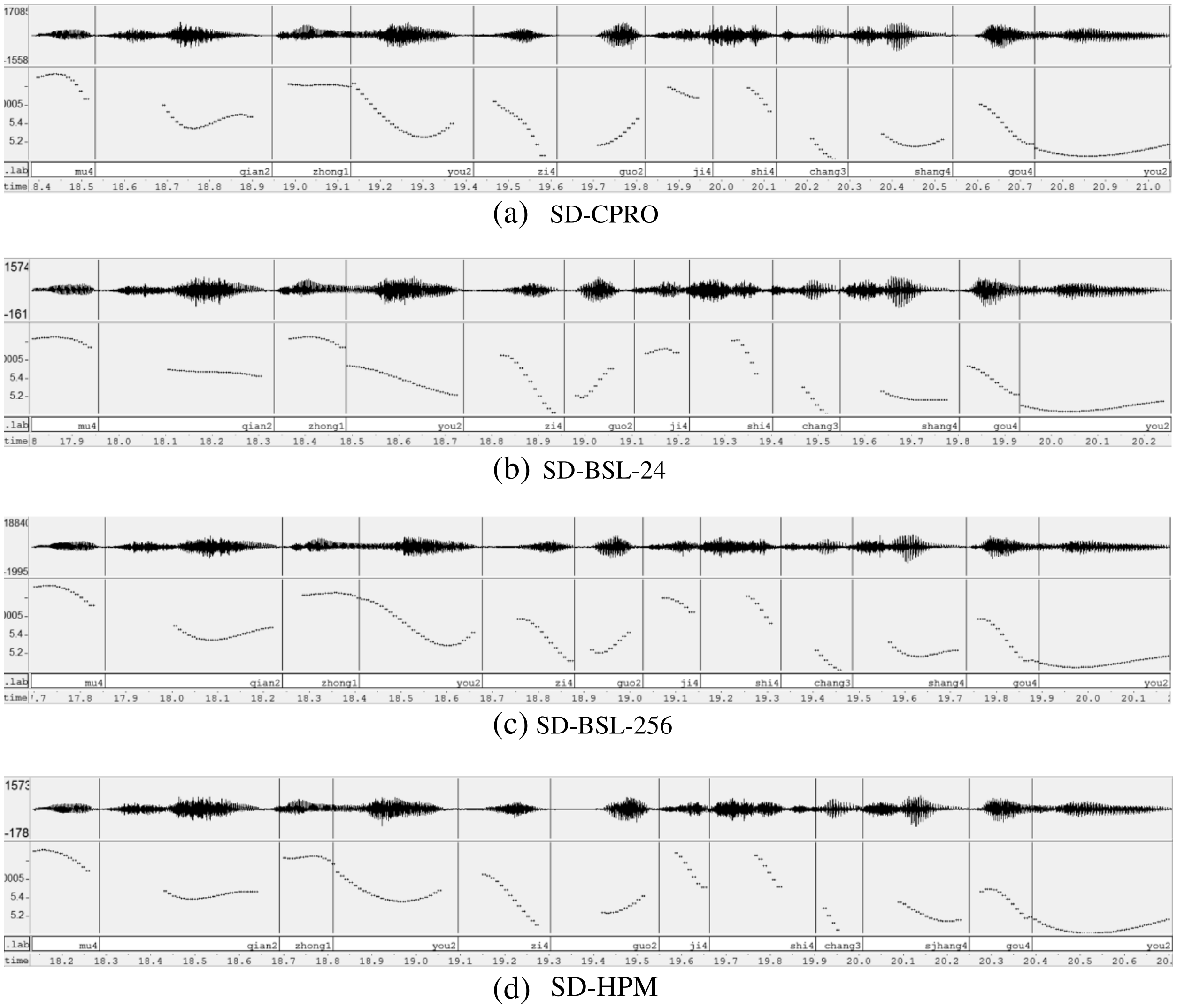
Fig. 7 Boxplots for the approaches of BSL with various codeword sizes, HPM, and CPRO in the SD case (up) and the SI case (down)

APs describe the logF0 patterns of neighboring tones' interaction; and the prosodic state APs represent global logF0 patterns of intonation.

Figure 8 shows four examples of the reconstructed speech waveform segments and their corresponding syllable boundaries and logF0 contours by the prosody coding approach of SD-CPRO, SD-BSL-24, SD-BSL-256, and SD-HPM. Taking SD-CPRO (Fig. 8a) as an reference, some of the logF0 contours generated by SD-BLS-24 (Fig. 8b) are stylized and discontinuous at syllable junctures. Although the approaches of SD-BSL-24 and SD-HPM have the similar RMSE performances, the speech of SD-BLS-24 sounds unnatural, especially for the second syllable "qian2" and the fourth syllable "you2" which are tone 2 syllables. In this case, they sounds like tone 3 syllables. The logF0 contours of SD-HPM (Fig. 8d) are not as vivid or dynamic as the ones of SD-CPRO. The tones of all the syllables of SD-CPRO, however, sound correct in tone perception and natural because the logF0 contours are continuous at the syllable junctures (e.g., the junctures between the third and fourth syllables "zhong1-you2" and the last two syllables "gou4-you2") and have similar logF0 slopes and curvatures to the ones of SD-CPRO (e.g., the second and fourth syllables "qian2" and "you2"). Although the logF0 contours of SD-BSL-256 (Fig. 8c) are very close to the ones of SD-CPRO (Fig. 8a), the fourth syllable "you2" sounds like a tone 3 syllable. It is found that the RMSE for the fourth syllable "you2" of SD-BSL-256 is smaller than the ones of SD-HPM. The downward concave logF0 in the first half of the fourth syllable "you2" of SD-BSL-256, nevertheless, makes the syllable sound like tone 3. On the other hand, the logF0 contour of the fourth syllable "you2" of SD-HPM concaves upward, matches well with the logF0 trend of SD-CPRO and sounds natural. These examples can partially illustrate the reason why the proposed prosody coding scheme can perform better than the conventional *k*-Mean coding scheme.

