A statistics-based pitch contour model for Mandarin speech

Sin-Horng Chen
Department of Communication Engineering, National Chiao Tung University, Taiwan

Wen-Hsing Lai[a)]
Department of Communication Engineering, National Chiao Tung University, Taiwan and Chunghwa Telecommunication Laboratories, Taiwan

Yih-Ru Wang
Department of Communication Engineering, National Chiao Tung University, Taiwan

(Received 11 April 2004; revised 26 August 2004; accepted 4 November 2004)

A statistics-based syllable pitch contour model for Mandarin speech is proposed. This approach takes the mean and the shape of a syllable log-pitch contour as two basic modeling units and considers several affecting factors that contribute to their variations. The affecting factors include the speaker, prosodic state (which essentially represents the high-level linguistic components of F0 and will be explained more clearly in Sec. I), tone, and *initial* and *final* syllable classes. The parameters of the two modeling units were automatically estimated using the expectation-maximization (EM) algorithm. Experimental results showed that the root mean squared errors (RMSEs) obtained in the closed and open tests in the reconstructed pitch period were 0.362 and 0.373 ms, respectively. This model provides a way to separate the effects of several major factors. All of the inferred values of the affecting factors were in close agreement with our prior linguistic knowledge. It also gives a quantitative and more complete description of the coarticulation effect of neighboring tones rather than conventional qualitative descriptions of the tone *sandhi* rules. In addition, the model can provide useful cues to determine the prosodic phrase boundaries, including those occurring at intersyllable locations, with or without punctuation marks. © *2005 Acoustical Society of America.* [DOI: 10.1121/1.1841572]

PACS numbers: 43.72.Ar [DDO] Pages: 908–925

I. INTRODUCTION

Prosody is an inherent supra-segmental feature of human speech. It carries stress, intonation patterns, and timing structures of continuous speech which, in turn, determine the naturalness and understandability of an utterance. How to automatically generate, analyze, and recognize prosody in speech is one of the unresolved problems confronting researchers who study speech synthesis and recognition. Although it is known that prosody is affected by many factors, such as the phonetic context, sentence type, syntactical structure, semantics, and the emotional status of the speaker, the relationship between these affecting factors and prosody are not totally understood.

Among all the features known to carry prosodic information, pitch is the most important one. It has been reported that the F0 contour characterizes the speaking style and speaker.[1] Therefore, pitch plays a role in many speech related applications, like text-to-speech (TTS),[2–9] tone recognition,[10,11] prosodic labeling,[12,13] emotional state recognition,[14] speaker accent identification,[15] and so on. Adequate pitch control is very important for synthetic speech to be natural in TTS. If a TTS system generates a tone shape matching only the lexical expectation of each individual syllable, the lack of consideration of contextual tone variations will result in underarticulation of tones and lead to the generation of unnatural speech.

Pitch modeling is even more critical for Mandarin speech processing, as Mandarin is a tonal language and the information related to the tonality of a syllable appears, for the most part, on its pitch contour. Although there are only five lexical tones and a previous study[16] has concluded that the pitch contour of each of the first four tones can be simply represented by a single standard pattern, syllable pitch contour patterns in continuous speech vary highly and can deviate dramatically from their canonical forms (i.e., high-level tone, mid-rising tone, low-falling tone, high-falling tone, and low-energy tone). Many factors have been shown to have a major influence on the pitch contour of a tone. They include the effects of neighboring tones, referred to as *sandhi* rules,[17] coarticulation, stress, intonation type, semantics, emotional status, and so on. In addition, the pronunciation of tone 5 is usually highly context dependent and is relatively arbitrary. Thus, pitch modeling is not a trivial research issue for Mandarin speech processing.

Pitch modeling has been the subject of many recent research studies on various languages. The general goal of pitch modeling is to derive a computational model that describes the relationship between a set of affecting factors and pitch contour patterns. The related literature has been concerned with finding perceptual cues and intonational linguistic representations.[18,19] The pitch contour generation rules for synthesizing intelligible and natural-sounding speech,[2–9] and the automatic pitch or tone analysis for the purposes of

[a)] Author to whom correspondence should be addressed. Electronic mail: lwh@cht.com.tw je-4 (0) wei-4 (0) yue-1 (1) han-4 (1) huo-4 (3) pu-3 (7) jin-1 (7) sz-1 (7) da-4 (7) shiue-2 (8) ming-2 (8) yu-4 (8) jiay-4 (8) shou-4 (12) tzai-4 (1) di-4 (1) yi-1 (3) jie-4 (7) guo-2 (5) ji-4 (8) shing-4 (3) gau-1 (8) chau-2 (9) huei-4 (9) yi-4 (10) jung-1 (13) shuo-1 (14) , ta-1 (0) duei-4 (1) je-4 (1) yi-4 (5) shr-3 (4) yu-2 (7) yi-1 (3) jiou-3 (4) ba-1 (7) ling-2 (6) nian-2 (8) dai-4 (8) de-5 (6) shing-4 (5) chiu-1 (11) shr-4 (15) gan-3 (6) dau-4 (6) nan-2 (11) guo-4 (14).

FIG. 1. An example of prosodic states.

speech recognition,[10,11] speech understanding[20] and word finding,[21] have also been studied. Pitch modeling can be performed using two approaches that are rule based[2–4] or data driven.[5–8,22,23] The former approach is conventional; it uses linguistic expertise to manually infer some phonologic rules of pitch contour generation, based on observation of a large set of utterances. A prevalent method in the approach applied to TTS uses sequential rules to initially assign the pitch contour of a segment with an intrinsic value and then successively applies rules to modify it.[2–4] There are three main disadvantages to this approach. First, manually exploring the effect of mutual interaction among several linguistic features at different levels is highly complex. Second, the rule-inference process usually involves a controlled experiment, in which only a limited number of contextual factors is examined. The resulting inferred rules may, therefore, not be general enough for unlimited texts. Third, the rule-inference process is cumbersome. As a result, it is generally very difficult to collect enough rules without expending a great deal of effort.

The data-driven approach tries to construct a pitch model from a large speech corpus, usually by means of statistical methods[6,23] or artificial neural network (ANN) techniques.[5,22] It first designs a computational model to describe the relationship between pitch contour patterns and some affecting factors and then trains the model, using the speech corpus. The training goals are to automatically deduct phonologic rules from the speech corpus and to implicitly incorporate them into the model's parameters or into the ANN's weights. The primary advantage of this approach is that the rules can be automatically established based on the training data set during the training process, without the help of linguistic experts. The recurrent neural network (RNN)-based method[5,22] is a popular method which uses an RNN to learn the mapping between the pitch parameters and some linguistic features. The main criticism raised against this method is the difficulty of interpreting the hidden structures of the model. Other methods include the hidden Markov model (HMM),[23] regression analysis,[6] vector quantization,[7] and the tree-based approach.[8] In addition, an approach that adopts the concept of separating an utterance's pitch contour into a global trend and a locally variational term has been applied in recent pitch modeling studies, e.g., those on superpositional modeling[24,25] and two-stage modeling.[9,26]

In this paper, a new pitch modeling approach for Mandarin speech is proposed. It takes the mean and shape of a syllable log-pitch contour as two basic modeling units and uses statistical methods to model them separately while considering several affecting factors that control their variation. The reason for using parameters of the syllable pitch contour as modeling units lies in the fact that the syllable is the basic pronunciation unit of Mandarin speech and each syllable is lexically marked with a lexical tone, which is a factor that strongly affects pitch in Mandarin. But it is well known that the prosody of an utterance is better modeled with an interval that is much longer than a syllable. Therefore, we use the neighboring tones and the prosodic states to measure the impact. The affecting factors used include the speaker, prosodic state, tone, and initial and final syllable classes. Here, the prosodic state is conceptually defined as the state of a syllable in a prosodic phrase. In continuous speech, speakers tend to group words into phrases whose boundaries are marked by durational and intonational cues. Those phrases are usually referred to as prosodic phrases. Many phonological rules limit their operation within prosodic phrases. While it is generally agreed that the prosodic structure of an utterance has some relationship with its syntactic structure, the two are not isomorphic. In the model, the prosodic state is used as a substitute for high-level linguistic information, like a word, phrase, or syntactic boundaries. Our purpose in using the prosodic state to replace conventional high-level linguistic information is to divide the complicated pitch modeling task into two subtasks. The first one involves modeling the pitch parameters by considering the effects of some low-level linguistic features and the prosodic state. The second subtask involves exploring the relationship between the prosodic state and high-level linguistic cues. Through this two-stage pitch modeling approach, some unsolved problems can be avoided. Problems such as the inconsistency of prosodic and syntactic structures, the ambiguity of word-segmentation and word-chunking for Mandarin Chinese, and the difficulty of performing automatic syntactic analysis on unlimited natural texts can be prevented in the first subtask. In the second subtask, the researcher can focus on modeling the global effect of mapping high-level linguistic features to the prosodic state, since interference caused by low-level linguistic features has already been removed in the first subtask. In this paper, we attack the first subtask only, leaving the second subtask to be dealt with in the future. Due to the fact that the prosodic state of a syllable is not explicitly given, it has been treated as a hidden variable and expectation-maximization (EM) algorithms have been applied to estimate all the parameters of the two pitch models based on a large training set. A by-product of the EM algorithm is the determination of the hidden prosodic states of all the syllables in the training set. This is an additional advantage because prosodic labeling has recently become an interesting research topic.[12] An example is given as Fig. 1. This example shows that the term prosodic state could be made more understandable. Figure 1 shows the phonetic transcription, tone (after dash), and the prosodic states (in parentheses) of each syllable, which are assigned automatically by our model. For each syllable, in our experiment, 1 of 16 prosodic states was assigned. From the sequence of prosodic states, some high-level linguistic phenomenon could be observed, like the possible prosodic phrase boundaries. The prosodic state essentially represents the high-level linguistic components of F0, so the results reported in this paper apply to the prediction of the low-level linguistic component (tone, *initial/final* class, and speaker factors in our model) given the prediction of high-level linguistic components of F0 (the second subtask mentioned above).

This paper is organized as follows. Section II discusses, in detail, the proposed pitch modeling approach for Mandarin speech. Section III presents the experimental results. Detailed analyses of the inferred affecting factors are given in Sec. IV. An application of the proposed syllable pitch contour model to pitch prediction of TTS is given in Sec. V. In the last section, we offer concluding remarks and suggestions for future research.

II. THE PROPOSED PITCH MEAN AND SHAPE MODELS

In the proposed pitch modeling approach, we first perform rough speaker normalization to the pitch period. Our purpose is to adjust the pitch levels and dynamic ranges of all the speakers so that they are approximately the same, in order to improve the efficiency of the subsequent syllable log-pitch contour modeling. In Ref. 27, a Gaussian normalization was used to perform a mapping from the reference pitch values to the desired frequencies, and the authors found that pitch contour moved in the proper direction. We use the same idea to normalize the pitch period of a speaker as follows:

$$f(t)=\frac{f'(t)-\mu_k}{\sigma_k}\cdot\sigma_{all}+\mu_{all}, \tag{1}$$

where $f'(t)$ and $f(t)$ are the original and normalized pitch periods of frame t; μ_k and σ_k are the mean and standard deviation of the pitch period distribution of speaker k; and μ_{all} and σ_{all} are the average mean and average standard deviation of the pitch period distribution of all the training speakers. We then take the logarithm of the normalized pitch period, and the resulting log-pitch contour of each utterance is subsequently divided into a sequence of syllable log-pitch contours. Each syllable log-pitch contour was then decomposed into two parts, the mean and the shape, using a third-order orthogonal polynomial expansion, with the zeroth-order coefficient representing the mean and the other three higher order coefficients representing the shape. We then take the syllable's pitch mean and shape as basic modeling units and employ the two separate statistical models to consider several major affecting factors. Some parts of the pitch modeling approach are discussed in detail in the following.

A. Discrete orthogonal polynomial expansion

Since all syllable log-pitch contours are smooth curves, a third-order orthogonal polynomial expansion is employed to represent them. Actually, in some previous studies,[2,5] orthogonal polynomials, up to the third order, were shown to be good enough to represent Mandarin syllable pitch contours. The four basis polynomials used are normalized, in length, to $[0,1]$ and can be expressed as follows:[5]

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

for $0\leqslant i\leqslant M$, where $M+1$ is the length of the current syllable log-pitch contour and $M\geqslant 3$. They are, in fact, discrete Legendre polynomials. A syllable log-pitch contour, $f(i/M)$, can then be approximated by

$$\hat{f}\left(\frac{i}{M}\right)=\sum_{j=0}^{3}\alpha_j\cdot\phi_j\left(\frac{i}{M}\right),\quad 0\leqslant i\leqslant M, \tag{3}$$

where

$$\alpha_j=\frac{1}{M+1}\sum_{i=0}^{M}f\left(\frac{i}{M}\right)\cdot\phi_j\left(\frac{i}{M}\right). \tag{4}$$

B. Affecting factors

In naturally spoken Mandarin Chinese, pitch varies considerably, depending on various linguistic/nonlinguistic factors. In this study, we considered some factors that may have major effects on control of the variation of the pitch contour. The specific affecting factors chosen for the pitch mean and shape models are discussed in the following.

