Unsupervised joint prosody labeling and modeling for Mandarin speech

Chen-Yu Chiang[a] and Sin-Horng Chen[b]
Department of Communication Engineering, National Chiao Tung University, 1001 Ta-Hsueh Road, Hsinchu 300, Taiwan, Republic of China

Hsiu-Min Yu[c]
Language Center, Chung Hua University, 707, Sec. 2, Wu-Fu Road, Hsinchu 300, Taiwan, Republic of China

Yih-Ru Wang[d]
Department of Communication Engineering, National Chiao Tung University, 1001 Ta-Hsueh Road, Hsinchu 300, Taiwan, Republic of China

(Received 11 September 2007; revised 3 December 2008; accepted 3 December 2008)

An unsupervised joint prosody labeling and modeling method for Mandarin speech is proposed, a new scheme intended to construct statistical prosodic models and to label prosodic tags consistently for Mandarin speech. Two types of prosodic tags are determined by four prosodic models designed to illustrate the hierarchy of Mandarin prosody: the break of a syllable juncture to demarcate prosodic constituents and the prosodic state to represent any prosodic domain's pitch-level variation resulting from its upper-layered prosodic constituents' influences. The performance of the proposed method was evaluated using an unlabeled read-speech corpus articulated by an experienced female announcer. Experimental results showed that the estimated parameters of the four prosodic models were able to explore and describe the structures and patterns of Mandarin prosody. Besides, certain corresponding relationships between the break indices labeled and the associated words were found, and manifested the connections between prosodic and linguistic parameters, a finding further verifying the capability of the method presented. Finally, a quantitative comparison in labeling results between the proposed method and human labelers indicated that the former was more consistent and discriminative than the latter in prosodic feature distributions, a merit of the method developed here on the applications of prosody modeling.
 [DOI: 10.1121/1.3056559]

PACS number(s): 43.72.Ar [DOS] Pages: 1164–1183

I. INTRODUCTION

The term *prosody* refers to certain inherent suprasegmental properties that carry melodic, timing, and pragmatic information of continuous speech, encompassing accentuation, intonation, rhythm, speaking rate, prominences, pauses, and attitudes or emotions intended to express. Prosodic features are physically encoded in the variations in pitch contour, energy level, duration, and silence of spoken utterances. Prosodic studies have indicated that these prosodic features are not produced arbitrarily, but rather realized after a hierarchically organized structure which demarcates speech flows into domains of varying lengths by boundary or break cues such as pre- and postboundary lengthening, pitch and energy change, pauses, etc. Therefore, prosodic structure in English, for example, functions to set up syntagmatic contrasts to mark a prosodic word (PW), an intermediate phrase, or an intonational boundary.[1–3] On the other hand, the prosodic structure of Mandarin Chinese also parses continuous speech into different prosodic constituents by breaks that reflect different levels of Chinese linguistic processing: phonetic, lexical, syntactic, and pragmatic. As a result, successive words with related prosodic feature variations are aggregated to form prosodic phrases (PPhs), and contiguous PPhs are, in turn, integrated to form PPhs of a higher level. Consequently, deep exploration and an appropriate description of speech prosody are essential to the study of the speech processing of any language given. To provide a possible specification of prosodic features of utterances, a three-layer structure comprising PWs, intermediate phrases (or PPhs) and intonational phrases are commonly used, especially at the sentential level.[4–6] Some recent studies[7] proposed to integrate PPhs into PPh groups to interpret the contributions of higher-level discourse information to the wider-range and larger variations in the prosodic features of utterances of long texts. In the science of speech processing, to model prosody is to exploit a framework or a computational model to represent a hierarchy of PPhs of speech and to describe its relationship with the syntactic structure of the associated text.

In the past many prosody modeling methods have been proposed for various applications, including generation of prosodic information for text to speech (TTS),[8–10] segmenta-

[a]Electronic mail: gene.cm91g@nctu.edu.tw
[b]Electronic mail: schen@mail.nctu.edu.tw
[c]Electronic mail: kuo@chu.edu.tw
[d]Electronic mail: yrwang@cc.nctu.edu.tw