3.3.3 Analysis of bit rates

Tables 11 and 12 show the data rates of the proposed system and the baseline systems in the SD speech coding task in the units of bits/syllable and bits/second, respectively. In the non-compression case (denoted as NC), the baseline system reaches the same numbers of bits per syllable and



(a) SD-CPRO

(b) SD-BSL-24

(c) SD-BSL-256

(d) SD-HPM

Fig. 8 Four examples of the reconstructed prosodic-acoustic features by the prosody coding approach of (**a**) SD-CPRO, (**b**) SD-BSL-24, (**c**) SD-BSL-256, and (**d**) SD-HPM. The texts are "mu4-qian2 zhong1-you2 zi4 guo2-ji4 shi4-chang3 shang4 gou4 you2 (目前中油自國際市場上購油)."

per second when logF0 is encoded with 128 codewords. This means that the propose system can perform better than the baseline system at the same bit rate (27 bits/syllable or 102.6 bits/second) in terms of subjective tests. We then apply entropy coding scheme to compress the source symbols, i.e., the syllable type for both of the baseline and the proposed systems, the codewords of {sp_n, sd_n, se_n, and pd_n} solely for the baseline system, and tone, pitch/duration/energy prosodic state tags, and break type tags, for the proposed system. We first assume each of these symbol types is the Markov information source of zero order. Specifically, the symbols are encoded by the Huffman coding method with the information of the unigram probability for each type of the symbols obtained in the training set, i.e., TrainTB set. The bit rates are listed in the columns denoted by M0 (zero-order Markov) in Tables 11 and 12. In compared with the non-compressed case (NC), this entropy coding can reduce average bits/syllable and bits/second, standard deviations, and the minimum of bits/second. In the test set, the bit rate of the proposed scheme with Huffman coding with zero-order Markov source assumption (23.04 bits/syllable or 87.5 bits/second) is slightly lower than the one of the baseline system SD-BSL-256 (24.64 bits/syllable or 93.6 bits/second) whose subjective performance is most close to the proposed approach.

To further reduce bit rates, the symbols are encoded by the Huffman coding scheme with the information of the bigram probabilities (the Markov information source of first order.). For the proposed approach, the symbols related to base syllable types, tones, and break types are encoded by their bigram probabilities. The prosodic states are encoded by the probabilities provided by the

Table 11 Statistics of bits per syllable of the proposed approach and the baselines with various logF0 codebooks for the SD speech coding case

	Bits per syllable						
	Average			Standard deviation		Maximum/Minimum	
	NC	M0	M1	M0	M1	M0	M1
(a) Inside							
SD-HPM	27	23.23	20.65	0.30	0.38	24.23/22.57	22.09/19.59
SD-BSL-24	25	20.95	20.12	0.20	0.32	21.84/20.44	21.21/19.16
SD-BSL-32	25	21.29	20.45	0.20	0.32	22.14/20.73	21.52/19.35
SD-BSL-64	26	22.31	21.20	0.20	0.31	23.20/21.81	22.14/20.24
SD-BSL-128	27	23.09	21.83	0.20	0.32	23.96/22.54	22.71/20.78
SD-BSL-256	28	24.10	22.12	0.20	0.32	24.95/23.55	22.93/21.14
(b) outside							
SD-HPM	27	23.04	21.42	0.29	0.39	23.72/22.45	22.31/20.48
SD-BSL-24	25	21.31	21.01	0.19	0.34	21.96/21.02	21.62/20.19
SD-BSL-32	25	21.59	21.38	0.20	0.37	22.25/21.26	22.05/20.35
SD-BSL-64	26	22.65	22.72	0.21	0.56	23.37/22.34	23.82/21.48
SD-BSL-128	27	23.59	27.70	0.19	1.61	24.29/23.28	31.69/24.63
SD-BSL-256	28	24.64	54.37	0.21	6.60	25.37/24.31	67.40/41.36

three prosodic state transition models (illustrated in Eq. (14)), i.e., $P(p_n| p_{n-1}, B_{n-1})$,$P(q_n| q_{n-1}, B_{n-1})$, and $P(r_n| r_{n-1}, B_{n-1})$, which directly well describe the first order Markov property of the prosodic states conditioned on the break types. For the baseline approach, the symbols related to base syllable types and the codewords of the pause duration are encoded by their bigram probabilities. The codewords of pitch contour, syllable duration, and syllable energy level are encoded by the probabilities of $P(c(sp_n)| c(sp_{n-1}), c(pd_{n-1}))$, $P(c(sd_n)| c(sd_{n-1}), c(pd_{n-1}))$, and $P(c(se_n)| c(se_{n-1}), c(pd_{n-1}))$, where $c(x)$ is the codeword of the feature x. It is noted that the roles of $c(sp_n)$, $c(sd_n)$, and $c(se_n)$ are analogous to the ones of p_n, q_n, and r_n, and the role of $c(pd_{n-1})$ is analogous to the one of B_{n-1}. The design of the bigram probabilities for the Huffman coding of the baseline approach is therefore fair for comparing with

Table 12 Statistics of bits per second of the proposed approach and the baselines with various logF0 codebooks for the SD speech coding case

	Bits per second								
	Average			Standard deviation			Maximum/Minimum		
	NC	M0	M1	NC	M0	M1	NC	M0	M1
(a) Inside									
SD-HPM	104.1	89.5	79.6	4.8	4.0	3.6	103.9/81.2	118.1/71.7	91.5/62.9
SD-BSL-24	96.4	80.7	77.3	4.4	3.6	3.3	93.1/75.1	109.3/64.2	88.1/61.0
SD-BSL-32	96.4	81.8	78.6	4.4	3.6	3.3	94.4/75.1	109.3/65.2	89.2/62.2
SD-BSL-64	100.3	85.4	81.7	4.6	3.8	3.5	98.6/78.2	113.7/67.9	92.7/63.8
SD-BSL-128	104.1	88.8	83.8	4.8	4.0	3.6	102.4/81.2	118.1/70.2	94.9/65.3
SD-BSL-256	108.0	92.7	84.7	5.0	4.1	3.7	106.5/84.2	122.4/73.6	95.8/66.1
(b) Outside									
SD-HPM	102.6	87.5	81.4	4.3	3.5	3.3	95.0/88.1	110.6/76.1	88.7/71.8
SD-BSL-24	95.0	80.8	79.4	4.0	3.4	3.0	87.3/81.5	102.4/69.5	84.7/69.7
SD-BSL-32	95.0	82.1	81.1	4.0	3.5	3.0	88.6/81.5	102.4/70.6	87.1/71.7
SD-BSL-64	98.8	85.7	86.3	4.2	3.6	3.5	92.4/84.8	106.5/73.7	94.5/76.1
SD-BSL-128	102.6	89.3	105.2	4.3	3.8	6.7	96.3/88.1	110.6/76.8	116.3/91.0
SD-BSL-256	106.4	93.6	206.1	4.5	4.0	23.7	101.0/91.3	114.7/80.5	257.1/147.0