1. Affecting factors for the pitch mean model

The pitch mean is mainly affected by intonation, while the pitch shape is affected mainly by lexical tones. A brief summary of the major factors affecting intonation contours was given in Ref. 28. They include declination, downstep, final lowering, accents and tones, segmental effects, and intonation type. In our pitch mean model, the affecting factors considered include the tones of the previous, current, and following syllables; the *initial* and *final* classes of the current syllable; the prosodic state of the current syllable; and the speaker's level shift and dynamic range scaling factors. Their influence on the syllable pitch mean is discussed below.

Mandarin Chinese is a tonal and syllable-based language. The syllable is the basic pronunciation unit. Each character is pronounced as a syllable. Only about 1300 phonetically distinguishable syllables, comprising the set of all legal combinations of 411 base-syllables and five tones, exist. The tonality of a syllable is mainly characterized by its pitch contour, loudness, and duration. We, therefore, consider the tone of the current syllable as an affecting factor. Coarticulations from the neighboring tones, which are known as sandhi rules, also exist. Thus, the tones of the previous and following syllables are also chosen as affecting factors.

Mandarin base-syllables have a very regular phonetic structure. Each base-syllable is composed of an optional *consonant initial* and a *final*. The *final* can be further broken down into an optional *medial*, a *vowel nucleus*, and an optional *nasal ending*. As discussed in Refs. 28 and 29, many types of observed F0 movement are caused by these segmental effects. We, therefore, consider the broad *initial* and *final* classes of the current syllable as affecting factors and investigate their effects on pitch mean variation.

Aside from the linguistic factors mentioned above, other high-level linguistic components, such as word-level and syntactic-level features, can also seriously affect the pitch contour of an utterance. As discussed in Sec. I, the prosodic state is in our approach used to account for the influence of all high-level linguistic features. Here, the prosodic state simply means the state of the syllable in a prosodic phrase. The pitch level of a syllable can vary drastically in different parts of a prosodic phrase. The declination effect of the global downtrend, referring to the tendency of F0 to decline over the course of an utterance, is a well-known example. There are two advantages of using the prosodic state to replace high-level linguistic features. First, pitch information is a kind of prosodic feature, so the variation of the syllable pitch contour should better match the prosodic phrase structure than the syntactic phrase structure. Second, as mentioned above, some unsolved problems, such as the ambiguity of word-segmentation and word-chunking in Mandarin Chinese and the difficulty of performing automatic syntactic analysis on unlimited natural texts, can be avoided in the current pitch modeling approach. This prevents us from using improper or incomplete high-level linguistic information. The main problem with using the prosodic state is the lack of large speech corpora with prosodic tags that have been properly labeled. Thus, we have to treat the prosodic state of a syllable as a hidden or unknown variable. Fortunately, we are able to solve this problem by using the expectation-maximization (EM) algorithm, which is a technique of maximum likelihood (ML) estimation from incomplete data. A by-product of the approach is the automatic determination of prosodic states for all the syllables in the training set. This is an additional advantage because prosodic labeling has recently become an interesting research topic.[12,13] In addition, such prosodic phrasal information provides clues for resolving syntactic ambiguity in automatic speech understanding[20,21,30,31] and for improving the naturalness of TTS.[32,33]

Lastly, the pitch contour of an utterance is also significantly affected by the speaker. Speakers have different pitch levels and dynamic ranges. Rough speaker normalization is performed in the preprocessing stage in order to suppress the speaker effect and allow the pitch period distributions of all the speakers to have the same mean and standard deviation. However, we also use two speaker affecting factors in the pitch mean model to examine whether the Gaussian-normalized syllable log-pitch contour is still speaker dependent.

2. Affecting factors for the pitch shape model

Pitch shapes are relatively tone determined. Production studies of Chinese tones have shown that tone shapes in natural continuous speech often deviate from their canonical shapes. They suffer from large deformations due to tone coarticulation, also known as tone *sandhi*. This situation is particularly common in conversation, where the boundaries among tonal categories are blurred. It has been suggested in Ref. 17 that *sandhi* contour patterns of poly-tonal groups are rather invariant and can be treated as the basic units of pitch contour analysis/generation. Therefore, lexical tone combinations are used here to consider the effect of tone coarticulation. To give further consideration to the coupling/noncoupling effect of neighboring syllables, we considered one-, two-, and three-syllable tone combinations as affecting factors in the pitch shape model.

Other affecting factors chosen for the pitch shape model includ the *initial* and *final* classes of the current syllable for the segmental effect, the prosodic state of the current syllable for the effects of high-level linguistic features, and the pitch level shifting effect of speakers.

C. The pitch models

In pitch modeling, we take the mean and shape of the syllable log-pitch contour as basic modeling units and use two separate models to exploit their variations. Because the complicated high-level linguistic components of F0 are represented by prosodic states, only acoustic factors are considered. Therefore, simple additive models are adopted in our study. They are discussed in detail in the following.

1. The pitch mean model

The pitch mean model was constructed by first considering the two affecting factors of the speaker, expressed as

$$Z_n=(Y_n+\beta_{s_n})\gamma_{s_n}, \tag{5}$$

where Z_n is the observed mean (i.e., the zeroth-order coefficient α_0 of the orthogonal polynomial transform) of the log-pitch contour of the nth (current) syllable; β_{s_n} and γ_{s_n} are the companding (compressing-expanding) factors (CFs) of the two affecting factors of the speaker, representing, respectively, the effects of level shift and dynamic range scaling on Z_n; and Y_n is the speaker effect-compensated pitch mean. Here, CF means the effect of a factor on the expansion (increase) or compression (reduction) of the pitch mean. The model goes on to further consider other affecting factors, expressed as

$$Y_n=X_n+\beta_{t_n}+\beta_{pt_n}+\beta_{ft_n}+\beta_{i_n}+\beta_{f_n}+\beta_{p_n}, \tag{6}$$

where X_n is the normalized pitch mean of the nth syllable and is modeled as a normal distribution with mean μ and variance ν; β_r is the CF for affecting factor r; t_n, pt_n, and ft_n represent the lexical tones of the current, previous, and following syllables, respectively; i_n and f_n are broad *initial* and *final* classes of the current syllable; and p_n represents the

TABLE I. (a) CFs of the affecting factors in the pitch mean model. (b) CF vectors of the affecting factors in the pitch shape model.

(a)	
γ_{s_n}	CF of the dynamic range scaling of the speakers
β_{s_n}	CF of the level shift of speakers
β_{t_n}	CF of the current lexical tone
β_{pt_n}	CF of the previous lexical tone
β_{ft_n}	CF of the following lexical tone
β_{i_n}	CF of the initial class
β_{f_n}	CF of the final class
β_{p_n}	CF of the pitch-mean prosodic state

(b)	
$\mathbf{b}_{s_n}$	CF vector of the speakers
$\mathbf{b}_{tc_n}$	CF vector of the lexical tone combination of the current syllable and its two neighbors
$\mathbf{b}_{i_n}$	CF vector of the initial class
$\mathbf{b}_{f_n}$	CF vector of the final class
$\mathbf{b}_{q_n}$	CF vector of the pitch-shape prosodic state

prosodic state of the current syllable. Note that t_n ranges from 1 to 5, while both pt_n and ft_n range from 0 to 5 with 0 denoting cases with punctuation marks or the nonexistence of a preceding or succeeding syllable. The affecting factors for $pt_n=0$ and $ft_n=0$ are simply set to zero because we do not want to include the effect of tone across punctuation marks. All the affecting factors in the pitch mean model and their notations are summarized in Table I(a).

2. The pitch shape model

The pitch shape model is expressed as

$$\mathbf{Z}_n=\mathbf{X}_n+\mathbf{b}_{tc_n}+\mathbf{b}_{q_n}+\mathbf{b}_{s_n}+\mathbf{b}_{i_n}+\mathbf{b}_{f_n}, \tag{7}$$

where $\mathbf{Z}_n$ is the observed pitch shape vector $[\alpha_1\ \alpha_2\ \alpha_3]^T$ for the nth syllable; $\mathbf{X}_n$ is the normalized pitch shape vector of the nth syllable and is modeled as a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{R}$; $\mathbf{b}_r$ is the CF vector for affecting factor r; tc_n represents a lexical tone combination of the current syllable and its two nearest neighbors; and q_n represents the pitch-shape prosodic state of the current syllable. Here, a lexical tone combination, instead of individual tones, is used because we want to consider the aggregative influence of the current tone and its two nearest neighboring tones. The invoking of the preceding and succeeding tones in the tone combination depends on whether or not long intersyllable pauses exist before and/or after the current syllable, respectively. In a case where both the pre- and postpauses of the current syllable are not long, we consider the effects of both the preceding and succeeding tones, and use a tri-tone combination. When the prepause and/or the postpause are equal to or longer than a predetermined threshold (=13 frames or 65 ms in this study), we ignore the influence of the preceding and/or succeeding syllables, and use a single-tone/bi-tone combination. All the affecting factors in the pitch shape model and their notations are summarized in Table I(b).

D. Training the pitch models

1. Training the pitch mean model

To estimate the parameters of the pitch mean model, an EM algorithm is adopted. The derivation of the EM algorithm is based on treating the prosodic state as an unknown variable. An auxiliary function is first defined in the expectation step (E-step) as follows:

$$Q(\bar{\lambda},\lambda)=\sum_{n=1}^{N}\sum_{p_n=1}^{P} p(p_n|Z_n,\bar{\lambda})\log p(Z_n,p_n|\lambda), \tag{8}$$

where N is the total number of training samples, P is the total number of prosodic states, $p(p_n|Z_n,\lambda)$ and $p(Z_n,p_n|\lambda)$ are conditional probabilities, $\lambda=\{\mu,\nu,\beta_t,\beta_{pt},\beta_{ft},\beta_i,\beta_f,\beta_p,\beta_s,\gamma_s\}$ is the set of parameters to be estimated, and λ and λ are the new and old parameter sets, respectively. Based on the assumption that the normalized pitch mean X_n is normally distributed, $p(Z_n,p_n|\lambda)$ can be derived from the assumed model given in Eqs. (5) and (6) and expressed as

$$p(Z_n,p_n|\lambda)=N(Z_n;(\mu+\beta_{t_n}+\beta_{pt_n}+\beta_{ft_n}+\beta_{i_n}+\beta_{f_n}+\beta_{p_n}+\beta_{s_n})\gamma_{s_n},\nu\gamma_{s_n}^2), \tag{9}$$

where $N(Z;a,b)$ denotes a normal distribution of Z with mean a and variance b. Similarly, $p(p_n|Z_n,\bar{\lambda})$ can be expressed as

$$p(p_n|Z_n,\bar{\lambda})=\frac{p(Z_n,p_n|\bar{\lambda})}{\sum_{p_n'=1}^{P}p(Z_n,p_n'|\bar{\lambda})}. \tag{10}$$

Then, sequential optimizations of these parameters can be performed in the maximization step (M-step).

A drawback of the above EM algorithm is that it may produce a nonunique solution. To solve this problem, we modify each optimization procedure in the M-step to constrained optimization by introducing a global constraint. The auxiliary function is then changed to

$$Q(\bar{\lambda},\lambda)=\sum_{n=1}^{N}\sum_{p_n=1}^{P} p(p_n|Z_n,\bar{\lambda})\log p(Z_n,p_n|\lambda) + \eta\left(\sum_{n=1}^{N}(\mu+\beta_{t_n}+\beta_{pt_n}+\beta_{ft_n}+\beta_{i_n}+\beta_{f_n}+\beta_{p_n}+\beta_{s_n})\gamma_{s_n}-N\mu_Z\right), \tag{11}$$

where μ_Z is the average of Z_n and η is a Lagrange multiplier. The constrained optimization is finally solved via the Newton–Raphson method.

To execute the EM algorithm, initializations of the parameter set λ are needed. This is done by estimating each individual parameter independently. Specifically, the initial multiplicative/additive CF for a specific value of an affecting factor is assigned to be the ratio/difference of the mean of Z_n with the affecting factor equaling the value to the mean of all Z_n. Notice that, in the initialization of the CFs for the affecting factors of the prosodic states, each syllable is preassigned a prosodic state by means of vector quantization. Following, all the parameters are sequentially updated in each iteration. The iterative procedure is continued until convergence is reached. The prosodic state can, finally, be assigned as

$$p_n^* = \arg\max_{p_n} p(p_n|Z_n,\lambda). \tag{12}$$

The EM algorithm is summarized below:

(1) Compute the initial values of λ by independently estimating each individual parameter from the training set.

(2) Do this for each iteration k:

(a) Update $\overline{\lambda} = \lambda$.

(b) E-step: Use Eqs. (9)–(11) to calculate $Q(\overline{\lambda},\lambda)$.

(c) M-step: Find the optimal λ as follows:

$$\lambda = \arg\max_{\lambda} Q(\overline{\lambda},\lambda). \tag{13}$$

(d) Termination test: If $L(k)-L(k-1)<\varepsilon$ or $k \geqslant K$, then stop, where

$$L(k) = \sum_{n=1}^{N} \log p(Z_n|\lambda) \tag{14}$$

is the total log-likelihood for iteration k and K is the maximum number of iterations.

(3) Assign prosodic states to all the syllables using Eq. (12).