tion of untranscribed speech into sentences or topics,[11–13] generation of punctuations from speech,[14–16] detection of interrupt points in spontaneous speech,[11,17–19] automatic speech recognition (ASR),[20–26] and so forth. It can be found from those prosody modeling studies that four main issues have been intensively addressed. The first one is concerning representing a hierarchical PPh structure indirectly by tags marking important prosodic events. Among various prosodic events explored in the relevant literature,[27–32] break type and tone pattern are the most important ones: the break types of all word boundaries can determine the hierarchical PPh structure of an utterance, and the tonal patterns of all syllables/words can indicate the accented syllables/words of an utterance and may specify the pitch contour patterns of the prosodic constituents. Several prosody representation systems have been proposed in the past. They include tones and breaks indices (ToBI) (a standard prosody transcription system for American English utterances),[27] PROSPA,[29] INTSINT,[30] and TILT.[31] Among them, ToBI and its modifications to other languages, such as Pan-Mandarin ToBI (Ref. 32) and C-ToBI,[33] are most popular conventions for Mandarin Chinese prosodic tagging. The second main issue is about realizing the constituents of a hierarchical PPh structure by using prosodic feature patterns. This is mainly used in TTS for the generation of prosodic information from prosodic tags. A common approach is to use a multicomponent representation model to superimpose several prototypical contours of multilevel PPhs for each prosodic feature.[34–36] In Ref. 34, three components of sentence-specific contours, word-specific contours, and tone-specific contours are superimposed to form the synthesized contours of pitch and syllable duration for Mandarin TTS. The third main issue is related to exploring the relationship between prosodic tags (or boundary types) and the acoustic features surrounding the associated word juncture. Patterns of pause duration, pitch, and energy around word junctures are modeled for each prosodic tag or boundary type to help speech segmentation,[11–13] topic identification,[13] punctuation generation,[14–16] interrupt point detection,[11,17–19] and ASR (Refs. 20–26) based on word-based features. The last issue is upon modeling the relationship between prosodic structure and syntactic structure. It is known that prosodic structure is closely related to syntactic structure although they are not identical. Usually, only the relationship between a prosodic tag, such as break or prominence, and contextual linguistic features of syntactic structure is built. A good break-syntax model should be very useful in predicting breaks of various levels from input text for TTS. Main methods of building a break-syntax model for TTS are hierarchical stochastic model,[37,38] N-gram model,[39] classification and regressive tree (CART),[38,40–42] Markov model,[43] artificial neural networks,[44] maximum entropy model,[45–48] etc. In the popular Markov model-based approach, emission probabilities can be generated by CART (Ref. 42) or maximum entropy model.[48]

In all those studies, prosody modeling has been proved to be useful in above-mentioned applications, and the most commonly adopted approach by the previous studies is a supervised one to construct prosodic model from an annotated speech database with tags marking prosodic events being prelabeled manually. However, the supervised prosody modeling based on human labeling unavoidably arises such problems as diseconomy due to labeler training and manual labeling labor, and interlabelers' and intralabeler's inconsistency caused by individual subjectivity and fatigue during long time labeling, respectively. This inconsistency may mislead prosody modeling to obtain erroneous results, and hence lead to unwanted degradation of modeling performance. Even in the studies where prosody labeling can be automatically done by machine, their model is still trained with a manually annotated speech corpus,[26,28,49–54] so the performance of machine labeling is still subject to the quality of human prosody labeling.

To tackle the problems arising from the supervised prosody modeling with manual labeling, this work proposes a new unsupervised approach of prosody modeling to jointly perform prosody modeling and labeling for Mandarin speech based on an unlabeled speech database. The basic idea is to properly model data and then let the modeled data determine prosodic tags by themselves. The task is to automatically determine two types of prosodic tags for all utterances of a corpus and to build four prosodic models simultaneously. The two types of prosodic tags are (1) the break types of intersyllable locations (or syllable junctures) which can be used to demarcate the constituents of a hierarchy of Mandarin speech prosody and (2) the prosodic states of syllables which can be used to construct the pitch contour patterns of prosodic constituents. As will be discussed later, the prosodic state of a syllable is defined as a quantized and normalized pitch level affected by the current tone and the coarticulations from the two nearest neighboring syllables being properly eliminated. Since it mainly carries the information of PPhs, we therefore name it to refer to the state in a PPh. In Sec. IV, we will demonstrate its capability on realizing the pitch contour patterns of multilevel PPhs. It should be mentioned that in this study only pitch information is considered in the prosodic-state tag labeling. We will extend the study to consider the other two features of syllable duration and energy level in the future. The four prosodic models are introduced to describe the various relationships between the two types of prosodic tags and all available information sources including acoustic prosodic features and syntactic structure features. The first model, referred to as the syllable pitch contour model, describes the variations in syllable pitch contours controlled by several major affecting factors. The next one, referred to as the break-acoustics model, describes the relationship between the break type of a syllable juncture and nearby acoustic features. The third one describes the relationship between the break type of a syllable juncture and contextual linguistic features. It is referred to as the break-syntax model. Finally, the last model describes the relationship between the prosodic states of syllables and the break types of neighboring syllable junctures and is referred to as the prosodic-state model. A sequential optimization training algorithm is designed to iteratively estimate parameters of the four prosodic models and find all prosodic tags using an unlabeled speech corpus. Three advantages of the proposed method can be found. First, prosody modeling and labeling are accomplished jointly and automatically without using

human-labeled training corpus. Second, all information sources, including acoustic and linguistic features, are systematically used (via introducing the four prosodic models) in the prosody labeling. We therefore expect that the result of the prosodic labeling is more consistent than that done by human, which will in turn make the four prosodic models more accurate. Third, the four prosodic models constructed address all the four main issues of prosody modeling discussed above. So they are useful models and may be directly used or extended to be used in those applications mentioned above.