Table 13 Statistics of bit rates per syllable of the proposed approach and the baselines with various logF0 codebooks for the SI speech coding case

	Bits per syllable						
	Average			Standard deviation		Maximum/Minimum	
	NC	M0	M1	M0	M1	M0	M1
(a) Inside							
SI-HPM	27	22.62	18.08	0.37	0.46	24.33/21.69	20.07/16.84
SI-BSL-10	23	18.40	17.61	0.21	0.32	19.15/17.63	19.00/16.64
SI-BSL-16	23	18.97	18.16	0.21	0.34	19.74/18.20	19.67/17.03
SI-BSL-32	24	19.67	19.00	0.21	0.37	20.43/18.91	20.44/17.99
SI-BSL-64	25	20.55	19.84	0.21	0.39	21.30/19.77	21.50/18.68
SI-BSL-128	26	21.51	20.61	0.21	0.38	22.22/20.71	22.67/19.53
SI-BSL-256	27	22.50	21.15	0.21	0.35	23.20/21.68	23.00/20.10
(b) Outside							
SI-HPM	27	22.87	18.94	0.35	0.44	23.98/22.04	20.32/17.82
SI-BSL-10	23	18.58	18.11	0.18	0.33	19.13/18.07	19.09/17.31
SI-BSL-16	23	19.13	18.67	0.18	0.34	19.66/18.60	19.95/17.94
SI-BSL-32	24	20.07	19.54	0.19	0.40	20.65/19.57	20.73/18.45
SI-BSL-64	25	20.85	20.60	0.18	0.60	21.40/20.34	23.40/19.26
SI-BSL-128	26	21.95	23.17	0.19	1.70	22.51/21.40	29.80/20.46
SI-BSL-256	27	22.80	35.23	0.18	6.64	23.33/22.26	61.12/23.55

Table 14 Statistics of bit rates per second of the proposed approach and the baselines with various logF0 codebooks for the SI speech coding case

	Bits per second								
	Average			Standard deviation			Maximum/Minimum		
	NC	M0	M1	NC	M0	M1	NC	M0	M1
(a) Inside									
SI-HPM	103.1	86.4	69.0	13.3	11.0	8.6	137.7/52.7	163.6/45.4	108.6/37.1
SI-BSL-10	87.9	70.3	67.2	11.3	9.0	8.3	111.5/44.9	139.4/35.6	105.5/35.0
SI-BSL-16	87.9	72.4	69.1	11.3	9.3	8.5	114.8/44.9	139.4/36.7	108.6/36.4
SI-BSL-32	91.7	75.0	72.5	11.8	9.6	8.9	119.1/46.8	145.4/37.9	113.7/38.0
SI-BSL-64	95.5	78.6	75.7	12.3	10.1	9.4	124.8/48.8	151.5/39.9	118.7/39.8
SI-BSL-128	99.3	82.3	78.7	12.8	10.6	9.8	130.6/50.8	157.5/41.6	124.3/40.9
SI-BSL-256	103.1	86.0	80.8	13.3	11.0	10.2	136.6/52.7	163.6/43.7	127.4/41.5
(b) Outside									
SI-HPM	103.6	87.7	72.7	13.3	11.2	9.4	121.2/77.6	142.6/64.6	102.4/52.4
SI-BSL-10	88.3	71.2	69.5	11.4	9.0	8.7	99.2/66.1	121.5/53.1	96.5/51.2
SI-BSL-16	88.3	73.3	71.2	11.4	9.3	8.9	102.1/66.1	121.5/54.7	98.2/52.2
SI-BSL-32	92.1	76.9	74.9	11.9	9.7	9.2	106.8/69.0	126.7/57.4	103.8/54.6
SI-BSL-64	95.9	80.2	79.2	12.4	10.2	9.4	111.3/71.9	132.0/60.0	109.3/56.8
SI-BSL-128	99.8	84.1	88.8	12.9	10.7	9.9	116.5/74.8	137.3/62.9	120.0/66.7
SI-BSL-256	103.6	87.5	136.7	13.3	11.1	22.2	121.3/77.6	142.6/65.4	204.0/93.9

the proposed approach. The bit rates for the bigram assumption are listed in the columns denoted by M1 (first order Markov) in Tables 11 and 12. It can be seen from the tables that the bit rates can be further reduced from the ones of the unigram assumption (M0) in the training set. The bit rates of M1 for the SD-BSL-64, SD-BSL-128, and SD-BSL-256 in the test set, however, are greater than the ones of M0, indicating that the bigram probabilities from the training set are overfitted although we have already apply standard probability smoothing technique to the estimations of these bigram probabilities. Concluding the bit rates shown in Tables 11 and 12, the lowest bit rates for the best proposed approach and the baseline in terms of the subjective tests are respectively 21.42 bits/syllable (or 81.4 bits/second) in the M1 encoding case and 24.64 bits/syllable (or 93.6 bits/second) in the M0 encoding case, showing the compactness of the proposed prosody coding scheme.

Tables 13 and 14 show the data rates of the proposed system and the baseline systems in the SI speech coding task in the units of bits/syllable and bits/second, respectively. The Huffman coding scheme with the the Markov information source of zero order (M0) and first order (M1) are also applied to reduce the data rates. The proposed approach can achieve the lowest data rate of 18.94 bits/syllable (or 72.7 bits/second) in the M1 encoding case, which is lower than the data rate of the best baseline, i.e., 22.80 bits/syllable (or 87.5 bits/second) in the M0 encoding case.