2. Training the pitch shape model

The pitch shape model is trained using the same EM algorithm. An auxiliary function with a global constraint was first defined as follows:

$$\mathbf{Q}(\overline{\boldsymbol{\lambda}},\boldsymbol{\lambda}) = \sum_{n=1}^{N} \sum_{q_n=1}^{P} p(q_n|\mathbf{Z}_n,\overline{\boldsymbol{\lambda}}) \log p(\mathbf{Z}_n,q_n|\boldsymbol{\lambda}) + \mathbf{L}^T \times \left(\sum_{n=1}^{N} (\boldsymbol{\mu} + \mathbf{b}_{tc_n} + \mathbf{b}_{i_n} + \mathbf{b}_{f_n} + \mathbf{b}_{q_n} + \mathbf{b}_{s_n}) - N\boldsymbol{\mu}_Z \right), \tag{15}$$

where $\mathbf{L}$ is a 3×1 Lagrange multiplier vector and $\boldsymbol{\lambda} = \{\boldsymbol{\mu},\mathbf{R},\mathbf{b}_{tc},\mathbf{b}_i,\mathbf{b}_f,\mathbf{b}_q,\mathbf{b}_s\}$ is the set of parameter vectors to be estimated. Based on the assumption that the normalized pitch shape vector $\mathbf{X}_n$ is normally distributed, $p(\mathbf{Z}_n,q_n|\boldsymbol{\lambda})$ can be expressed as

$$p(\mathbf{Z}_n,q_n|\boldsymbol{\lambda}) = \mathrm{MVN}(\mathbf{Z}_n;\boldsymbol{\mu} + \mathbf{b}_{tc_n} + \mathbf{b}_{i_n} + \mathbf{b}_{f_n} + \mathbf{b}_{q_n} + \mathbf{b}_{s_n},\mathbf{R}), \tag{16}$$

where $\mathrm{MVN}(\mathbf{Z};\mathbf{a},\mathbf{B})$ denotes a multivariate normal distribution of $\mathbf{Z}$ with mean vector $\mathbf{a}$ and covariance matrix $\mathbf{B}$. By maximizing the auxiliary function, we can get the optimal parameter set. The training procedure is similar to that for the pitch mean model.

E. Testing the pitch models

1. Testing the pitch mean model

Although we obtain CFs for all affecting factors through the above training procedure, some information still must be discovered in the testing phase. This includes the CFs of the two speaker-affecting factors and the prosodic state of each syllable. The following testing procedure is used to estimate these unknown parameters:

(1) Initialization:

(a) Freeze the CFs for the current, previous, and following tones, for the *initial* and *final* classes, and for the prosodic state, the mean, and variance of the normalized pitch mean to their trained values, and form a parameter set $\overline{\lambda}_1 = \{\overline{\mu},\overline{\nu},\overline{\beta}_t,\overline{\beta}_{pt},\overline{\beta}_{ft},\overline{\beta}_i,\overline{\beta}_f,\overline{\beta}_p\}$.

(b) Compute the initial CFs for the parameter set $\lambda_2 = \{\beta_s,\gamma_s\}$.

(2) Do this for each iteration k:

(a) Update $\overline{\lambda}_2 = \lambda_2$.

(b) E-step: Calculate

$$Q(\overline{\lambda}_2,\lambda_2) = \sum_{n=1}^{N} \sum_{p_n=1}^{P} p(p_n|Z_n,\overline{\lambda}_1,\overline{\lambda}_2) \times \log p(Z_n,p_n|\overline{\lambda}_1,\lambda_2). \tag{17}$$

(c) M-step: Find the optimal λ_2 via

$$\lambda_2 = \arg\max_{\lambda_2} Q(\overline{\lambda}_2,\lambda_2). \tag{18}$$

(d) Termination test: If $L(k)-L(k-1)<\varepsilon$ or $k \geqslant K$, then stop, where

$$L(k) = \sum_{n=1}^{N} \log p(Z_n|\overline{\lambda}_1,\lambda_2) \tag{19}$$

is the total log-likelihood for iteration k.

(3) Assign prosodic state by means of

$$p_n^* = \arg\max_{p_n} p(p_n|Z_n,\overline{\lambda}_1,\lambda_2). \tag{20}$$

After performing the above procedure, we can derive the two speaker CFs for each testing speaker and determine the prosodic state of each syllable.

2. Testing the pitch shape model

In the testing phase, a similar procedure is employed to estimate the unknown parameters of the pitch shape model from the testing data set, with all the known parameters being fixed. Here, the unknown parameters are the CF vector of the speaker affecting factor and the prosodic state of each syllable. In this case, the fixed parameter set $\overline{\boldsymbol{\lambda}}_1 = \{\overline{\boldsymbol{\mu}},\overline{\mathbf{R}},\overline{\mathbf{b}}_{tc},\overline{\mathbf{b}}_i,\overline{\mathbf{b}}_f,\overline{\mathbf{b}}_q\}$ and the unknown parameter set $\boldsymbol{\lambda}_2 = \{\mathbf{b}_s\}$ are used in the testing procedure.

III. EXPERIMENTAL RESULTS

A. Databases

The effectiveness of the proposed syllable pitch modeling method was examined through simulations on two databases. The first database was a high-quality, reading-style, microphone speech database, which was recorded in a sound-proof booth. It is referred to as the TL database. It was generated by five native Chinese speakers, including two males and three females; among these five, two were profes-

TABLE II. TL database statistics.

Data Set	Speaker	Sentence	Paragraph	Syllable
Training	Male A	1-455	1–200	34 670
Training	Female B	1-455	1–50	12 945
Training	Male C	1-455	1–100	20 748
Training	Female D	1-455	1–200	34 166
Testing	Female E	None	201–300	22 109

sional radio announcers. The database consisted of two types of data. The first type of data comprised sentential utterances with texts belonging to a well-designed, phonetic-balanced corpus of 455 sentences. The lengths of these sentences ranged from 3 to 75 syllables with an average of 13 syllables. The other types of data were longer utterances with texts belonging to a corpus of 300 paragraphs, which covered a wide range of topics, including news, primary school textbooks, literature, essays, etc. The lengths of these paragraphs ranged from 24 to 529 syllables with an average length of 170 syllables. The database was divided into two parts: a training set and a test set. Table II shows the database statistics. The training set contained, in total, 102 529 syllables, and the test set contained 22 109 syllables. The speakers and text content in the test set were different from those in the training set.

After recording was completed, all speech signals in the database were converted into 16-bit data at a 20-kHz sampling rate. They were then manually segmented into *initial* and *final* subsyllables. The phonetic transcription was generated automatically by a linguistic processor, with an 80 000-word lexicon. All the transcription errors were manually corrected. The pitch period was then automatically detected by the ESPS software, with large errors being detected by the program and corrections made by hand. A four-step preprocessing procedure was then applied to extract the four modeling parameters. The four steps included frame-based speaker normalization, frame-based logarithm operation, dividing the utterances' log-pitch contours into syllable segments, and performing orthogonal expansion of syllable log-pitch contours. The statistics for the observed mean and shape of the syllable log-pitch contour can be found in Table III(a).

TABLE IV. The mean and (co)variance statistics of (a) the observed and (b) the normalized mean and shape of the syllable log-pitch contour with 16 prosodic states for the TCC database (unit of pitch period: ms).

	mean	(co)variance	RMSE
(a)			
Pitch mean α_0	1.840	0.0209	
Pitch shape $[\alpha_1\ \alpha_2\ \alpha_3]^T$ (×100)	$\begin{bmatrix}2.797\\-0.593\\-0.018\end{bmatrix}$	$\begin{bmatrix}32.392 & 0.341 & -2.680\\0.341 & 9.740 & -0.199\\-2.680 & -0.199 & 3.289\end{bmatrix}$	
(b)			
Pitch mean α_0	1.842	0.000 739	0.0275
Pitch shape $[\alpha_1\ \alpha_2\ \alpha_3]^T$ (×100)	$\begin{bmatrix}2.810\\-0.577\\-0.020\end{bmatrix}$	$\begin{bmatrix}7.037 & -0.561 & -1.791\\-0.561 & 3.657 & -0.403\\-1.791 & -0.403 & 2.165\end{bmatrix}$	$\begin{bmatrix}2.653\\1.912\\1.471\end{bmatrix}$

The second database was a 100-speaker, microphone speech data set, which was a subset of TCC-300, provided by the Association of Computational Linguistics and Chinese Language Processing. It is referred to as the TCC database. The database was generated by 50 males and 50 females. Each speaker uttered several paragraphs of differing content. The speech data were all directly, digitally recorded in a laboratory in 16-kHz, 16-bit linear PCM. The total number of syllables in the database was 141 991. After recording was completed, all the speech signals were automatically segmented, using 100-*initial* and 39-*final* HMM models. Then, the pitch period was automatically detected by WaveSurfer software, the large errors being excluded by the program. The same four-step preprocessing procedure was then applied to extract the four modeling parameters. In Table IV(a), the statistics for the observed mean and shape of the syllable log-pitch contour are shown.

TABLE III. The mean and (co)variance statistics of (a) the observed and (b) the normalized mean and shape of the syllable log-pitch contour with 16 prosodic states for the TL database (unit of pitch period: ms).

(a)

	Training set		Test set	
	Mean	(Co)variance	Mean	(Co)variance
Pitch mean α_0	1.949	0.0372	1.948	0.0345
Pitch shape $[\alpha_1\ \alpha_2\ \alpha_3]^T$ (×100)	$\begin{bmatrix}3.545\\-0.982\\-0.056\end{bmatrix}$	$\begin{bmatrix}58.550 & 3.229 & -5.140\\3.229 & 9.671 & -0.106\\-5.140 & -0.106 & 2.900\end{bmatrix}$	$\begin{bmatrix}4.012\\-0.749\\-0.142\end{bmatrix}$	$\begin{bmatrix}49.489 & 3.653 & -4.007\\3.653 & 12.460 & 0.276\\-4.007 & 0.276 & 4.356\end{bmatrix}$

(b)

	Training set			Test set		
	Mean	(Co)variance	RMSE	Mean	(Co)variance	RMSE
Pitch mean α_0	1.948	0.000 402	0.0203	1.948	0.000 344	0.0183
Pitch shape $[\alpha_1\ \alpha_2\ \alpha_3]^T$ (×100)	$\begin{bmatrix}3.660\\-0.996\\-0.104\end{bmatrix}$	$\begin{bmatrix}9.865 & -0.354 & -0.076\\-0.354 & 1.907 & 0.232\\-0.076 & 0.232 & 1.251\end{bmatrix}$	$\begin{bmatrix}3.143\\1.381\\1.120\end{bmatrix}$	$\begin{bmatrix}3.861\\-0.906\\-0.085\end{bmatrix}$	$\begin{bmatrix}12.885 & 0.955 & 1.073\\0.955 & 3.101 & 0.808\\1.073 & 0.808 & 2.263\end{bmatrix}$	$\begin{bmatrix}3.603\\1.762\\1.505\end{bmatrix}$

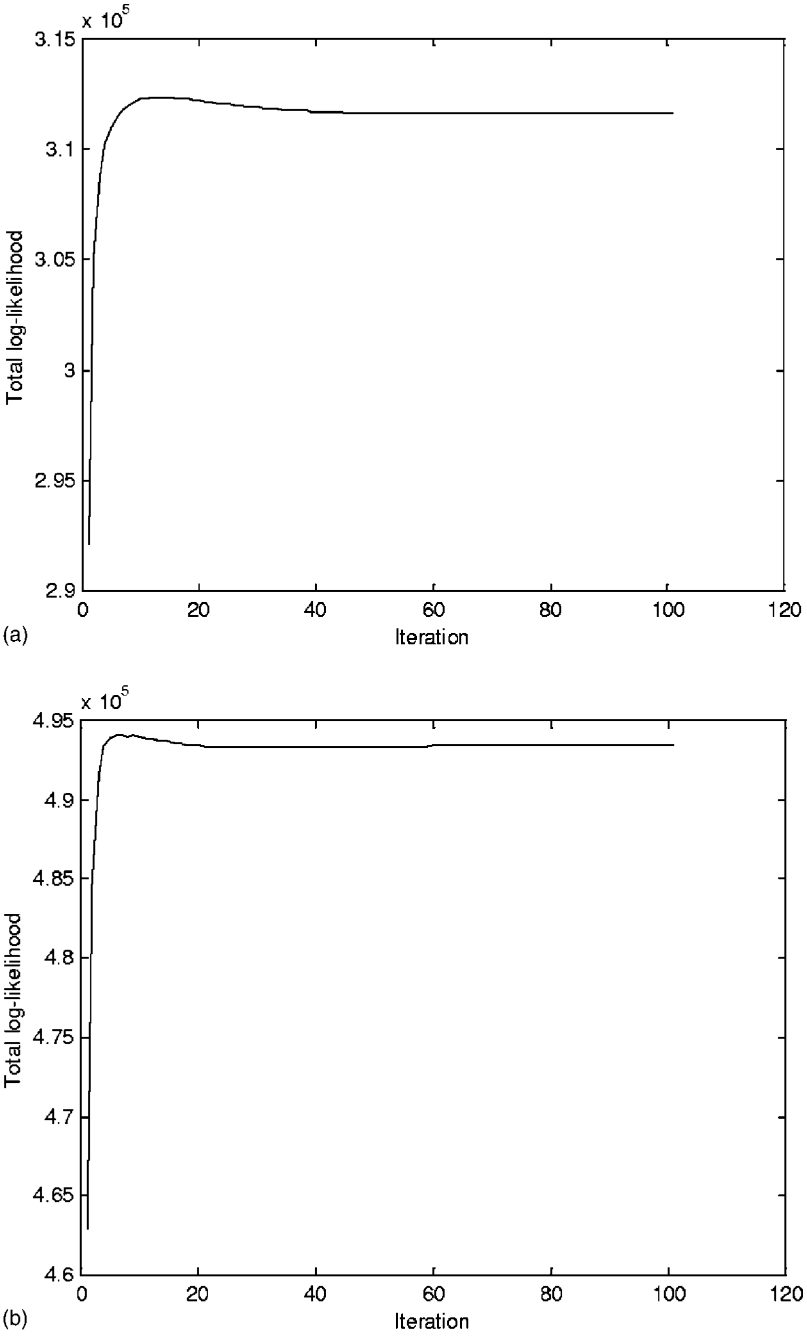


FIG. 2. The plot of the total log-likelihood versus the iteration number for the training of the pitch mean model of (a) the TL database and (b) the TCC database.