The remainder of this paper is organized as follows. Section II briefly describes the prosodic structure of Mandarin speech. Section III presents the proposed method. In Sec. IV experimental results are discussed, and in Sec. V some conclusions are drawn.

II. THE HIERARCHY OF MANDARIN SPEECH PROSODY

Much literature on Chinese prosody has shown that the prosody of Mandarin speech can be organized into hierarchical structures. A commonly agreed and used structure consists of four layers, including, from the lowest layer to the highest one, syllable layer, PW layer, PPh layer (or intermediate phrase), and intonation phrase.[38,41,42,44,45,48] As far as the major prosodic information relevant to each of the layers is concerned, given that Mandarin is a monosyllabic and tonal language, where each syllable with its inherent tone contains a lexical meaning, and each tone carries a lexically contrastive role, the features of every syllabic tone of an utterance are the most important prosodic information for the lowest layer; besides, tone along with syllable constituents affects syllable duration and energy level as well. As for the second prosodic layer, a PW refers to disyllabic and multisyllabic words or phrases composed of words syntactically and semantically closely related or most frequently collocated, so the words or phrases are uttered as a single unit as in hen "very"+*bu* 'not'+*zhuan-ye* "professional" (*not very professional*). As for the third prosodic layer, PPh is composed of one or several PWs and it usually ends with a perceptible but unobvious break. Finally, intonation phrase is at the top layer of the Mandarin prosodic structure. It determines the pitch contour of the intonation of a sentence containing one or several PPhs and it ends with an obvious break. Basically, the four-layer prosodic structure interprets the pitch and duration variations in syllable well for sentential utterances.

Recently, Tseng *et al.*[7] proposed to integrate contiguous PPhs into PPh groups to interpret the contributions of higher-level discourse information to the wider-range and larger variations in syllable pitch and duration of long utterances in paragraphs. Figure 1 displays the hierarchical prosodic phrase grouping (HPG) model of Mandarin speech proposed by Tseng *et al.* It is a five-layer structure. The first three layers in the hierarchy proposed by Tseng *et al.* are the same as those of the four-layer prosodic structure discussed above, which are referred to as syllable (SYL), PW, and PPh in the system of Tseng *et al.*, respectively. The fourth layer, breath group (BG), is formed by combining a sequence of PPhs,

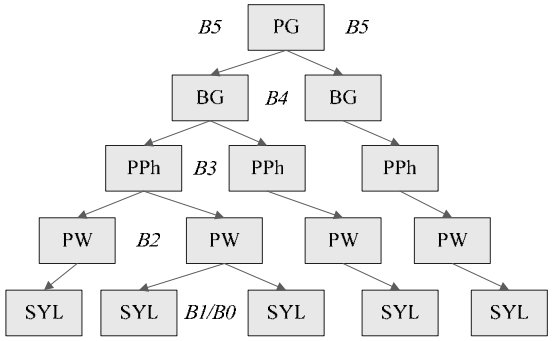

FIG. 1. A conceptual prosody hierarchy of Mandarin speech proposed by Tseng *et al.* in Ref. [7].

and a sequence of BGs, in turn, constitutes the fifth layer, prosodic phrase group (PG). The above five prosodic units are delimited by different types of the six breaks proposed by Tseng *et al.* First, *B*0 and *B*1 are defined for SYL boundaries within PW. Here, *B*0 represents reduced syllabic boundary and *B*1 represents normal syllabic boundary. Usually no identifiable pauses exist for both *B*0 and *B*1. Second, *B*4 and *B*5 are defined for BG and PG boundaries, respectively. *B*4 is a breathing pause and *B*5 is a complete speech paragraph end characterized by final lengthening coupled with weakening of speech sounds. Third, *B*2 and *B*3 are perceivable boundaries defined for PW and PPh boundaries, respectively.

In this study, we adopt the prosodic structure of Tseng *et al.* because our speech database also consists of long Mandarin utterances of paragraphs. However, we modify the break-type labeling scheme of HPG model by dividing *B*2 into two types, *B*2-1 and *B*2-2., and combining *B*4 and *B*5 into one denoted simply by *B*4. Here, *B*2-2 represents syllabic boundary of *B*2 perceived by pause, while *B*2-1 is *B*2 with *F*0 movement. The reason of dividing *B*2 into *B*2-1 and *B*2-2 is due to the difference of their acoustic cues to be modeled. On the contrary, the combination of *B*4 and *B*5 owes to the similarity of their acoustic characteristics. So, the break-type tags used is in $\Lambda=\{B0, B1, B2\text{-}1, B2\text{-}2, B3, B4\}$. These six break-type tags can be used to delimit four types of prosodic units: SYL, PW, PPh, and BG/PG. These four units are the constituents of our hierarchical prosodic structure.