The data compression ratios (uncompressed/compressed) for the SD-HPM and SI-HPM by the M1 encoding case are

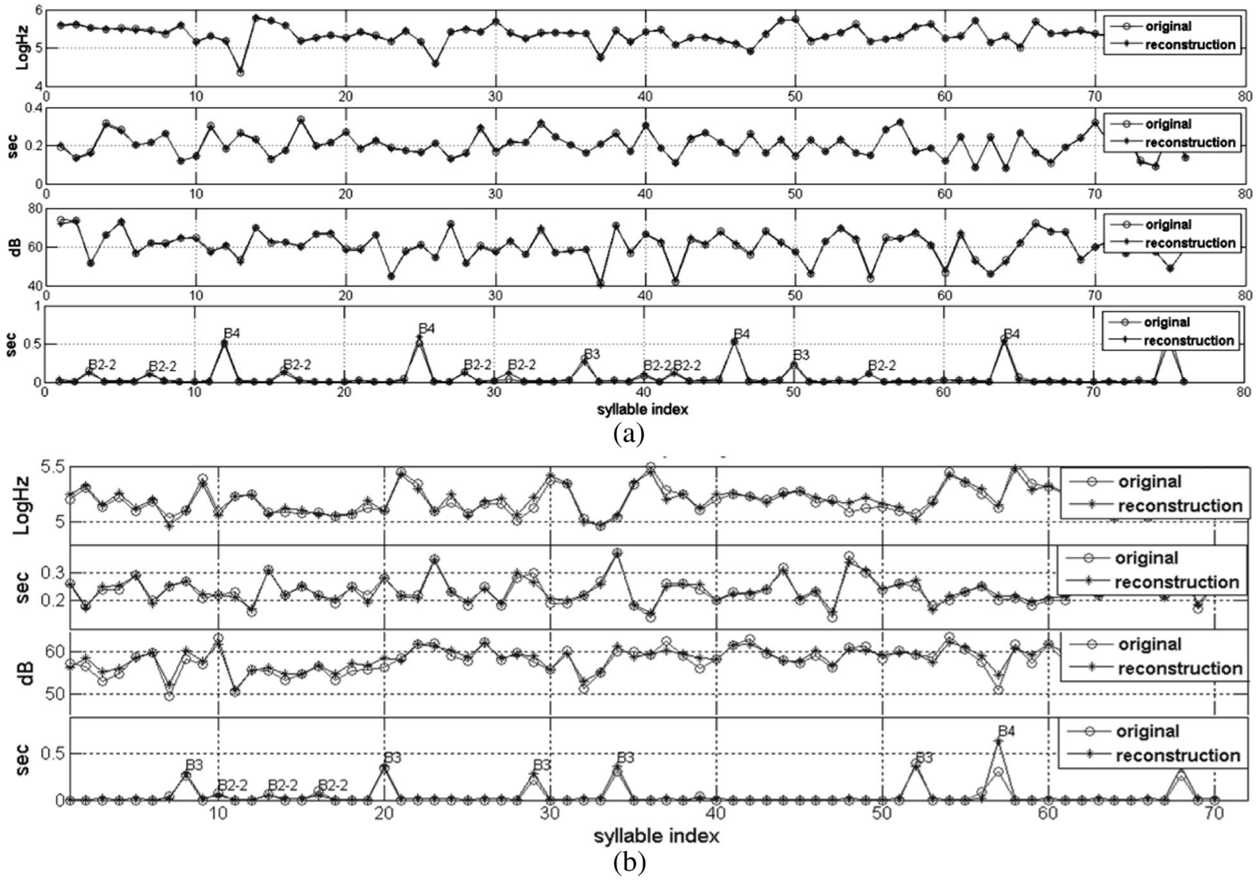


Fig. 9 Two examples of the reconstructed prosodic features for (**a**) an utterance in Treebank and (**b**) an utterance in TCC300. From top to bottom: syllable pitch mean, syllable duration, syllable energy level, and pause duration. (open circle: reference, star: reconstructed) The texts are (**a**) qin-yi gong-si xian-jin zeng-zi yi-dian liu-yi yuan,shen-gou ri-qi jie-zhi shi-yi-yue wu-ri wei-zhi。gai gong-si jin-nian-du xian-jin zeng-zi-gu jiang yu yuan-you gu-fen fen-kai gua-pai,er zeng-zi-gu yu yuan-you gu-fen quan-li yi-wu jian bu-tong de shi,dui ben nian-du ying-yu fen-pei de quan-li,(勤益 公司 現金 增資 一點 六億 元,申購 日期 截至 十一月 五日 為止。該 公司 今年度 現金 增資股 將 與 原有 股份 分開 掛牌,而 增資股 與 原有 股份 權利 義務 間 不同 的 是,對 本年度 盈餘 分配 的 權利,); and (**b**) lian-ri lai gai qiao zhi yin-dao yin zhi pu yi-ceng bo-bo de bo-you lu-mian jing zhong-xing sha-shi-che zhi zhan ya lu-mian yi sun-huai qian-tian wan-jian ceng fa-sheng qi-che yin lu-kuang bu. shou zhuang-che shi-jian suo-xing wei fan-fu shu-lin zhen-gong-suo ri-zuo pai-yuan kan-cha fa-xian … (連日 來 該 橋 之 引道 因 只 鋪 一層 薄薄 的 柏油 路面,經 重型 砂石車 之 輾 壓 路面 已 損壞,前天 晚間 曾 發生 汽車 因 路況 不 熟 撞車 事件,所幸 未 翻覆。樹林 鎮 公所 日昨 派員 勘查,發現…)

1.26 and 1.43, respectively. The data compression ratios for the SD-BSL-256 and SI-BSL-256 by the M0 encoding case are 1.14 and 1.18, respectively. The higher data compression ratio achieved by the proposed coding scheme is mainly because of the high autocorrelation of the prosodic state sequence. The prosodic state sequence encompass the information of syllable prosodic-acoustic features being subtracted by the local frustration resulted from tone, base syllable type, and coarticulation effect. The patterns of the prosodic states are therefore smoother than the ones of the directly encoded observed prosodic-acoustic features, i.e., $c(sp_n)$, $c(sd_n)$, and $c(se_n)$. Besides, the M1 encoding for the HPM symbols just matches the properties of model parameters of the HPM in Eq. (14), i.e., $P(p_n| p_{n-1}, B_{n-1})$,$P(q_n| q_{n-1}, B_{n-1})$, and $P(r_n| r_{n-1}, B_{n-1})$. Evidently, these above-mentioned properties make the proposed coding scheme encode speech more efficiently than the baseline system does when the proposed approach has better subjective test results than the baseline in the SD case, and equal subjective results to the baseline in the SI case.