B. Experimental results of pitch modeling

The effect of the proposed pitch modeling method was examined first, with the number of prosodic states set to 16. Table III(b) shows the experimental results of pitch mean and shape modeling. It can be seen from the third and sixth columns of Table III(b) that the (co)variances of the normalized mean and shape of the syllable log-pitch contour were greatly reduced for both the closed and open tests, when compared with those shown in Table III(a). The RMSEs of the reconstructed mean and shape of the syllable log-pitch contour are shown in the fourth and seventh columns of Table III(b). Here, the reconstructed mean (shape) was calculated based on the well-trained pitch mean (shape) model by assigning the most probable prosodic state to each syllable and setting the normalized mean (shape) parameter(s) to its (their) mean value(s). By combining the results of the reconstructed pitch mean and shape, we could reconstruct the pitch contour for each syllable. The RMSEs of the reconstructed pitch contour were 0.362 and 0.373 ms/frame for the closed and open tests, respectively. Notice that these two values included RMSEs of 0.17 and 0.19 ms/frame, which resulted from applying orthogonal transformation to the closed and open test data sets, respectively.

Figure 2(a) shows a plot of the total log-likelihood $L(k)$ versus the iteration number k. It can be seen from Fig. 2(a) that the EM algorithm quickly converged in the first several iterations. The histograms of the observed and normalized syllable log-pitch mean for the training set are plotted in Figs. 3(a) and 3(b). It can be seen from these two figures that the variation of the syllable log-pitch mean was greatly reduced after the influence of the affecting factors considered in the model was eliminated. Based on the above results, we

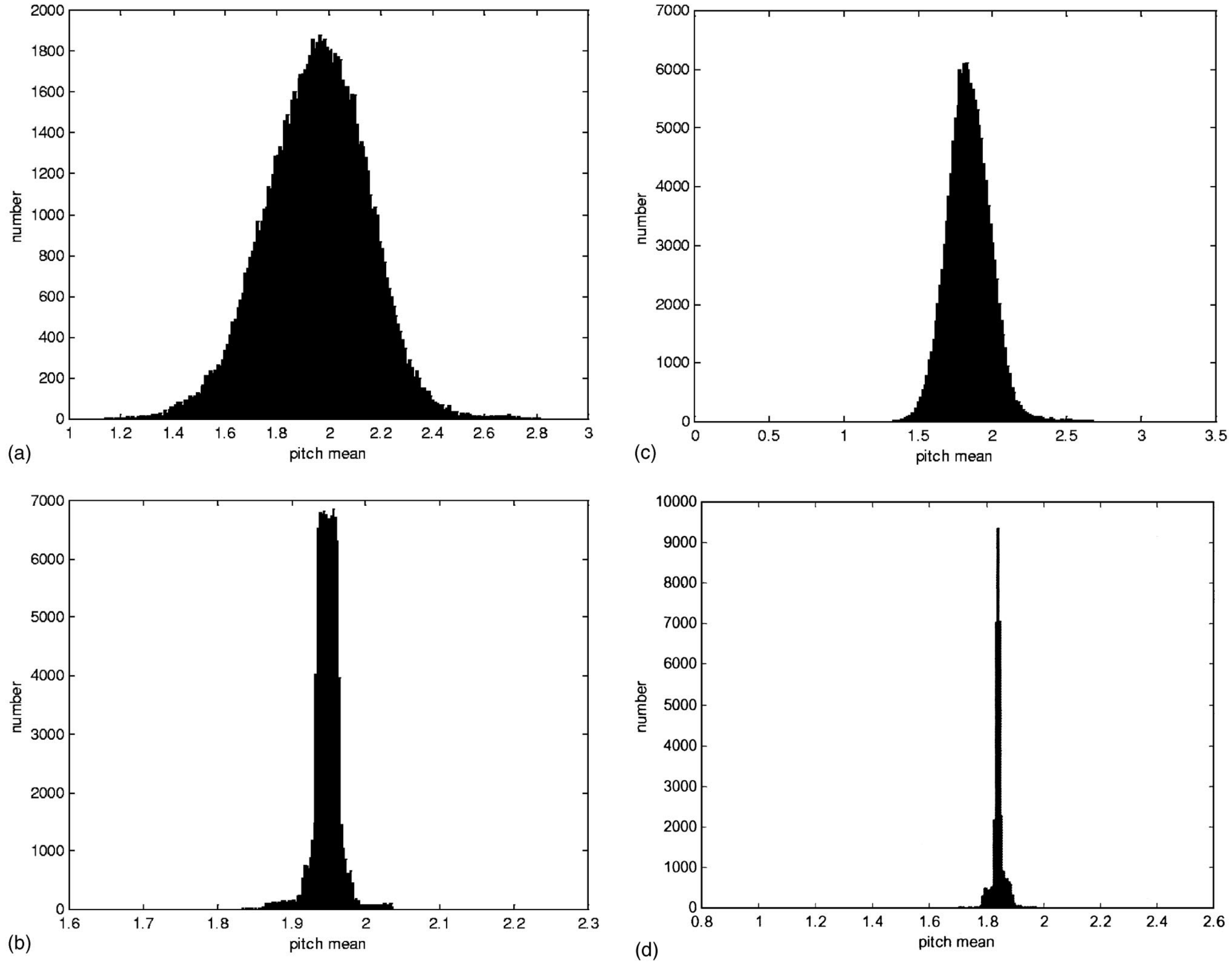


FIG. 3. The histograms of (a) the observed and (b) the normalized pitch means for the training set of the TL database and the histograms of (c) the observed and (d) the normalized pitch means for the TCC database.

concluded that the proposed pitch mean modeling method was effective.

We then examined a case in which the number of prosodic states changed. The resulting variance of the normalized syllable log-pitch mean is shown in Fig. 4(a). As can be seen, the variance of the normalized pitch mean decreased as the number of prosodic states increased. This implies that the pitch mean model became more accurate as the number of prosodic states increased. The improvement reached saturation when the number of prosodic states was greater than 16. Similar findings were observed for the corresponding RMSEs of the reconstructed pitch mean shown in Fig. 4(b).

Figure 5 shows two typical examples of reconstructed pitch contours of two utterances based on the pitch mean and shape models with 16 prosodic states. It can be seen from these two figures that all the reconstructed syllable pitch contours closely resembled their original counterparts. Actually, they were the smoothed versions of the originals as three-order orthogonal polynomial transformation was used. Further evaluation of the performance of the reconstructed pitch contours was conducted by means of two subjective tests: the AB test and the mean opinion score (MOS) test. The synthesized speech recordings, with both the original pitch contours and the reconstructed pitch contours, were presented to the listeners involved in the tests. The inside/outside test could show whether test sentences are from the training or testing set. In the inside test, the original and reconstructed pitch contours of utterances of speaker A (see Table II), a male professional announcer, were used, while in the outside test, the utterances of speaker E, a female speaker, were used. All the testing utterances were generated by the PSOLA algorithm using two acoustic inventories containing the waveform templates of 414 monosyllables. These two acoustic inventories were generated by speakers A and B for the inside and outside tests, respectively. It should be noted that the acoustic inventory of speaker B, who is a professional female announcer, was used in the outside test because the acoustic inventory of speaker E was lacking. All the other prosodic parameters, including the syllable duration, syllable log-energy level, and intersyllabic pause duration, were estimated from the training database using a regression model. Five different long test sentences were used in both the inside and outside tests. Combined with the two kinds of synthesized speech, there were, in total, 20 test sample utterances. Sixteen listeners, university students, were involved in the two tests. In the AB test, each listener was given a pair of

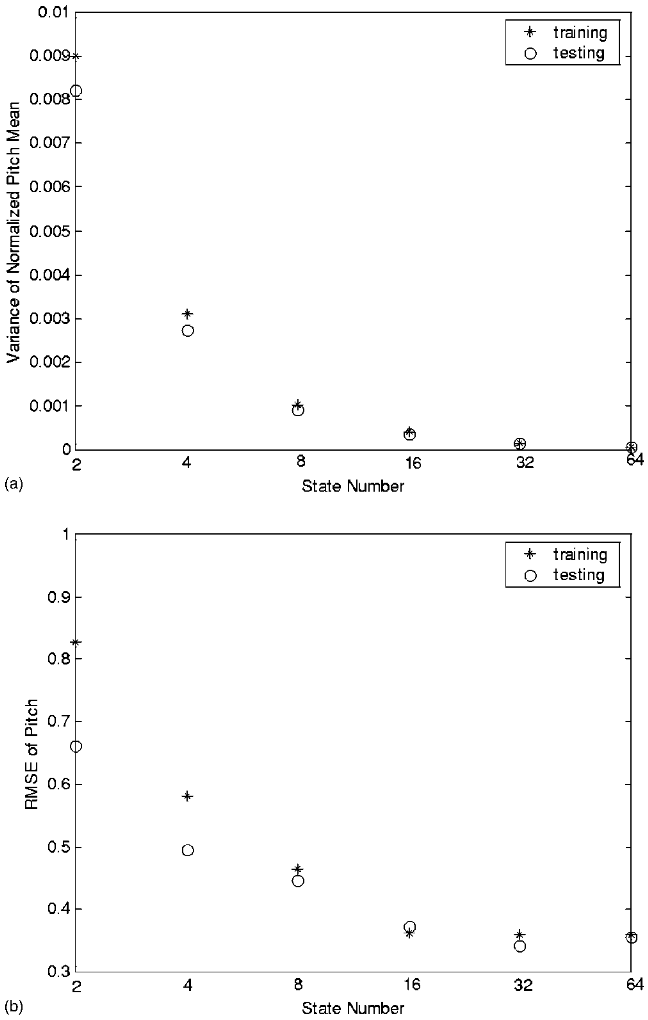


FIG. 4. Plots of (a) the variance of the normalized pitch mean versus the number of prosodic states, and (b) the RMSE of the reconstructed pitch mean versus the number of prosodic states.

synthesized utterances, along with the original and reconstructed pitch contours for each testing sentence, and asked to vote for the better one. Experimental results showed that 41.25% (22.5%) of the synthesized speech recordings, with the original pitch contours, were found by the listeners to sound better; 25% (27.5%) of the synthesized speech recordings, with the reconstructed pitch contours, were found to sound better; and 33.75% (50%) of the two speech recordings were found to sound equivalent for the inside (outside) test, respectively. In the MOS test, absolute category rating was conducted on a scale from 1 (‘‘bad’’) to 5 (‘‘excellent’’). Experimental results showed that average MOSs of 3.94 (3.34) and 3.68 (3.4) were obtained for the synthesized speech recordings with the original and reconstructed pitch contours, respectively, in the inside (outside) test. From the results of these two subjective tests, we concluded that the reconstructed pitch contours functioned almost as well as their original counterparts.

We then checked whether it was necessary to include the speaker affecting factors in both pitch mean and shape models, besides frame-based speaker normalization, which was performed in the preprocessing stage. An experiment, which excluded the two speaker affecting factors used in the pitch mean model and the speaker affecting factor used in the pitch shape model, was conducted. RMSEs of 0.362 and 0.372 ms were obtained in the closed and open tests, respectively. These results were almost the same as those for the previous cases, which used these three speaker affecting factors in the pitch mean and shape models. This showed that rough speaker normalization was good enough to eliminate the speaker’s influence.

We then checked whether the pitch mean and shape models could share the same set of prosodic states. An experiment, in which the prosodic state of every syllable in the pitch shape model was forced to be the same as that in the pitch mean model, was then conducted. RMSEs of 0.504 and 0.478 ms were obtained in the closed and open tests, respectively. These results were worse than those obtained using separate sets of prosodic states in the pitch mean and shape models. Figure 6 shows the 16 patterns of unified prosodic states. The patterns are plotted from left to right in increasing order of the prosodic state index. The vertical axis is pitch period (ms). Sixteen syllable pitch contour patterns were formed using the CFs of the prosodic states, and the average values of the normalized syllable log-pitch mean and shape can be found in this figure. It can also be found in Fig. 6 that the lower-indexed states had a lower pitch mean and smaller pitch slope; they represented the beginning part of a prosodic phrase. On the other hand, the higher-indexed states had a higher pitch mean and larger pitch slope; they represented the ending part of a prosodic phrase.

Finally, we examined the effectiveness of pitch modeling via the TCC database. The same training procedure used with the TL database was applied. The number of prosodic states was set to 16. Table IV(b) shows the experimental results obtained for the mean and (co)variance of the normalized pitch mean and shape, and the RMSEs of the reconstructed pitch mean and shape. It can be seen from the third column in Tables IV(a) and (b) that the variance of the normalized pitch mean and the covariance of the normalized pitch shape were greatly reduced, when compared with those of the original pitch mean and shape. By combining the results for the reconstructed pitch mean and shape, we could reconstruct the pitch contour of each syllable. The RMSEs of the reconstructed pitch contours were 0.384 ms/frame, which included the RMSEs of 0.172 ms that resulted from applying orthogonal transformation. A plot of the total log-likelihood $L(k)$ versus the iteration number k is shown in Fig. 2(b). The histograms of the observed and normalized syllable log-pitch mean for TCC are plotted in Figs. 3(c) and (d). The results were still quite promising even though the pitch variation, due to the large population of speakers, was very high, and the accuracy of the observed data, due to the automatic segmentation performed by the HMM models, was not as high as that achieved by applying manual segmentation to the TL database.