To further specify the four-layer prosodic structure, a representation of its constituents using prosodic features is needed. Two main approaches of representation can be considered. One is direct representation approach to represent each individual prosodic constituent by multiple prototypical patterns for each prosodic feature of syllable pitch contour, duration, or energy level.[7–9,34–36] The other is indirect representation approach[55,56] by using some tags which carry the information of prosodic constituents and are treated as hidden. Due to the following two reasons, we do not adopt direct representation approach in the prosody modeling and labeling study. First, the technique of direct representation approach is still not mature enough to produce a good direct representation for the hierarchy of Mandarin speech prosody. The modeling errors, defined as the ratio of root mean square errors of direct representations to the standard deviations of the raw data, are still as high as about 30% for the multilayer representations of syllable duration, and energy using the

HPG model.[7] Second, a good direct representation is not easy to be realized for the case of joint prosody modeling and labeling using an unlabeled speech corpus in which the prosodic structures of all utterances are not well determined in advance. Degeneration may occur because break labeling errors may produce inaccurate representation patterns of prosodic constituents, which in turn may cause more break labeling errors to occur. Instead, we adopt an indirect representation approach to employ a new prosodic tag to represent the aggregative contributions of the constituents of the upper three layers on syllable pitch level. This tag is defined as a quantized and normalized syllable pitch level with the affections from the current tone and the two nearest neighboring tones being properly eliminated. So it carries mainly the pitch-level information of the upper three layers of the prosodic structure, i.e., PW, PPh, and BG/PG. We call it *prosodic state* to roughly mean the state in the pitch contour of a PPh (PW, PPh, or BG/PG). Two advantages of using the prosodic-state tag can be found. First, the tag is defined for each individual syllable so that the affection of a labeling error is limited to the current syllable only. No degeneration in the joint prosody modeling and labeling process will occur. Second, the tag carries the full information of pitch-level variation in the upper three layers of the prosodic structure. In Sec. IV, we will show the capability of the prosodic-state tag on constructing the pitch contour patterns of PW, PPh, and BG/PG. It is worthy to note that prosodic states of syllable duration and energy level can be similarly defined and added to the joint prosody labeling and modeling study. But for simplicity we only consider the prosodic state of syllable pitch level in this study.

III. THE PROPOSED METHOD

The proposed method first treats the problem as a model-based prosody labeling problem to define the four prosodic models to describe various relationships between the prosodic tags to be labeled and the available information sources of acoustic and syntactic features. It then extends the formulation for the joint prosody labeling and modeling problem and applies a sequential optimization procedure to jointly label prosodic tags and estimate the model parameters using an unlabeled speech corpus. We discuss these two parts in detail as follows.

A. The design of the four prosodic models

The prosody labeling problem can be generally formulated as a parametric optimization problem to find the best prosodic tag sequence $\mathbf{T}^*$ given with the acoustic feature sequence $\mathbf{A}$ of the input speech utterance and the linguistic feature sequence $\mathbf{L}$ of the associated text:

$$\mathbf{T}^* = \arg\max_{\mathbf{T}} P(\mathbf{T}|\mathbf{A},\mathbf{L}) = \arg\max_{\mathbf{T}} P(\mathbf{T},\mathbf{A}|\mathbf{L}). \quad (1)$$

Two types of prosodic tags which carry the information of prosodic structure of Mandarin speech are considered in this study. One is the break type of syllable juncture. A set of six break types, defined in Sec. II, is used. It is denoted as $\{B0, B1, B2\text{-}1, B2\text{-}2, B3, B4\}$. These six break types are used to define a hierarchy of speech prosody comprising four constituents of SYL, PW, PPh, and BG/PG. Another is the prosodic state of syllable defined as a quantized and normalized syllable pitch level with the affections of the current tone and the two nearest neighboring tones being properly eliminated. As discussed in Sec. II, it is an indirect representation of the prosodic constituents to carry the pitch-level information of PW, PPh, and BG/PG. So, $\mathbf{T}$ can be refined to comprise a break-type sequence $\mathbf{B}$ and a prosodic-state sequence $\mathbf{p}$.