It is interesting to find that the proposed approach in the SD case needs a higher bit rate (81.4 bps) than the SI case does (72.7 bps) to achieve the subjective qualities which are closed to the ones by the oracle prosody encoder with correct prosodic-acoustic features (SD-CPRO and SI-CPRO). This finding is quite counter-intuitive because speech coding for the SI cases usually needs more codewords (or bit rates) than the SD cases to model speakers' variabilities in spectrum and prosody. Since the proposed coding method generates spectral information by the HPM parameters with the side information of the CD-HMM for modeling state duration and MGC, we disregarded the bit rates related to the spectrum coding part that usually contributes the majority of the bitstream in conventional speech coding cases. When considering prosody coding solely in this study, the higher bit rate for prosody indicates that the richer prosodic variation made by a speaker to convey more information. Recall that the SD and SI cases in this study encoded the prosodies of the TTS speech corpus (Treebank speech corpus) and the ASR speech corpus (TCC300), respectively. The professional announcer tended to utter the TTS corpus with rich prosodic variation for a better conveying of linguistic, paralinguistic, and non-linguistic information while the amateur speakers tended to utter the ASR corpus with flat prosody for just conveying linguistic information. So, the counter-intuitive result that the SD case needs a higher bit rate than the SI case is reasonable in this study. Figure 9 shows two typical examples of the reconstructed prosodic-acoustic features of two utterances of the outside test. As shown in the figure, most reconstructed prosodic features were close to their reference values. This shows that the proposed prosody coding approach is very promising.

4 Application to speaking rate conversion

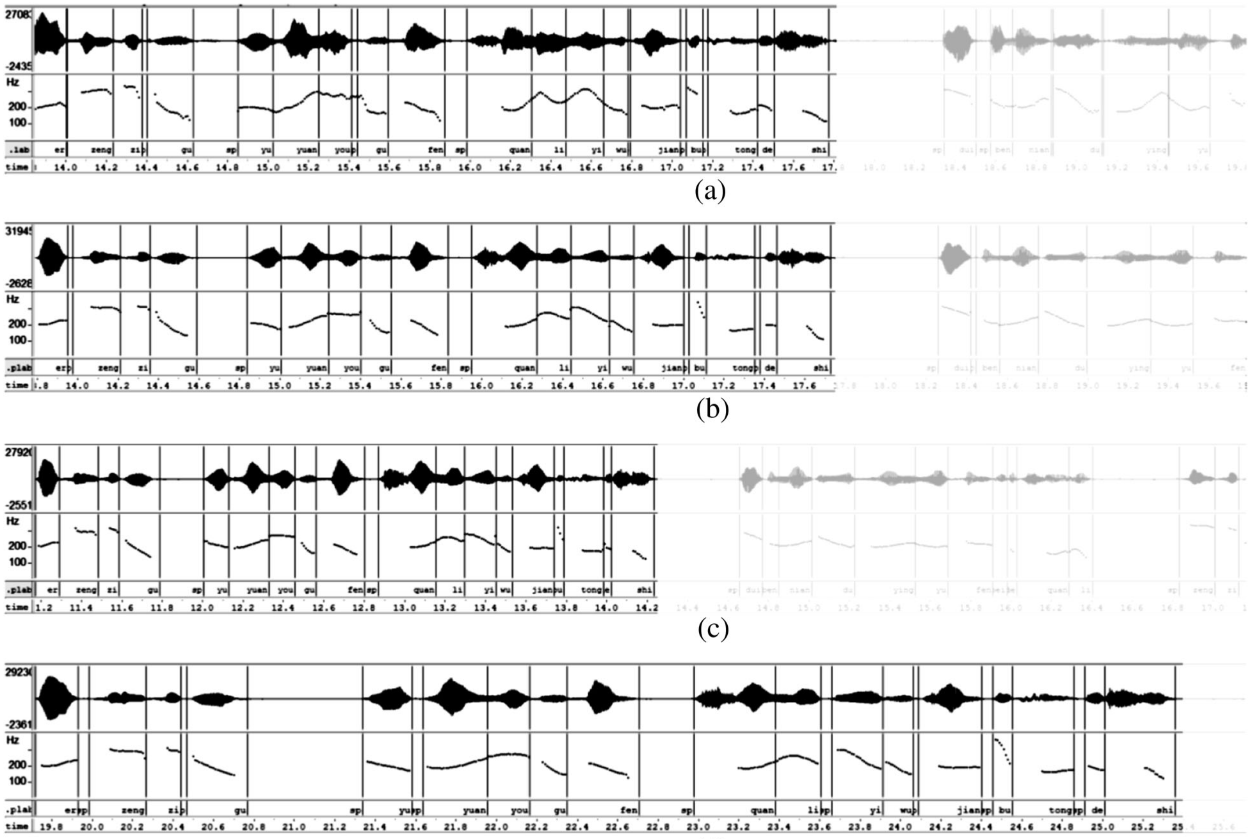
An example of modifying the speaking rate of a reconstructed speech via directly replacing the HPM parameters with those of two different speaking rates is illustrated. The task is to convert the source prosody of a reconstructed speech in the SD prosody coding task into the target prosody of a slower or faster speaking rate. This is realized by replacing the side information of the HPM used in the synthesis operation of the decoder with that of the HPM trained in the target speaking rate. Here, we simply assume that the source and target speeches have the same prosodic phrase structure so as to let the converted utterance share the same decoded break and prosodic-state tags of the source utterance. Since the source Treebank database was recorded in the most comfortable speed for the announcer, it is therefore regarded as the normal speaking-rate speech corpus. By taking the normal speech as a reference, two other parallel corpora, FastTB of the fast rate and SlowTB of the slow rate, were then recorded. Table 15 displays the statistics of these speech corpora. Notice that the speech rate (SR) is defined as the average number of syllable uttered per second, while the articulation rate (AR) is defined as the average number of syllable uttered per second excluding all pauses. Figure 10 displays the original waveform, the synthetic waveforms of normal, slow, and fast speaking rates, and their corresponding pitch contours. It can be seen from the figure that both syllable durations and pause durations were changed largely to match the given speaking rate, while waveforms and pitch contour shapes were mostly kept unchanged. An informal listening test confirmed that both converted speech of high and low speaking rate sounded very fluently and naturally.