IV. ANALYSES OF THE INFERRED MODEL PARAMETERS

We then analyzed, in detail, the inferred model parameters in order to gain a better understanding of the effects of the affecting factors. Before discussing this, we will briefly

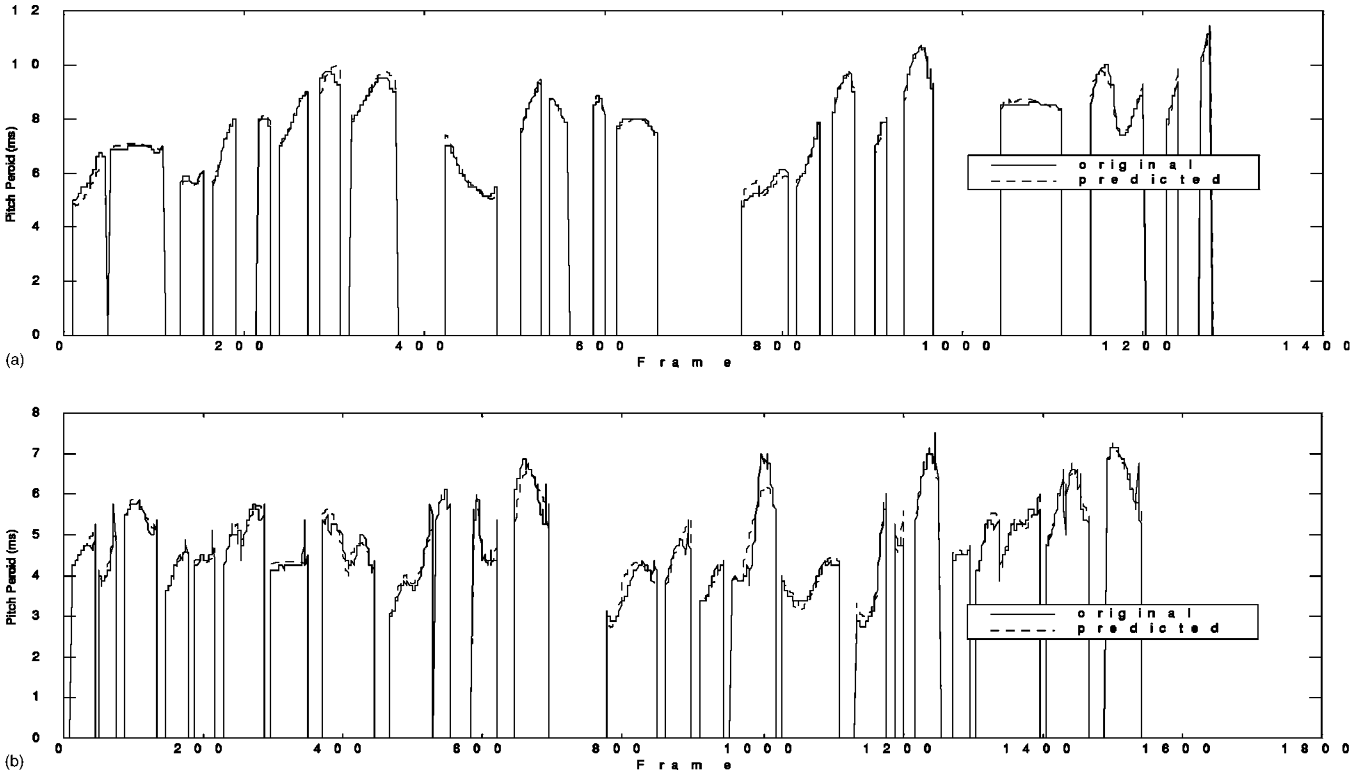


FIG. 5. Examples of reconstructed pitch contours for (a) an inside test utterance: "tzai-4 guo-2 ren-2 shiau-1 fei-4 shi-2 guan-4 gai-3 bian-4, guo-2 min-2 suo-3 de-2 ti-2 gau-1, shin-4 yung-4 dai-4 kuan-3 shr-4 chang-2, cheng-2 wei-2 chian-2 li-4 shr-4 chang-2" and (b) an outside test utterance: "tzai-4 yi-4 guo-2 jeng-4 jing-1 huen-4 luan-4 jung-1 lin-2 wei-2 shou-4 ming-4 de-5 chi-2 an-1 pei-2, wei-4 lai-2 tzai-4 jeng-4 jing-1 liang-3 fang-1 mian-4 dou-1 you-3 bu-4 shau-3 jian-1 kuen-4 ren-4 wu-4 dai-4 wan-2 cheng-2."

introduce *a priori* knowledge of tone patterns in Mandarin speech. As reported in Ref. 16, tone 1 is a high-level tone that starts in a speaker's high F0 range and remains high; tone 2 is a mid-rising tone that starts in the speaker's mid F0 range, remains level or drops slightly during the first half of the vowel, and then rises to a high-level tone at the end; tone 3 is a low-falling tone that starts in the speaker's mid range and falls to the low range; tone 4 is a high-falling tone that usually peaks around the vowel onset and then falls to the low F0 range at the end; and tone 5 is a low-energy tone that has a relatively arbitrary pitch contour pattern. The F0 contour of each of the first four tones can be represented by a simple single standard pattern, as shown in Fig. 7. However, syllable pitch contour patterns in continuous speech vary highly and can deviate dramatically from their canonical forms.

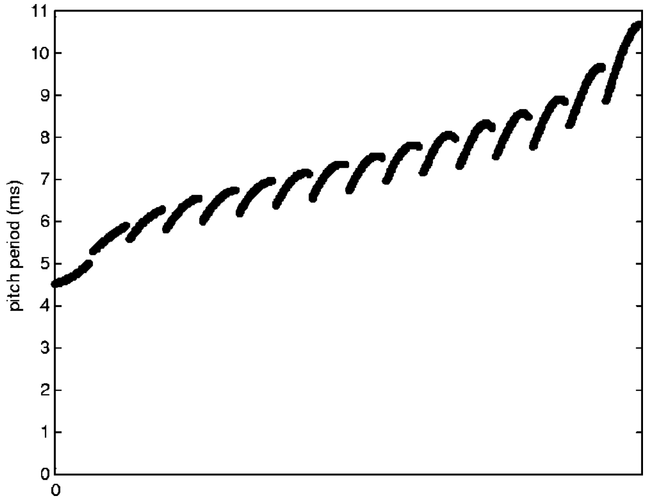


FIG. 6. The effect on the syllable pitch contour of the 16 unified prosodic states of the pitch mean and shape models. Patterns are plotted from left to right in increasing order of the prosodic state index.

Table V shows the CFs of the affecting factors of the previous, current, and following tones in the pitch mean model. As can be seen in Table V, the CFs of the affecting factors of the current tone had negative values for tones 1 and 4, and a positive value for the other three tones. Due to the fact that the effect of a positive (negative) CF was to decrease (increase) the F0 mean, the CFs of the affecting factors of the current tone were well matched with the *a priori* phonologic knowledge discussed above. It was also reported in Ref. 34 that all tones, preceding a tone 3, had a

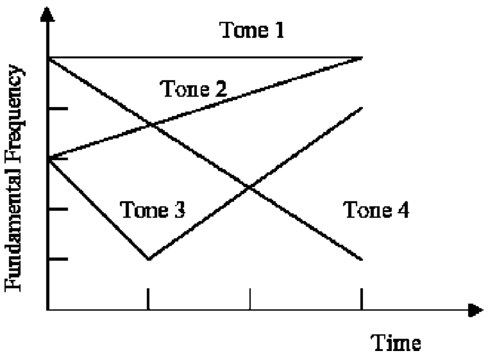


FIG. 7. Standard F0 contour patterns of the first four tones.

TABLE V. The inferred CFs for the affecting factors of the current, preceding and following tones in the pitch mean model (unit of pitch period: ms).

Tone	1	2	3	4	5
β_t CF of current tone	−0.154	0.054	0.160	−0.035	0.128
β_{pt} CF of previous tone	−0.022	−0.034	0.018	0.024	0.029
β_{ft} CF of following tone	0.022	−0.003	−0.047	0.011	0.013

much higher F0 level than they did when they preceded other tones, and all tones had a slightly lower F0 level when they preceded a tone 1. In addition, all tones following tone 1 or 2 had a higher F0 level than they did when they followed tone 3 or 4. These phenomena corroborated the results shown in Table V. Specifically, the effect of the relatively large negative CF ($\beta_{ft}=-0.047$) for $ft=3$ greatly increased the F0 level of the current syllable when it preceded a tone 3, while the positive CF ($\beta_{ft}=0.022$) for $ft=1$ decreased the F0 level of the current syllable when it preceded a tone 1. Similarly, $\beta_{pt}=-0.022$ for $pt=1$ and $\beta_{pt}=-0.034$ for $pt=2$ increased the F0 level of the current syllable when it followed a tone 1 or 2, while the positive CFs ($\beta_{pt}=0.018,0.024,0.029$) decreased the F0 level of the current syllable when it followed a tone 3, 4, or 5.

An advantage of the proposed pitch modeling method is that it provides a quantitative and more complete description of the coarticulation effect of neighboring tones rather than conventional qualitative descriptions of some of the *sandhi* rules. This can be illustrated by reconstructing the pitch contour patterns using the CFs of tone-related affecting factors and the average values of the pitch mean and shape models, while ignoring the CFs of all the other affecting factors. Specifically, the pitch contour pattern of the current tone t_c with the preceding tone t_p and the following tone t_f can be calculated, based on the proposed pitch mean and shape models, as follows:

$$\tilde{f}\left(\frac{i}{M}\right)=e^{\hat{f}(i/M)}, \quad 0\leqslant i\leqslant M, \tag{21}$$

where

$$\hat{f}\left(\frac{i}{M}\right)=\sum_{j=0}^{3}\hat{\alpha}_j\cdot\phi_j\left(\frac{i}{M}\right), \quad 0\leqslant i\leqslant M, \tag{22}$$

$$\hat{\alpha}_0=\mu+\beta_{pt=t_p}+\beta_{t=t_c}+\beta_{ft=t_f}, \tag{23}$$

$$\begin{bmatrix}\hat{\alpha}_1\\ \hat{\alpha}_2\\ \hat{\alpha}_3\end{bmatrix}=\boldsymbol{\mu}+\boldsymbol{\beta}_{tc=t_pt_ct_f}. \tag{24}$$

Figure 8 shows two examples. Figure 8(a) displays the reconstructed patterns for the current tones in tone combinations of 033, 133, 233, 333, 433, 533, 030, and 020. It should be noted that 0 denotes a case in which the effect of the previous or following tone is ignored. It can also be seen that all six patterns of tone 3 following tone 3 (i.e., 033, 133, 233, 333, 433, and 533) more closely resemble a pure tone 2 (i.e., 020) than a pure tone 3 (i.e., 030). This corroborates the well-known *sandhi* rule for a 33-tone pair, which says that a tone 3 will change to a tone 2 when it precedes a tone 3. In addition, these six patterns also show their dependence on the preceding tone. Roughly, their beginning parts were adjusted in order to be more smoothly concatenated with the patterns of the preceding tones. Figure 8(b) displays the reconstructed patterns for tone combinations of 044, 144, 244, 344, 444, 544, and 040; it shows that all six patterns of tone 4 following tone 4 (i.e., 044, 144, 244, 344, 444, and 544) have a smaller slope and lower ending point, which agrees with a previous finding.[3] These six patterns also show that they depend on the preceding tone.

We then examined the effects of the *initial* and *final* of the current syllable. We divided all 22 *initials* into seven broad classes, and 40 *finals* into seven broad classes, according to the manner of articulation. *Initial* classes included $I_0=\{\text{null}\ \mathit{initial}\}$, $I_1=\{\text{b,d,g}\}$, $I_2=\{\text{f,s,sh,shi,h}\}$, $I_3=\{\text{m,n,l,r}\}$, $I_4=\{\text{ts,ch,chi}\}$, $I_5=\{\text{p,t,k}\}$, and $I_6=\{\text{tz,j,ji}\}$. *Final* classes included $F_0=\{\text{low vowels}\}$, $F_1=\{\text{middle vowels}\}$, $F_2=\{\text{high vowels}\}$, $F_3=\{\text{compound vowels}\}$, $F_4=\{\text{vowels with nasal ending}\}$, $F_5=\{\text{retroflexion}\}$, and $F_6=\{\text{null vowels}\}$. Table VI(a) shows the CFs for these seven *initial* classes and seven *final* classes in the pitch mean model. It can be found in Table VI(a) that the positive CFs for {b,d,g}, {f,s,sh,shi,h}, {ts,ch,chi}, and {tz,j,ji} lowered the syllable F0 mean, while all the others raised the syllable F0 mean. As for the *finals*, the positive CFs of the low vowels, compound vowels, and null vowels lowered the syllable F0 mean, while the negative CFs of the middle vowels, high vowels, nasal endings, and retroflexion raised the syllable F0 mean. However, all these 14 CFs were relatively small, compared to the CFs of the other affecting factors. This shows that the *initial* and *final* of the current syllable were not major factors affecting the syllable pitch level. Table VI(b) shows the CFs of these seven *initial* classes and seven *final* classes in the pitch shape model. It can also be seen that all the CFs are relatively small, so they also are not major factors affecting the syllable pitch shape.