Two types of acoustic features can be considered. One is the prosodic features which carry the information of prosodic constituents. Acoustic features of this type are assumed to be closely related to the prosodic-state tags and loosely related to or independent of the break-type tags. Primary features of this type include syllable pitch contour, syllable duration, and syllable energy level. For simplicity we only consider syllable pitch contour in this study and will extend the study to include the other two in the future. Another is the acoustic features used to specify the break type of syllable juncture. Acoustic features of this type are assumed to be closely related to the break-type tags and loosely related to or independent of the prosodic-state tags. Primary features of this type include pause duration and energy-dip level of syllable juncture, energy, and pitch jumps across syllable juncture, lengthening factor of syllable duration, etc. Among them, pitch jump has been implicitly considered via the use of prosodic-state tag, energy jump is somewhat a redundant feature as energy-dip level is used, and lengthening factor will be considered together with the syllable duration modeling in the future. We therefore only consider the two features of pause duration and energy-dip level in this study. From above discussions, $\mathbf{A}$ can be refined to comprise a syllable pitch contour sequence $\mathbf{sp}$, a pause duration sequence $\mathbf{pd}$, and an energy-dip level sequence $\mathbf{ed}$.

The linguistic features used span a wide range from syllable level, such as syllable tone and initial type; word level, such as syllable juncture type (intraword and interword), word length, part of speech (POS), and type of punctuation mark (PM); to syntactic tree level, such as size of syntactic phrase and syntactic juncture type (intraphrase and interphrase). Since syllable tone is an important linguistic feature and mainly used in the modeling of syllable pitch contour, we separate it from other linguistic features. So, $\mathbf{L}$ is refined to include a syllable tone sequence $\mathbf{t}$ and a reduced linguistic feature set $\mathbf{l}$.

Based on above discussions, we rewrite $P(\mathbf{T},\mathbf{A}|\mathbf{L})$ by

$$\begin{aligned} P(\mathbf{T},\mathbf{A}|\mathbf{L}) &= P(\mathbf{B},\mathbf{p},\mathbf{sp},\mathbf{pd},\mathbf{ed}|\mathbf{l},\mathbf{t}) \\ &= P(\mathbf{sp},\mathbf{pd},\mathbf{ed}|\mathbf{B},\mathbf{p},\mathbf{l},\mathbf{t})P(\mathbf{B},\mathbf{p}|\mathbf{l},\mathbf{t}), \end{aligned} \quad (2)$$

where $P(\mathbf{sp},\mathbf{pd},\mathbf{ed}|\mathbf{B},\mathbf{p},\mathbf{l},\mathbf{t})$ is a general prosodic feature model describing the variations in acoustic prosodic features $(\mathbf{sp},\mathbf{pd},\mathbf{ed})$ controlled by the prosodic tags $(\mathbf{B},\mathbf{p})$ representing the prosodic structure and the linguistic features $(\mathbf{l},\mathbf{t})$ representing the syntactic structure, and $P(\mathbf{B},\mathbf{p}|\mathbf{l},\mathbf{t})$ is a general prosody-syntax model which describes the relationship between $(\mathbf{B},\mathbf{p})$ and $(\mathbf{l},\mathbf{t})$.

Since the break-type tag sequence, $\mathbf{B}$, has already carried the prosodic cues related to syllable junctures, we there-

fore assume that the observed syllable-based acoustic feature, **sp**, and the juncture-based acoustic features, (**pd**,**ed**), are independent as **B** is given. So we split $P(\mathbf{sp},\mathbf{pd},\mathbf{ed}|\mathbf{B},\mathbf{p},\mathbf{l},\mathbf{t})$ into two terms:

$$P(\mathbf{sp},\mathbf{pd},\mathbf{ed}|\mathbf{B},\mathbf{p},\mathbf{l},\mathbf{t}) \approx P(\mathbf{sp}|\mathbf{B},\mathbf{p},\mathbf{l},\mathbf{t})P(\mathbf{pd},\mathbf{ed}|\mathbf{B},\mathbf{p},\mathbf{l},\mathbf{t}). \tag{3}$$

Here $P(\mathbf{sp}|\mathbf{B},\mathbf{p},\mathbf{l},\mathbf{t})$ is a syllable pitch contour model describing the variation in syllable pitch contour controlled by $(\mathbf{B},\mathbf{p},\mathbf{l},\mathbf{t})$ and $P(\mathbf{pd},\mathbf{ed}|\mathbf{B},\mathbf{p},\mathbf{l},\mathbf{t})$ is a break-acoustics model describing the acoustic cues of syllable junctures for different break types. In this study, the syllable pitch contour model is realized using a modified version of the syllable pitch contour model proposed previously.[55] It models the pitch contour of each syllable separately and considers four main affecting factors, including the current prosodic state p_n, the current tone t_n, and the coarticulations from the two nearest neighboring tones, t_{n-1} and t_{n+1}, conditioned, respectively, on the break types, B_{n-1} and B_n, of the syllable junctures on both sides. Specifically, the model is expressed by

$$P(\mathbf{sp}|\mathbf{B},\mathbf{p},\mathbf{l},\mathbf{t}) \approx P(\mathbf{sp}|\mathbf{B},\mathbf{p},\mathbf{t}) \approx \prod_{n=1}^{N} P(\mathbf{sp}_n|p_n,B_{n-1}^{n},t_{n-1}^{n+1}), \tag{4}$$