5 Conclusions

A novel parametric prosody coding approach for Mandarin speech has been discussed in this paper. Its novelty lies in employing a sophisticated hierarchical prosodic model as the prosody-generating model to analyze the prosodic-acoustic features and extract the representing parameters for encoding in the encoder, and to synthesize the prosodic-acoustic features from the decoded representing parameters with the help of the hierarchical prosodic model in the decoder. Low

Table 15 The Statistics of Speech Corpora FastTB and SlowTB

Corpus	Utt#	Syl#	Hours	AR (syllables/seconds)	SR (syllables/seconds)
FastTB	368	50,691	3.4	5.52	4.40
TrainTB	376	51,868	3.9	5.05	3.82
TestTB	44	3895	0.3	4.89	3.78
SlowTB	372	51,231	6.0	3.78	2.46



(a)

(b)

(c)

(d)

Fig. 10 An example of speaking rate conversion: speech signal and pitch contour of (**a**) the original, (**b**) the reconstructed of the same speaking rate (normal), (**c**) the reconstructed of the faster speaking rate, and (**d**) the reconstructed of the slower speaking rate. The text is "er zeng-zi-gu yu yuan-you gu-fen quan-li yi-wu jian bu-tong de shi ...(而增資股與原有股份權利義務間不同的是...)"

average data rates of 81.4 and 72.7 bps have been reached respectively for the SD and SI cases at the condition of low RMSEs of the reconstructed prosodic-acoustic features with good synthesized speech quality. These data rates are lower than the ones encoded by the conventional segment-based prosody coding with scalar or vector quantization scheme. It is interesting to find this counter-intuitive result that the SD case needs a higher bit rate than the SI case does. The reason for the result may be that the SD speech corpus uttered by the professional announcer (for constructing a TTS) is richer in prosodic variation than the SI speech corpus uttered by the amateur speakers (for constructing ASR systems). The higher bit rate may indicate the more information to be conveyed in a speech with the richer prosodic variation. The use of the hierarchical prosodic model also provided the proposed approach an additional advantage of easily manipulating the prosody of the synthetic speech. An example to convert the speaking rate of the synthetic speech via changing the parameters of the hierarchical prosodic model in the decoding stage has been illustrated. This makes the proposed approach useful in some applications such as e-book reader.

A drawback of the proposed approach lies in the need of the text associated with the encoding speech. This can be cured by using a front-end automatic speech recognizer (ASR) to extract the linguistic information as well as to segment the speech signal. A prosody-assisted ASR [61, 62] can be used to help to solve the problem.

Abbreviations

AP: Affecting pattern; AR: Articulation rate; ASBC: Academia Sinica Balanced Corpus of Modern Chinese; ASR: Automatic speech recognition; BG/PG: Breathe/prosodic phrase group; BSL: Baseline; CD-HMM: Context-dependent hidden Markov model; CELP: Code-excited linear prediction; CMAPLR: Constrained maximum a posteriori estimation linear regression; CPRO: Encoding prosody by the proposed HPM with Correct PROsodic-acoustic features; F0: Fundamental frequency; HMM: Hidden Markov model; HPM: Hierarchical prosodic model; HTK: Hidden Markov model toolkit; LPC: Linear predictive coding; M0: The Huffman coding scheme with the Markov information source of zero order; M1: The Huffman coding scheme with the Markov information source of first order; MELP: Mixed excitation linear prediction; MGC: Mel-generalized cepstral coefficient; MLSA: Mel log spectrum approximation; MQ: Matrix quantization; MUSHRA: MUltiple Stimuli with Hidden Reference and Anchor; NAT: Vocoded NAtural speech with 24-dimensional Mel-generalized cepstral coefficients; PCM: Pulse Code Modulation; PLA: piecewise linear approximation; PLM: Prosody labeling and modeling; PPh: Prosodic phrase; PW: Prosodic word; SD: Speaker dependent; SI: Speaker independent; SR: Speech rate; SYL: Syllable; TTS: Text-to-speech; VQ: Vector quantization

Acknowledgements
The author would like to thank Prof. Sin-Horng Chen of National Chiao Tung University, Taiwan, for his generous guidance.

Funding
This work was supported by the Ministry of Science and Technology of Taiwan under Contract Nos. NSC-102-2221-E-305-005-MY3, MOST-105-2622-E-305-003-CC3, and MOST-106-2221-E-305 -010.

Availability of data and materials
All the speech utterances used in the MUSHRA subjective tests can be found at the website: http://cychiang.tw/pc_hpm/HPM_prosody_coding.htm.

Author's contribution
CYC is the single author of this paper. The author read and approved the final manuscript.

Authors' information
Chen-Yu Chiang obtained the B.S., M.S., Ph.D. degrees in communication engineering from National Chiao Tung University, Hsinchu, Taiwan, in 2002, 2004, and 2009, respectively. He is currently the Director of the Speech and Multimedia Signal Processing Lab and an Assistant Professor in the Department of Communication Engineering, National Taipei University, Taiwan. His research interests include speech processing, focusing on prosody modeling, automatic speech recognition, and text-to-speech systems (website: http://cychiang.tw/).

Competing interests
The author declares that there are no competing interests.