Table VII shows the estimated CFs of the three affecting factors for the four training speakers. As observed in Table VII(a), the four CFs of the dynamic range scaling factor in the pitch mean model were all close to 1 for the four speakers, while the four CFs of level shift were all close to 0. In addition, all the CFs of shape shift shown in Table VII(b) were relatively small. This shows that the use of additional speaker affecting factors, other than the frame-based speaker normalization performed in the preprocessing stage, had little effect on the improvement of the pitch mean and shape models. Actually, we have already shown in Sec. III B. that the RMSEs of the reconstructed pitch contour, formed by the proposed pitch mean and shape models, were almost the same when we excluded these three speaker affecting factors.

We then examined the prosodic states of the pitch mean model, labeled by the EM algorithm, in more detail. As mentioned in Sec. I, the prosodic state is conceptually defined as the state of the current syllable in a prosodic phrase. From this definition, one can expect the prosodic phrase structure of an utterance to be characterized by its prosodic state se-

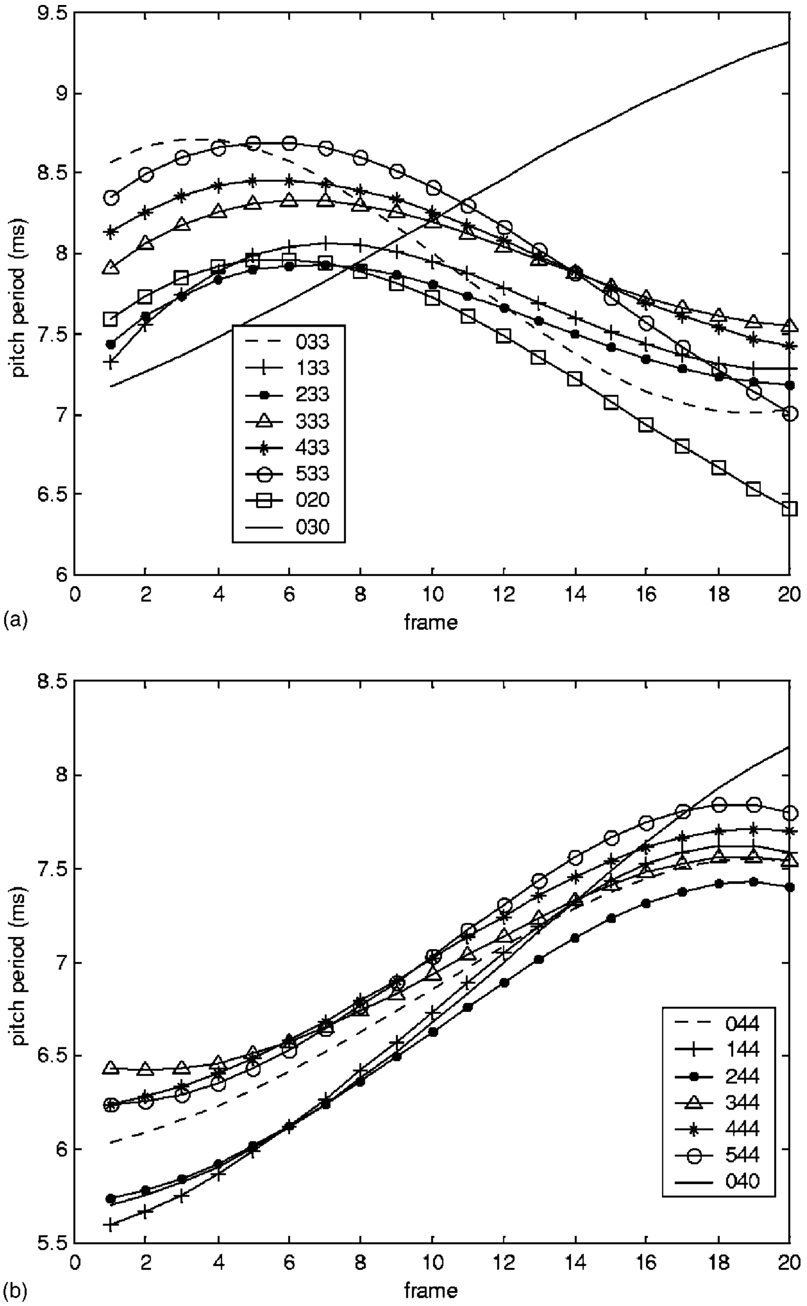


FIG. 8. (a) A comparison of the patterns of tone 3 preceding another tone 3 with the canonical patterns of tone 2 and tone 3. (b) A comparison of the patterns of tone 4 preceding another tone 4 with the canonical pattern of tone 4.

quence. First, a brief description of the characteristics of prosodic phrases will be given here. It is well known that the global downtrend tendency of F0 is to decline over the course of an utterance.[28] It is also known that a slight pitch reset of the bottom line of intonation will occur at a prosodic word boundary, and that a significant pitch reset of the bottom line of intonation will occur at an intonational phrase boundary.[35] The pitch mean sequence of an utterance will, therefore, show repeating patterns of smooth uptrend curves, starting with lower pitch levels and ending at higher pitch levels, representing the prosodic phrase structure of the utterance. With interference due to the tone effect, however, the prosodic phrase patterns are not as apparent as they are when they are observed based on the original pitch mean sequence of an utterance. A typical example is displayed in Fig. 9, where one can see that the original pitch mean sequence of the utterance exhibited a repeating uptrend pattern, while some had large zigzag variations. To eliminate the tone effect, we formed a reconstructed pitch mean sequence of the utterance by calculating the sum of the CFs of the prosodic state sequence and the mean value of the normalized pitch mean. The reconstructed pitch mean sequence is also displayed in Fig. 9, where it is clearly shown that the reconstructed pitch mean sequence was a better representation of the smooth repeating uptrend patterns of the prosodic phrases than the original pitch mean sequence was, because the large zigzag variations caused by the tone effect had been largely eliminated. Figure 10 shows the autocorrelation functions of the original and reconstructed pitch mean sequences. The higher autocorrelation values shown in Fig. 10 imply that the uptrend prosodic phrase patterns, represented by the reconstructed pitch mean sequence, were smoother. The figure also shows that the lowest autocorrelation value occurred at the 6-syllable lag. This agrees with the fact that the aver-

TABLE VI. The inferred CFs for the affecting factors of 7 *initial* and 7 *final* classes in the (a) pitch mean and (b) pitch shape models (unit of pitch period: ms).

(a)

Class	{null *initial*}	{b,d,g}	{f,s,sh,shi,h}	{m,n,l,r}	{ts,ch,chi}	{p,t,k}	{tz,j,ji}
β_i	−0.008	0.004	0.011	−0.013	0.003	−0.014	0.003
β_f	0.011	−0.001	−0.004	0.008	−0.005	−0.019	0.004

(b)

Class	{low vowels}	{middle vowels}	{high vowels}	{compound vowels}	{vowelswith nasal ending}	{retroflexion}	{null vowels}
$\mathbf{b}_i$ (×100)	$\begin{bmatrix}-0.971\\1.125\\-0.548\end{bmatrix}$	$\begin{bmatrix}0.522\\0.015\\-0.020\end{bmatrix}$	$\begin{bmatrix}0.509\\-0.440\\0.321\end{bmatrix}$	$\begin{bmatrix}-0.520\\0.506\\-0.697\end{bmatrix}$	$\begin{bmatrix}-1.270\\-0.666\\0.648\end{bmatrix}$	$\begin{bmatrix}-0.111\\-0.627\\0.389\end{bmatrix}$	$\begin{bmatrix}0.722\\-0.161\\0.075\end{bmatrix}$
$\mathbf{b}_f$ (×100)	$\begin{bmatrix}0.224\\-0.131\\0.182\end{bmatrix}$	$\begin{bmatrix}0.641\\0.280\\-0.095\end{bmatrix}$	$\begin{bmatrix}-0.278\\0.865\\-0.076\end{bmatrix}$	$\begin{bmatrix}0.978\\-0.017\\-0.094\end{bmatrix}$	$\begin{bmatrix}-0.640\\-0.703\\0.166\end{bmatrix}$	$\begin{bmatrix}-1.266\\0.891\\-0.080\end{bmatrix}$	$\begin{bmatrix}-0.354\\0.696\\-0.291\end{bmatrix}$

age length of prosodic phrases is 6.14 syllables, as evaluated based on a 1743-syllable subset of the TL database, with major and minor breaks labeled manually. Based on the above evidence, the validity of the prosodic state definition was confirmed.

Table VIII(a) shows the inferred CFs of the 16 prosodic states in the pitch mean model. It should be noted that these 16 CFs are sorted in increasing order, with state 0 having the smallest CF value and state 15 having the largest. Thus, the lower-indexed states correspond to the beginning part of a prosodic phrase, while the higher-indexed states correspond to the ending part of a prosodic phrase. From detailed analyses, we found that the prosodic states of syllables in a prosodic phrase generally varied from small to large and were reset when they crossed prosodic phrase boundaries. This means that a change of the state's index, from large to small, indicated a possible prosodic phrase boundary. We, therefore, set the following rules to detect minor and major prosodic phrase boundaries:

$$\text{location following syllable } n = \begin{cases} \text{major boundary} & \text{if } 10 \leqslant p_n - p_{n+1} \leqslant 15, \\ \text{minor boundary} & \text{if } 4 \leqslant p_n - p_{n+1} \leqslant 9, \\ \text{nonboundary} & \text{otherwise.} \end{cases} \qquad (25)$$

Figure 11 shows some examples of prosodic labeling performed using the above rules, with "*" representing a major boundary and "&" representing a minor boundary. As shown in Fig. 11, almost all the location of PMs (punctuation marks) were marked with major or minor prosodic phrase boundaries. This closely agrees with prior knowledge that a PM is a good location for a break in the pronunciation of a long text. It can also be seen in Fig. 11 that some major and minor prosodic phrase boundaries were detected at non-PM intersyllable locations. From detailed analyses, we found that most of those locations were boundaries of long words. Table IX shows the prosodic labeling statistics. As shown, 80.7% of the location of major PMs belonging to the set {comma, period, exclamation mark, semicolon, question mark} and 69.5% of the location of the secondary major PMs belonging to the set {pause—mark in Chinese punctuation used to set off items in a series, colon} were marked with major or minor prosodic phrase boundaries. On the other hand, only 42.3% of the location of the minor PMs belonging to the set {brace, bracket, dot} and 10.8% of the location of the non-PMs were marked with major or minor prosodic phrase boundaries. From detailed analyses, we found that most of the major/minor prosodic phrase boundaries occurring at non-PM locations were breathing breaks or long-phrase boundaries; most of the major and secondary major PMs labeled with nonboundaries occurred at the ends of very short sentences, at locations near other breaks, or at the ends of sentences whose pronunciation exhibited relatively flat pitch variation. These phenomena closely matched our prior linguistic knowledge. In order to more accurately evaluate the performance of automatic prosodic labeling, we manually processed a small data set containing 1743 syllables in order to determine whether each intersyllable location was a nonbreak, a minor break, or a major break. Table X shows a comparison of the two prosodic labeling methods, where it can be seen that the accuracy of the automatic prosodic labeling method was 94.1%. If we combine these two classes of minor and major breaks into one break class, the accuracy rate increases to 97.2%. The automatic prosodic labeling method is, therefore, promising.

TABLE VII. The inferred CFs for the four training speakers in the (a) pitch mean and (b) pitch shape models (unit of pitch period: ms).