where

$$\mathbf{sp}_n = \mathbf{sp}_n^r + \boldsymbol{\beta}_{t_n} + \boldsymbol{\beta}_{p_n} + \boldsymbol{\beta}^f_{B_{n-1},tp_{n-1}} + \boldsymbol{\beta}^b_{B_n,tp_n} + \boldsymbol{\mu} \quad \text{for } 1 \le n \le N \tag{5}$$

is the observed pitch contour of nth syllable (referred to as *syllable n* hereafter) represented by the first four orthogonally transformed parameters of syllable log $F0$ contour;[57] $B_{n-1}^{n}=(B_{n-1},B_n)$, $t_{n-1}^{n+1}=(t_{n-1},t_n,t_{n+1})$, $\mathbf{sp}_n^r$ is the normalized (or residual) version of $\mathbf{sp}_n$, and $\boldsymbol{\beta}_x$ represents the affecting pattern (AP) of affecting factor x. Here AP means the effect of a factor on increase or decrease in the observed syllable pitch contour vector $\mathbf{sp}_n$. $\boldsymbol{\beta}_{t_n}$ and $\boldsymbol{\beta}_{p_n}$ are the APs of affecting factors t_n and p_n, respectively; tp_n is the tone pair $t_n^{n+1}=(t_n,t_{n+1})$; $\boldsymbol{\beta}^f_{B_{n-1},tp_{n-1}}$ and $\boldsymbol{\beta}^b_{B_n,tp_n}$ are the APs of forward and backward coarticulations contributed from *syllable* $n-1$ and *syllable* $n+1$, respectively; and $\boldsymbol{\mu}$ is the AP of global mean. For taking care of utterance boundaries, two special break types, B_b and B_e, are assigned to the two ending locations of all utterances, i.e., $B_0=B_b$ and $B_N=B_e$, and two special APs of coarticulation, $\boldsymbol{\beta}^f_{B_b,t_1}=\boldsymbol{\beta}^f_{B_0,tp_0}$ and $\boldsymbol{\beta}^b_{B_e,t_N}=\boldsymbol{\beta}^b_{B_N,tp_N}$, are accordingly adopted to represent the effects of utterance onset and offset, respectively. In this study, $\boldsymbol{\beta}_{p_n}$ 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. By assuming that $\mathbf{sp}_n^r$ is zero mean and normally distributed, i.e., $N(\mathbf{sp}_n^r;\mathbf{0},\mathbf{R})$, we have

$$P(\mathbf{sp}_n|p_n,B_{n-1}^{n},t_{n-1}^{n+1}) = N(\mathbf{sp}_n;\boldsymbol{\beta}_{t_n} + \boldsymbol{\beta}_{p_n} + \boldsymbol{\beta}^f_{B_{n-1},tp_{n-1}} + \boldsymbol{\beta}^b_{B_n,tp_n} + \boldsymbol{\mu},\mathbf{R}) \quad \text{for } 1 \le n \le N. \tag{6}$$

It is noted that the affection from **l** is assumed to be implicitly included in the affection of **p** and hence is neglected. We also note that the coarticulation effect is elegantly treated to consider different degrees of coupling between two neighboring syllables via letting it depend on the break type of the syllable juncture.

The break-acoustics model $P(\mathbf{pd},\mathbf{ed}|\mathbf{B},\mathbf{p},\mathbf{l},\mathbf{t})$ is further elaborated via assuming that $(\mathbf{pd},\mathbf{ed})$ is independent of $(\mathbf{p},\mathbf{t})$ which mainly carries information of prosodic constituents rather than that of syllable juncture. So we have

$$P(\mathbf{pd},\mathbf{ed}|\mathbf{B},\mathbf{p},\mathbf{l},\mathbf{t}) \approx P(\mathbf{pd},\mathbf{ed}|\mathbf{B},\mathbf{l}) \approx \prod_{n=1}^{N-1} P(\mathrm{pd}_n,\mathrm{ed}_n|B_n,\mathbf{l}_n), \tag{7}$$

where pd_n and ed_n are the pause duration and energy-dip level of the juncture following syllable n (referred to as *juncture n* hereafter) and $\mathbf{l}_n$ is the contextual linguistic feature vector around juncture n. For mathematically tractable, $P(\mathrm{pd}_n,\mathrm{ed}_n|B_n,\mathbf{l}_n)$ is further simplified and realized by the product of a gamma distribution for pause duration and a normal distribution for energy-dip level:

$$P(\mathrm{pd}_n,\mathrm{ed}_n|B_n,\mathbf{l}_n) = g(\mathrm{pd}_n;\alpha_{B_n,\mathbf{l}_n},\beta_{B_n,\mathbf{l}_n})N(\mathrm{ed}_n;\mu_{B_n,\mathbf{l}_n},\sigma^2_{B_n,\mathbf{l}_n}). \tag{8}$$

In this study, $g(\mathrm{pd}_n;\alpha_{B_n,\mathbf{l}_n},\beta_{B_n,\mathbf{l}_n})$ and $N(\mathrm{ed}_n;\mu_{B_n,\mathbf{l}_n},\sigma^2_{B_n,\mathbf{l}_n})$ are concurrently generated by the decision tree method,[58] for each break type.