Publisher's Note

Received: 12 November 2017 Accepted: 2 July 2018
Published online: 11 July 2018

References

1. ITU-T (1993). *Pulse code modulation (PCM) of voice frequencies, technical report G.711*. Geneva: International Telecommunications Union.
2. ADPCM ITU-T (1990). *5-, 4-, 3- and 2-bits per Sample Embedded Adaptive Differential Pulse Code Modulation (ADPCM), Technical Report G.727*. Geneva: International Telecommunications Union.
3. Campbell, JP, & Tremain, TE (1986). Voiced/unvoiced classification of speech with applications to the U.S. government LPC-10E algorithm. In *Proc. ICASSP'86*, (pp. 473–476).
4. DDVPC, "LPC-10e speech coding standard, technical report FS-1015," U.S. Dept. of Defense Voice Processing Consortium, 1984.
5. McCree, AV, & Barnwell III, TP. (1995). A mixed excitation LPC vocoder model for low bit rate speech coding. *IEEE Trans. Speech Audio Process*, *3*(4), 242–250.
6. Schroeder, M, & Atal, B (1984). Code-excited linear prediction(celp): high-quality speech at very low bit rates. In *Proc. ICASSP'84*, (pp. 937–940).
7. Campbell, JP, Welch, VC, Tremain, TE (1989). An expandable error-protected 4800 BPS CELP coder (U.S. Federal Standard 4800 BPS voice coder). In *Proc. ICASSP'89*, (pp. 735–738).
8. DDVPC, "CELP speech coding standard, technical report FS-1016," U.S. Dept. of Defense Voice Processing Consortium, 1989.
9. ITU-T (1992). *Coding of speech at 16 kbit/s using low-delay code excited linear prediction, technical report G.728*. Geneva: International Telecommunications Union.
10. ETSI (1999). Universal Mobile Telecommunications System (UMTS); Mandatory Speech Codec Speech Processing Functions AMR Speech Codec; Transcoding Functions, 3G TS. 26.090 Version 3.1.0, Release 1999.
11. Roucos, S, Schwartz, R, Makhoul, J (1982). Segment quantization for very-low-rate speech coding. In *Proc. ICASSP'82*, (pp. 1565–1568).
12. Roucos, S, Schwartz, RM, Makhoul, J (1983). A segment vocoder at 150 b/s. In *Proc. ICASSP'83*, (pp. 61–64).
13. Roucos, S, & Wilgus, AM (1985). The waveform segment vocoder: a new approach for very-low-rate speech coding. In *Proc. ICASSP'85*, (pp. 236–239).
14. Tsao, C, & Gray, R. (1985). Matrix quantizer design for LPC speech using the generalized Llyod algorithm. *IEEE Trans. Acoust., Speech Signal Process.*, *33*(3), 537–545.
15. Shiraki, Y, & Honda, M. (1988). LPC speech coding based on variable length segment quantization. *IEEE Trans Acoust., Speech Signal Process.*, *36*(9), 1437–1444.
16. Cernocky, J, Baudoin, G, Chollet, G (1998). Segmental vocoder-going beyond the phonetic approach. In *Proc. ICASSP'98*, (pp. 605–608).
17. Holmes, WJ (1998). Towards a unified model for low bit-rate speech coding using a recognition-synthesis approach. In *Proc. ICSLP'98*.
18. Lee, K-S, & Cox, RV. (2001). A very low bit rate speech coder based on a recognition/synthesis paradigm. *IEEE Trans. Speech Audio Process.*, *9*(5), 482–491.
19. Cox, RV, & Lee, K-S. (2002). A segmental speech coder based on a concatenative tts. *Speech Commun.*, *38*(1), 89–100.
20. Baudoin, G, & El Chami, F (2003). Corpus based very low bit rate speech coding. In *Proc. ICASSP'03*, (pp. 792–795).
21. Chevireddy, S, Murthy, HA, Sekhar, CC (2008). Signal processing based segmentation and HMM based acoustic clustering for a syllable based segment vocoder at 1.4Kbps. In *Proc. EUSIPCO-2008*.
22. Harish, D, & Ramasubramanian, V (2008). Comparison of segment quantizers: VQ, MQ, VLSQ and unit-selection algorithms for ultra low bit-rate speech coding. In *Proc. ICASSP'2008*, (pp. 4773–4776).
23. Pradhan, A, Chevireddy, S, Veezhinathan, K, Murthy, H (2010). A low-bit rate segment vocoder using minimum residual energy criteria. In *Proc. NCC'10*.
24. Schwartz, R, Klovstad, J, Makhoul, J, Sorensen, J (1980). A preliminary design of a phonetic vocoder based on a diphone model. In *Proc. ICASSP'80*, (pp. 32–35).
25. Picone, J, & Doddington, GR (1989). A phonetic vocoder. In *Proc. ICASSP'89*, (pp. 580–583).
26. Ismail, M, & Ponting, K (1997). Between recognition and synthesis-300 bits/second speech coding. In *Proc. EUROSPEECH'97*, (pp. 441–444).
27. H.-C. Chen, C.-Y. Chen, K.-M. Tsou, and O. T.-C. Chen, "A 0.75 kbps speech codec using recognition and synthesis schemes," in Proc. IEEE Workshop Speech Coding Telecommunications, Sept. 1997, pp. 27–29.
28. Ribeiro, CM, & Trancoso, IM (1997). Phonetic vocoding with speaker adaptation. In *Proc. EUROSPEECH'97*, (pp. 1291–1294).
29. Tokuda, K, Masuko, T, Hiroi, J, Kobayashi, T, Kitamura, T (1998). A very low bit rate speech coder using HMM-based speech recognition/synthesis techniques. In *Proc. ICASSP'98*, (pp. 609–612).
30. Masuko, T, Tokuda, K, Kobayashi, T (1998). A very low bit rate speech coder using HMM with speaker adaptation. In *Proc. ICSLP'98*.
31. Hoshiya, T, Sako, S, Zen, H, Tokuda, K, Masuko, T, Kobayashi, T, Kitamura, T (2003). Improving the performance of HMM-based very low bitrate speech coding. In *Proc. ICASSP'03*, (pp. 800–803).
32. Halaly, I, & Bistritz, Y (2008). A phonetic vocoder with adaptation to selectable speaker codebooks. In *Proc EUSIPCO-2008*.
33. Halaly, I, & Bistritz, Y (2008). A phonetic vocoder with scalable adaptation to speaker codebooks. In *Proc. IEEEI'2008*, (pp. 684–688).
34. Benbassat, G, & Delon, X (1984). Low bit rate speech coding by concatenation of sound units and prosody coding. In *Proc. ICASSP'84*, (pp. 121–124).
35. Vepyek, P, & Bradley, AB (1997). Consideration of processing strategies for very-low-rate compression of wide band speech signal with known text transcription. In *Proc. EUROSPEECH'97*, (pp. 1279–1282).
36. Lee, K-S, & Cox, RV (1999). TTS based very low bit rate speech coder. In *Proc. ICASSP'99*, (pp. 181–184).
37. Wong, D, Juang, B-H, Gray Jr, AH. (1982). An 800 bit/s vector quantization LPC vocoder. *IEEE Trans. Acoust., Speech Signal Process.*, *30*(5), 770–780.
38. Paliwal, KK, & Atal, BS. (1993). Efficient vector quantization of LPC parameters at 24 bits/frame. *IEEE Trans. Speech Audio Process.*, *1*(1), 3–14.
39. Itakura, F. (1975). Line spectrum representation of linear predictive coefficients of speech signals. *J. Acoust. Soc. Amer.*, *57*(1), 35.
40. Atal, BS. (1982). Predictive coding of speech at low bit rates. *IEEE Trans. Commun.*, *30*(4), 600–614.
41. Chiang, C-Y, Chen, S-H, Yu, H-M, Wang, Y-R. (2009). Unsupervised joint prosody labeling and modeling for Mandarin speech. *J. Acoust. Soc. Amer.*, *125*(2), 1164–1183.
42. Chiang, C-Y, Chen, S-H, Wang, Y-R (2009). Advanced unsupervised joint prosody labeling and modeling for mandarin speech and its application to prosody generation for TTS. In *Proc. INTERSPEECH'09*, (pp. 504–507).
43. Chen, S-H, & Wang, Y-R. (1990). Vector quantization of pitch information in Mandarin speech. *IEEE Trans. Commun.*, *38*(9), 1317–1320.
44. Tseng, C-Y, Pin, S-H, Lee, Y-L, Wang, H-M, Chen, Y-C. (2005). Fluent speech prosody: Framework and modeling. *Speech Commun.*, *46*(3–4), 284–309.