Speaker	A	B	C	D
(a)				
γ_s	1.014	0.971	1.026	0.981
β_s	−0.030	0.049	−0.044	0.041
(b)				
$\mathbf{b}_s$ (×100)	$\begin{bmatrix}0.291\\0.134\\-0.012\end{bmatrix}$	$\begin{bmatrix}0.324\\0.302\\-0.125\end{bmatrix}$	$\begin{bmatrix}-0.216\\0.349\\0.348\end{bmatrix}$	$\begin{bmatrix}-0.301\\-0.472\\-0.152\end{bmatrix}$

V. AN APPLICATION TO PREDICT PITCH FOR TTS

A hybrid method, incorporating the above pitch mean and shape models with a linear regression method to predict

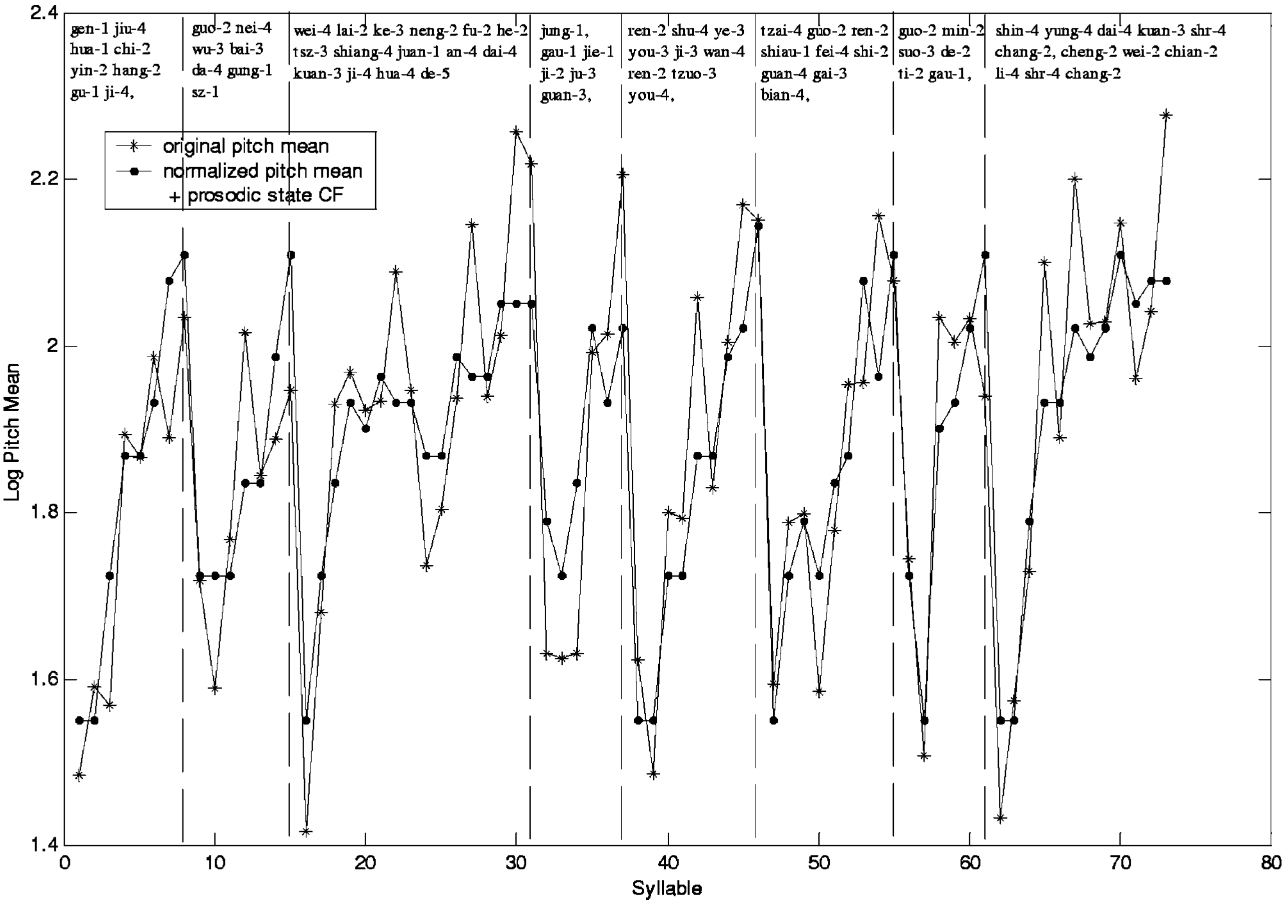


FIG. 9. A comparison between the original pitch mean sequence and the reconstructed pitch mean sequence formed by adding the mean value of the normalized pitch mean and prosodic state CFs. The sentence is "gen-1 jiu-4 hua-1 chi-2 yin-2 hang-2 gu-1 ji-4, guo-2 nei-4 wu-3 bai-3 da-4 gung-1 sz-1 wei-4 lai-2 ke-3 neng-2 fu-2 he-2 tsz-3 shiang-4 juan-1 an-4 dai-4 kuan-3 ji-4 hua-4 de-5 jung-1, gau-1 jie-1 ji-2 ju-3 guan-3, ren-2 shu-4 ye-3 you-3 ji-3 wan-4 ren-2 tzuo-3 you-4, tzai-4 guo-2 ren-2 shiau-1 fei-4 shi-2 guan-4 gai-3 bian-4, guo-2 min-2 suo-3 de-2 ti-2 gau-1, shin-4 yung-4 dai-4 kuan-3 shr-4 chang-2, cheng-2 wei-2 chian-2 li-4 shr-4 chang-2."

the syllable pitch contour for Mandarin TTS, was developed. Figure 12 shows a block diagram of the proposed method. It first estimates the prosodic state CF of each syllable from inputs of linguistic features using the linear regression technique. The linguistic features used here for this linear regression included (1) current word length: {1,2,3,>3}; (2) current syllable position in word: {first, intermediate, last}; (3) sentence length: {1,[2,5],[6,10],[11,15],[16,20],>20}; (4) current syllable position in sentence: {1st, 2nd, 3rd, [4th,5th], [6th,7th], [8th,11th], last, 2nd last, 3rd last, [5th last, 4th last], [7th last, 6th last], [11th last, 8th last], and others}, where the smaller count from the beginning or the end wins, with the count from the end breaking the tie; (5) punctuation mark after the current syllable (12 types+null); and (6) part of speech (53 types). This method then combines the predicted prosodic state CFs with the CFs of other affecting factors to form estimates of four orthogonal transform coefficients of the log-pitch contour for each syllable using the pitch mean and shape models. Here, the CFs of the tone- and syllable-related affecting factors were obtained directly by looking-up the corresponding CF tables constructed in the training phase. On the other hand, the three CFs of the speaker could be directly specified as additional inputs to control the dynamic range of pitch contour. In this study, in order to disregard the effect of the speaker's variability, the values of the three CFs of the speaker were assigned to the values obtained by the EM algorithm in training. In addition, the values of the normalized pitch mean and shape parameters, required to calculate the output orthogonal transform coefficients in Eqs. (6) and (7), could be obtained through

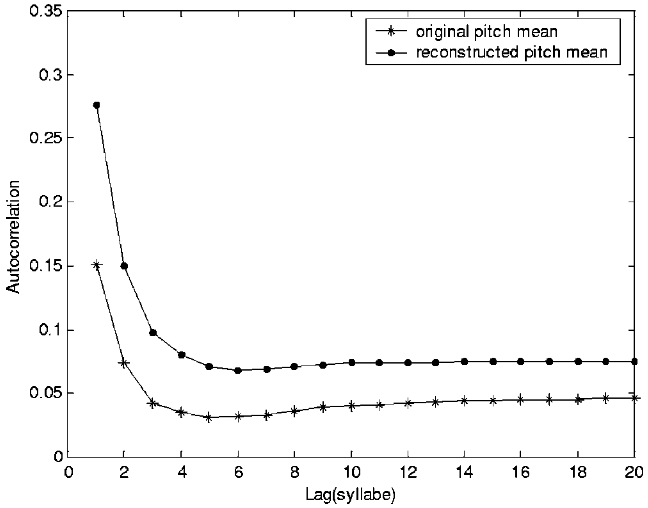


FIG. 10. Autocorrelation functions of the original pitch mean sequence and the reconstructed pitch mean sequence formed by adding the mean value of the normalized pitch mean and the prosodic state CFs.

TABLE VIII. The inferred CFs for the 16 prosodic states in the (a) pitch mean and (b) pitch shape models. The CFs in (a) are sorted from small to large (unit of pitch period: ms).

(a)

State	0	1	2	3	4	5	6	7
β_p	−0.400	−0.225	−0.159	−0.113	−0.081	−0.047	−0.016	0.014
State	8	9	10	11	12	13	14	15
β_p	0.039	0.073	0.102	0.130	0.161	0.196	0.265	0.348

(b)

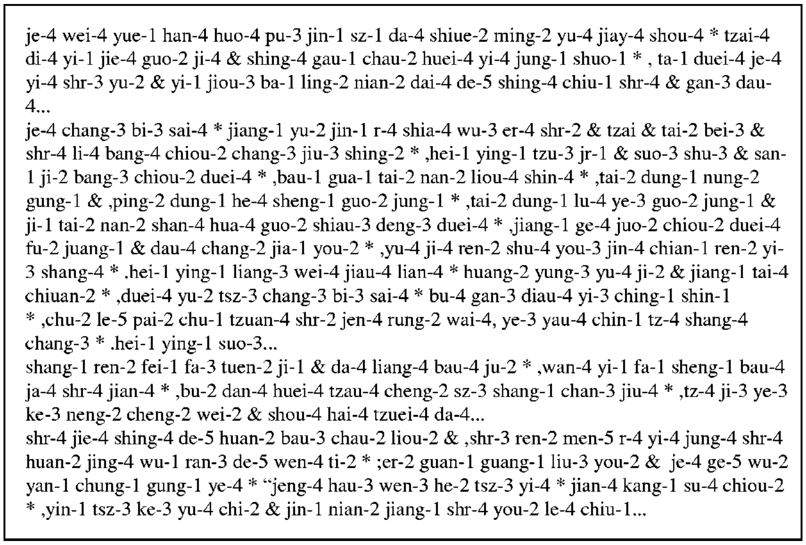
State	0	1	2	3	4	5	6	7
$\mathbf{b}_q$ (×100)	$\begin{bmatrix}-3.662\\-4.832\\-0.108\end{bmatrix}$	$\begin{bmatrix}0.047\\-0.179\\-1.535\end{bmatrix}$	$\begin{bmatrix}-1.167\\-3.221\\-0.436\end{bmatrix}$	$\begin{bmatrix}-2.297\\4.218\\0.346\end{bmatrix}$	$\begin{bmatrix}-2.245\\-0.591\\-0.267\end{bmatrix}$	$\begin{bmatrix}-1.558\\1.194\\-0.466\end{bmatrix}$	$\begin{bmatrix}-4.033\\0.582\\0.961\end{bmatrix}$	$\begin{bmatrix}-1.167\\-1.550\\0.248\end{bmatrix}$
State	8	9	10	11	12	13	14	15
$\mathbf{b}_q$ (×100)	$\begin{bmatrix}9.354\\1.249\\-1.476\end{bmatrix}$	$\begin{bmatrix}-0.164\\0.479\\0.304\end{bmatrix}$	$\begin{bmatrix}3.707\\0.295\\-0.773\end{bmatrix}$	$\begin{bmatrix}-1.340\\-0.798\\1.164\end{bmatrix}$	$\begin{bmatrix}0.849\\2.249\\0.184\end{bmatrix}$	$\begin{bmatrix}0.094\\1.469\\1.603\end{bmatrix}$	$\begin{bmatrix}1.550\\-2.455\\0.684\end{bmatrix}$	$\begin{bmatrix}-0.279\\-0.289\\0.106\end{bmatrix}$

similar linear regressive estimation. However, because of the fact that their variance was very small, we simply set their values to the means of these two models. Lastly, we generated the reconstructed syllable pitch contour by performing orthogonal polynomial expansion and frame-based speaker denormalization. Notice that the linguistic features used here were extracted from the input text by an automatic word tokenization algorithm, with an 80 000-word lexicon and a manual postcheck.

For a performance comparison, the conventional linear regression method was also implemented. It uses a linear combination of weighted input linguistic features to generate the four orthogonal transform parameters of the log-pitch contour for each syllable. To ensure a fair comparison, the input linguistic features used in the method comprised all the above features and some other syllable-level features, including the lexical tones (5×3 types) of the preceding, current, and succeeding syllables; the *initials* (21 types+null) of the current and succeeding syllables; the *medials* (3 types+null) of the current syllable; and the *finals* (14 types) of the preceding and current syllables.

Experimental results obtained using the TL database are shown in Table XI, where it can be clearly found that the hybrid method, with 16 prosodic states, outperformed the linear regression method. RMSEs of 0.996 and 0.865 ms/frame between the predicted and observed pitch periods were obtained in the closed and open tests, respectively. The results were better than those, 1.511 and 1.179 ms/frame, achieved using the linear regression method. Notice that the RMSEs resulting from orthogonal transformation were 0.17 and 0.19 ms for the closed and open test data sets, respectively.