Similarly, we simplify the general prosody-syntax model $P(\mathbf{B},\mathbf{p}|\mathbf{l},\mathbf{t})$ via assuming the independency of $(\mathbf{B},\mathbf{p})$ and **t**, and decomposing it into two models, i.e.,

$$P(\mathbf{B},\mathbf{p}|\mathbf{l},\mathbf{t}) \approx P(\mathbf{B},\mathbf{p}|\mathbf{l}) = P(\mathbf{p}|\mathbf{B},\mathbf{l})P(\mathbf{B}|\mathbf{l}) \approx P(\mathbf{p}|\mathbf{B})P(\mathbf{B}|\mathbf{l}), \tag{9}$$

where $P(\mathbf{p}|\mathbf{B})$ is a prosodic-state model describing the dynamics of **p** given with **B** and $P(B_n|\mathbf{l}_n)$ is a break-syntax model describing the relationship between **B** and the contextual linguistic feature sequence **l**. In this study, we realize $P(\mathbf{p}|\mathbf{B})$ by a Markov model:

$$P(\mathbf{p}|\mathbf{B}) \approx P(p_1)\left[\prod_{n=2}^{N} P(p_n|p_{n-1},B_{n-1})\right], \tag{10}$$

where $P(p_1)$ is the initial prosodic-state probability for syllable 1 and $P(p_n|p_{n-1},B_{n-1})$ is the prosodic-state transition probability from syllable $n-1$ to syllable n given B_{n-1}. We also simplify $P(\mathbf{B}|\mathbf{l})$ by separately modeling it for each syllable juncture:

$$P(\mathbf{B}|\mathbf{l}) = \prod_{n=1}^{N-1} P(B_n|\mathbf{l}_n). \tag{11}$$

Here $P(B_n|\mathbf{l}_n)$ is implemented by the decision tree method.[58]

B. Joint prosody labeling and modeling

A sequential optimization procedure based on the maximum likelihood (ML) criterion is proposed to jointly label the prosodic tags for all utterances of the training corpus and to estimate the parameters of the four prosodic models. It is

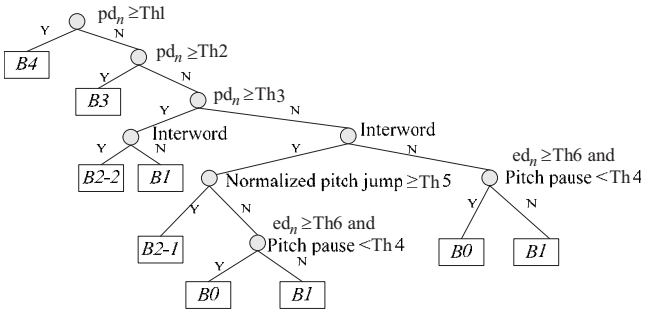

FIG. 2. The decision tree for initial break-type labeling.

divided into two main parts: *initialization* and *iteration*. The initialization part determines initial prosodic tags of all utterances and estimates initial parameters of the four prosodic models by a specially designed procedure. The iteration part first defines an objective likelihood function for each utterance by

$$Q = \left(\prod_{n=1}^{N} P(\mathbf{sp}_n | p_n, B_{n-1}^{n}, t_{n-1}^{n+1}) \right) \times \left(P(p_1) \prod_{n=2}^{N} P(p_n | p_{n-1}, B_{n-1}) \right) \times \left(\prod_{n=1}^{N-1} (P(\mathrm{pd}_n, \mathrm{ed}_n | B_n, \mathbf{l}_n) P(B_n | \mathbf{l}_n)) \right). \quad (12)$$

It then applies a multistep iterative procedure to update the labels of prosodic tags and the parameters of the four prosodic models sequentially and iteratively. In Secs. III B 1 and III B 2, we discuss the sequential optimization procedure in detail.