45. Breiman, L, Friedman, J, Olshen, R, Stone, C (1984). *Classification and Regression Trees*. Belmont: Wadsworth.
46. Tokuda, K, Zen, H, Black, AW (2004). HMM-based approach to multilingual speech synthesis. In S Narayanan, A Alwan (Eds.), *Text to speech synthesis: New paradigms and advances*. Upper Saddle River: Prentice Hall.
47. Yoshimura, T, Tokuda, K, Masuko, T, Kobayashi, T, Kitamura, T (1999). Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proc. EUROSPEECH'99*, (pp. 2347–2350).
48. T. Yoshimura, Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for HMM-based text-to-speech systems, Ph.D thesis, Nagoya Institute of Technology, 2002.
49. Zen, H, Nose, T, Yamagishi, J, Sako, S, Masuko, T, Black, AW, Tokuda, K (2007). The HMM-based speech synthesis system version 2.0. In *Proc. 6th ISCA Workshop Speech Synth*, (pp. 294–299).
50. Tokuda, K, Masuko, T, Kobayashi, T, Imai, S (1994). Mel-generalized cepstral analysis-a unified approach to speech spectral estimation. In *Proc. ICSLP '94*, (pp. 1043–1046).
51. Yamagishi, J, Kobayashi, T, Nakano, Y, Ogata, K, Isogai, J. (2009). Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. *IEEE Trans. Audio, Speech, Language Process.*, *17*(1), 66–83.
52. Tokuda, K, Yoshimura, T, Masuko, T, Kobayashi, T, Kitamura, T (2000). Speech parameter generation algorithms for HMM-based speech synthesis. In *Proc. ICASSP'00*, (pp. 1315–1318).
53. Imai, S (1983). Cepstral analysis synthesis on the mel frequency scale. In *Proc. of ICASSP'83*, (pp. 93–96).
54. Mandarin microphone speech corpus - TCC300 Corpus [Online] http://www.aclclp.org.tw/use_mat.php#tcc300edu. Accessed 9 July 2018.
55. Chen, F-Y, Tsai, P-F, Chen, K-J, Hunag, C-R. (1999). The construction of Sinica Treebank. *Comput Linguis Chin Lang Process*, *4*(2), 87–104.
56. Chen, K-J, Huang, C-R, Chang, L-P, Hsu, H-L (1996). Sinica Corpus: design methodology for balanced Corpra. In *Proc. PACLIC II*, (pp. 167–176).
57. Young, S, Evermann, G, Kershaw, D, Moore, G, Odell, J, Ollason, D, Povey, D, Valtchev, V, Woodland, PC (2007). *The HTK book (for HTK version 3.4)*. Cambridge: Cambridge University Press.
58. Sjölander, K, & Beskow, J (2000). Wavesurfer—an open source speech tool. In *Proc. ICSLP'00*, (pp. 464–467).
59. Sönmez, M, Heck, L, Weintraub, M, Shriberg, E (1997). A lognormal tied mixture model of pitch for prosody-based speaker recognition. In *Proc. EUROSPEECH'97*, (pp. 1391–1394).
60. ITU-R Rec. BS.1534-2 (2001–2014). Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems. Geneva: International Telecommunications Union.
61. JH Yang, MC Liu, HH Chang, CY Chiang, YR Wang, and SH Chen, "Enriching mandarin speech recognition by incorporating a hierarchical prosody model," in Proc. ICASSP 2011, Prague, Czech, May, 2011, pp 5052–5055.
62. Chiang, C-Y, Yang, J-H, Liu, M-C, Wang, Y-R, Liao, Y-F, Chen, S-H (2011). A new model-based Mandarin-speech coding system. In *Proc. Interspeech 2011, Florence, Italy*, (pp. 2561–2564).