Lastly, an AB test and an MOS perceptual test, similar to those discussed in Sec. III B, were employed to evaluate the performance of the proposed hybrid method and the conventional linear regression method. Synthesized speech recordings, with syllable pitch contours estimated using these two methods, were compared. The same 16 listeners were involved in these two tests. The experimental results of the AB test showed that 98.75% (100%) of the hybrid synthesized speech was found to sound better in the inside (outside) test, while 1.25% (0%) of the linear regression synthesized speech was found to sound better. The experimental results of the MOS test showed that average MOSs of 3.5 (3.18) and



je-4 wei-4 yue-1 han-4 huo-4 pu-3 jin-1 sz-1 da-4 shiue-2 ming-2 yu-4 jiay-4 shou-4 * tzai-4 di-4 yi-1 jie-4 guo-2 ji-4 & shing-4 gau-1 chau-2 huei-4 yi-4 jung-1 shuo-1 * , ta-1 duei-4 je-4 yi-4 shr-3 yu-2 & yi-1 jiou-3 ba-1 ling-2 nian-2 dai-4 de-5 shing-4 chiu-1 shr-4 & gan-3 dau-4...

je-4 chang-3 bi-3 sai-4 * jiang-1 yu-2 jin-1 r-4 shia-4 wu-3 er-4 shr-2 & tzai & tai-2 bei-3 & shr-4 li-4 bang-4 chiou-2 chang-3 jiu-3 shing-2 * ,hei-1 ying-1 tzu-3 jr-1 & suo-3 shu-3 & san-1 ji-2 bang-3 chiou-2 duei-4 * ,bau-1 gua-1 tai-2 nan-2 liou-4 shin-4 * ,tai-2 dung-1 nung-2 gung-1 & ,ping-2 dung-1 he-4 sheng-1 guo-2 jung-1 * ,tai-2 dung-1 lu-4 ye-3 guo-2 jung-1 & ji-1 tai-2 nan-2 shan-4 hua-4 guo-2 shiau-3 deng-3 duei-4 * ,jiang-1 ge-4 juo-2 chiou-2 duei-4 fu-2 juang-1 & dau-4 chang-2 jia-1 you-2 * ,yu-4 ji-4 ren-2 shu-4 you-3 jin-4 chian-1 ren-2 yi-3 shang-4 * .hei-1 ying-1 liang-3 wei-4 jiau-4 lian-4 * huang-2 yung-3 yu-4 ji-2 & jiang-1 tai-4 chiuan-2 * ,duei-4 yu-2 tsz-3 chang-3 bi-3 sai-4 * bu-4 gan-3 diau-4 yi-3 ching-1 shin-1 * ,chu-2 le-5 pai-2 chu-1 tzuan-4 shr-2 jen-4 rung-2 wai-4, ye-3 yau-4 chin-1 tz-4 shang-4 chang-3 * .hei-1 ying-1 suo-3...

shang-1 ren-2 fei-1 fa-3 tuen-2 ji-1 & da-4 liang-4 bau-4 ju-2 * ,wan-4 yi-1 fa-1 sheng-1 bau-4 ja-4 shr-4 jian-4 * ,bu-2 dan-4 huei-4 tzau-4 cheng-2 sz-3 shang-1 chan-3 jiu-4 * ,tz-4 ji-3 ye-3 ke-3 neng-2 cheng-2 wei-2 & shou-4 hai-4 tzuei-4 da-4...

shr-4 jie-4 shing-4 de-5 huan-2 bau-3 chau-2 liou-2 & ,shr-3 ren-2 men-5 r-4 yi-4 jung-4 shr-4 huan-2 jing-4 wu-1 ran-3 de-5 wen-4 ti-2 * ;er-2 guan-1 guang-1 liu-3 you-2 & je-4 ge-5 wu-2 yan-1 chung-1 gung-1 ye-4 * “jeng-4 hau-3 wen-3 he-2 tsz-3 yi-4 * jian-4 kang-1 su-4 chiou-2 * ,yin-1 tsz-3 ke-3 yu-4 chi-2 & jin-1 nian-2 jiang-1 shr-4 you-2 le-4 chiu-1...

FIG. 11. Examples of labeling minor (&) and major (*) prosodic phrase boundaries using rules based on the prosodic state differences of the pitch mean model.

1.34 (1.3) were obtained using the hybrid method and linear regression method, respectively, in the inside (outside) test. Based on the results of these two subjective tests, the proposed hybrid method was obviously better.

TABLE IX. Prosodic labeling statistics generated by Eq. (25). Here major PM={comma, period, exclamation mark, semicolon, question mark}; secondary major PM={pause, colon}, and minor PM={brace, bracket, dot}.

	Break		
PM	Nonboundary	Minor boundary	Major boundary
Non-PM	89.18%	9.80%	1.02%
Minor PM	57.73%	33.48%	8.80%
Secondary Major PM	30.52%	44.65%	24.83%
Major PM	19.31%	31.66%	49.02%

TABLE X. A comparison between the prosodic phrase boundaries automatically generated by the rules based on stated differences of the pitch mean model and the manually labeled ones using a 1743-syllable subset of the TL database (unit: syllable).

	Automatic		
Manual	Nonboundary	Minor boundary	Major boundary
Nonboundary	1463	34	2
Minor boundary	10	94	38
Major boundary	3	16	83

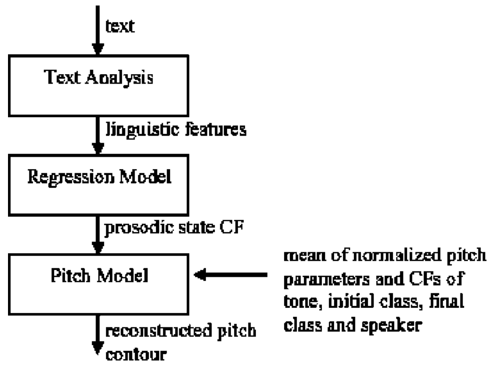

text
Text Analysis
linguistic features
Regression Model
prosodic state CF
Pitch Model
mean of normalized pitch parameters and CFs of tone, initial class, final class and speaker
reconstructed pitch contour

FIG. 12. A block diagram of the proposed hybrid method for syllable pitch contour prediction.

TABLE XI. The RMSEs of the hybrid method, with 16 prosodic states, and the linear regression method (unit: ms/frame).

RMSEs	Closed test	Open test
Hybridregression	0.996	0.865
Regression	1.511	1.179

VI. CONCLUSION AND FUTURE WORKS

This paper has presented a new statistics-based syllable pitch contour modeling method for Mandarin speech. Experimental results confirmed its effectiveness at separating several main factors that seriously affect the mean and shape of the syllable log-pitch contour of Mandarin utterances. All the inferred CFs of the affecting factors conformed well with our prior linguistic knowledge. In addition, the prosodic states labeled by the EM algorithm were linguistically meaningful, and the repeating uptrend pitch patterns of the prosodic phrase structure of an utterance were well represented by its prosodic state sequence. The proposed pitch contour modeling method is, therefore, extremely promising.

Some future work is well warranted. First, as discussed in Sec. I, only the first subtask of the complicated pitch modeling procedure was undertaken in the current study; this involved modeling the relationship between the syllable pitch contour features and some affecting factors, including local phonetic features, the speaker, and the prosodic state. The second subtask, which would explore the relationship between the prosodic state and high-level linguistic cues, is still untouched. We will undertake this second phase of research in the near future, using a tree-bank database. Second, by taking advantage of pitch modeling performed using only acoustic and simple phonetic features, we can apply the syllable pitch mean and shape models in such applications as tone recognition and prosodic labeling, which do not need *a priori* high-level linguistic information, such as word tokenization or syntactic features.

ACKNOWLEDGMENTS

This work was supported in part by the MOE under Contract No. EX-92-E-FA06-4-4 and in part by the NSC under Contract No. NSC91-2219-E-009-038.

[1] A. I. C. Monaghan and D. R. Ladd, ‘‘Manipulating Synthetic Intonation for Speaker Characterisation,’’ ICASSP (1991), S7.11, pp. 453–456.

[2] L.-S. Lee, C.-Y. Tseng, and C.-J. Hsieh, ‘‘Improved Tone Concatenation Rules in a Formant-Based Chinese Text-to-Speech System,’’ IEEE Trans. Speech Audio Process. **1**(3), 287–294 (1993).

[3] L.-S. Lee, C.-Y. Tseng, and M. Ouh-young, ‘‘The Synthesis Rules in a Chinese Text-to-speech System,’’ IEEE Trans. Acoust., Speech, Signal Process. **37**(9), 1309–1319 (1989).

[4] B. Ao, C. Shih, and R. Sproat, ‘‘A Corpus-Based Mandarin Text-To-Speech Synthesizer,’’ ICSLP (1994), S29, 8.1–8.4, pp. 1771–1774.

[5] S. H. Chen, S. H. Hwang, and Y. R. Wang, ‘‘An RNN-based prosodic information synthesizer for Mandarin text-to-speech,’’ IEEE Trans. Speech Audio Process. **6**(3), 226–239 (1998).

[6] N.-H. Pan, W.-T. Jen, S.-S. Yu, M.-S. Yu, S.-Y. Huang, and M.-J. Wu, ‘‘Prosody Model in a Mandarin Text-to-Speech System Based on a Hierarchical Approach,’’ IEEE International Conference on Multimedia and Expo (2000), Vol. 1, pp. 448–451.

[7] S.-H. Kim and J.-Y. Kim, ‘‘Efficient Method of Establishing Words Tone Dictionary for Korean TTS system,’’ Eurospeech, 1997.

[8] M. Dong and K.-T. Lua, ‘‘Pitch Contour Model for Chinese Text-to-Speech using CART and Statistical Model,’’ ICSLP (2002), pp. 2405–2408.

[9] Y. Ishikawa and T. Ebihara, ‘‘On the Global F0 Shape Model using a

Transition Network for Japanese Text-to-Speech Systems,'' Eurospeech, 1997.

[10] S.-H. Chen and Y.-R. Wang, ''Tone Recognition of Continuous Mandarin Speech Based on Neural Networks,'' IEEE Trans. Speech Audio Process. **3**(2), 146–150 (1995).

[11] W.-J. Yang, J.-C. Lee, Y.-C. Chang, and H.-C. Wang, ''Hidden Markov Model for Mandarin Lexical Tone Recognition,'' IEEE Trans. Acoust., Speech, Signal Process. **36**(7), 988–992 (1988).

[12] C. W. Wightman and M. Ostendorf, ''Automatic Labeling of Prosodic Patterns,'' IEEE Trans. Speech Audio Process. **2**(4), 469–481 (1994).

[13] A. Batliner, R. Kompe, A. Kiebling, H. Niemann, and E. Noth, ''Syntactic-Prosodic Labeling of Large Spontaneous Speech Data-Bases,'' ICSLP (1996), pp. 1720–1723.

[14] X. Lin, Y. Chen, S. Lim, and C. Lim, ''Recognition of Emotional State from Spoken Sentences,'' IEEE 3rd Workshop on Multimedia Signal Processing (1999), pp. 469–473.

[15] M. V. Chan, X. Feng, J. A. Heinen, and R. J. Niederjohn, ''Classification of Speech Accents with Neural Networks,'' *IEEE International Conference on Neural Networks* (1994), Vol. 7, pp. 4483–4486.

[16] C.-L. Shih, ''Tone and Intonation in Mandarin,'' *Working Papers of the Cornell Phonetics Laboratory*, No. 3, pp. 83–109, June 1988.

[17] Z.-J. Wu, ''Can Poly-Syllabic Tone-*Sandhi* Patterns be the Invariant Units of Intonation in Spoken Standard Chinese,'' ICSLP (1990), pp. 12.10.1–12.10.4.

[18] P. Taylor, ''Analysis and synthesis of intonation using the tilt model,'' J. Acoust. Soc. Am. **107**, 1697–1714 (2000).

[19] C. Shih, G. Kochanski, and E. Fosler-Lussier, ''Implications of Prosody Modeling for Prosody Recognition,'' *Proceedings of the ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding* (2001), pp. 133–138.

[20] N. M. Veilleux and M. Ostendorf, ''Probabilistic Parse Scoring with Prosodic Information,'' ICASSP (1993), pp. II-51–II-54.

[21] K. Iwano and K. Hirose, ''Prosodic Word Boundary Detection Using Statistical Modeling of Moraic Fundamental Frequency Contours and Its Use for Continuous Speech Recognition,'' ICASSP (1999), pp. 133–136.

[22] X. Sun, ''Predicting Underlying Pitch Targets for Intonation Modeling,'' Proc. of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis, Perthshire, Scotland (2001), pp. 143–148.

[23] A. Ljolje and F. Fallside, ''Synthesis of Natural Sounding Pitch Contours in Isolated Utterances Using Hidden Markov Models,'' IEEE Trans. Acoust., Speech, Signal Process. **ASSP-34**(5), 1074–1080 (1986).

[24] J. R. Bellegarda, K. E. A. Silverman, K. Lenzo, and V. Anderson, ''Statistical Prosodic Modeling: From Corpus Design to Parameter Estimation,'' IEEE Trans. Speech Audio Process. **9**(1), 52–66 (2001).

[25] J.-H. Ni, R.-H. Wang, and K. Hirose, ''Quantitative Analysis and Formulation of Tone Concatenation in Chinese F0 Contours,'' EUROSPEECH, 1997.

[26] M. Abe and H. Sato, ''Two-stage F0 Control Model Using Syllable Based F0 Units,'' ICASSP (1992), pp. II-53–II-56.

[27] D. T. Chappell and J. H. L. Hansen, ''Speaker-Specific Pitch Contour Modeling and Modification,'' ICASSP, 1998.

[28] C. Shih, ''Declination in Mandarin,'' *Intonation: Theory, Models and Applications*, Proceedings of an ESCA Workshop, Athens, Greece (1997), pp. 293–296.

[29] L. Aijun, ''Chinese Prosody and Prosodic Labeling of Spontaneous Speech,'' Speech Prosody, 2002.

[30] W.-J. Wang, Y.-F. Liao, and S.-H. Chen, ''Prosodic Modeling of Mandarin Speech and Its Application to Lexical Decoding,'' Eurospeech 99, Vol. 2, pp. 743–746.

[31] H.-Y. Hsieh, Ren-Y. Lyu, and L.-S. Lee, ''Use of Prosodic Information to Integrate Acoustic and Linguistic Knowledge in Continuous Mandarin Speech Recognition with Very Large Vocabulary,'' ICSLP (1996), Vol. 2, pp. 809–812.

[32] S. de Tournemire, ''Identification and Automatic Generation of Prosodic Contours for a Text-To-Speech Synthesis System in French,'' Eurospeech, 1997.

[33] F.-C. Chou, C.-Y. Tseng, K.-J. Chen, and L.-S. Lee, ''A Chinese Text-to-Speech Based on Part-of-Speech Analysis, Prosodic Modeling and Non-uniform Units,'' ICASSP (1997), pp. 923–926.

[34] C. Wang and S. Seneff, ''Improved Tone Recognition by Normalizing for Coarticulation and Intonation Effects,'' ICSLP, 2000.

[35] Y. Yufang and W. Bei, ''Acoustic Correlates of Hierarchical Prosodic Boundary in Mandarin,'' Speech Prosody, 2002.