1. Initialization

The initialization part is further divided into two subparts: (a) a specially designed procedure to determine initial break labels of all syllable junctures and (b) a ML estimation process to estimate initial parameters of the four prosodic models and to determine the initial prosodic-state labels of all syllables using the information of initial break labels determined in the first subpart.

a. Initial labeling of break indices The initial break index of each syllable juncture is determined by a decision tree (see Fig. 2) designed based on a prior knowledge about break labeling/modeling gained in previous studies.[7,26,38,49–55,59,60] It is known that pause duration is the most important acoustic cue to specify breaks. Most word junctures with PM have long pauses so that they are most likely labeled as major break, or in our case *B*3 and *B*4. On the other hand, most intraword syllable junctures have very short pause duration so that they are generally labeled as nonbreak, or in our case *B*0 and *B*1. Moreover, *B*0 represents tightly coupled syllable juncture so that it is distinguished from *B*1 by having very short pitch pause duration and high energy-dip level. In-between these extreme situations, non-PM interword junctures with medium pause duration and with medium pitch jump are likely labeled as *B*2-2 and *B*2-1, respectively. By using the prior knowledge, we develop the algorithms to determine all thresholds of the decision tree (Th1–Th6) in a systematic way to avoid doing it manually or by trial and error. Detail of the algorithm is given in the Appendix.

b. Estimation of the initial parameters of the four prosodic models and prosodic-state indices The initializations of the break-acoustics model and the break-syntax model can be done independently with initial break indices of all syllable junctures being given. We realize them by the CART algorithm.[58] For the initialization of the break-acoustics model, the CART algorithm with the node splitting criterion of maximum likelihood gain is adopted to classify pause duration pd_n and energy-dip level ed_n for each break type B according to a question set Θ_1 derived from the contextual linguistic features $\mathbf{l}_n$. Each leave node represents the product of a gamma distribution $g(\mathrm{pd}_n; \alpha_{B,\mathbf{l}_n}, \beta_{B,\mathbf{l}_n})$ and a normal distribution $N(\mathrm{ed}_n; \mu_{B,\mathbf{l}_n}, \sigma^2_{B,\mathbf{l}_n})$. For the initialization of the break-syntax model $P(B_n|\mathbf{l}_n)$, a decision tree is built by using another question set Θ_2 derived also from $\mathbf{l}_n$ to classify break types.

The initializations of the syllable pitch contour model and prosodic-state indices are integrated together and performed by a progressive estimation procedure. Since the syllable pitch contour model is a multiparametric representation model to superimpose several APs of major affecting factors to form the surface syllable pitch contour, the estimation of an AP may be interfered by the existence of the APs of other types. It is therefore improper to estimate all initial parameters independently. We hence adopt a progressive estimation strategy to first determine the initial APs which can be estimated most reliably and then eliminate their affections from the surface pitch contours for the estimations of the remaining APs. In this study, the order of initial AP estimation is listed as follows: global mean $\boldsymbol{\mu}$, five tones $\boldsymbol{\beta}_t$, coarticulation $\{\boldsymbol{\beta}^f_{B,tp}, \boldsymbol{\beta}^b_{B,tp}, \boldsymbol{\beta}^f_{B_b,t_1}$, and $\boldsymbol{\beta}^b_{B_e,t_N}\}$, and prosodic states $\boldsymbol{\beta}_p$. Notice that the initial prosodic-state indices are assigned by vector quantization (VQ) of the pitch-level components of the residue pitch contours, and the APs are set to be the codewords obtained by VQ. Lastly, the initialization of the prosodic-state model $P(\mathbf{p}|\mathbf{B})$ is done using the labeled prosodic-state indices and break indices.

2. Iteration

The iteration is a multistep iterative procedure listed below.

Step 1. Update the APs of five tones $\boldsymbol{\beta}_t$ with all other APs being fixed.

Step 2. Update the APs of coarticulation $\{\boldsymbol{\beta}^f_{B,tp}, \boldsymbol{\beta}^b_{B,tp}, \boldsymbol{\beta}^f_{B_b,t_1}$, and $\boldsymbol{\beta}^b_{B_e,t_N}\}$ with all other APs being fixed, and then update $\mathbf{R}$.

Step 3. Relabel the prosodic-state sequence of each utterance by using the Viterbi algorithm so as to maximize Q defined in Eq. (12). Then, update the APs of prosodic-state $\boldsymbol{\beta}_p$, the prosodic-state model $P(\mathbf{p}|\mathbf{B})$, and $\mathbf{R}$.

Step 4. Relabel the break-type sequence of each utterance by using the Viterbi algorithm so as to maximize Q. Then, update the prosodic-state model $P(\mathbf{p}|\mathbf{B})$ and $\mathbf{R}$.

Step 5. Reconstruct the decision trees to update $P(\mathrm{pd}_n, \mathrm{ed}_n|B_n, \mathbf{l}_n)$ and $P(B_n|\mathbf{l}_n)$ by the CART algorithm using the question sets Θ_1 and Θ_2, respectively.

Step 6. Repeat Steps 1–5 until a convergence is reached.

IV. EXPERIMENTAL RESULTS

The proposed method was evaluated using an unlabeled Mandarin speech database. The database contained read speech of a female professional announcer. Its texts were all short paragraphs composed of several sentences selected from the Sinica Treebank corpus.[61] The database consisted of 380 utterances which contained in total 52 192 syllables. In this experiment, the number of prosodic states was properly set to be 16 because the root mean squared error (RMSE) of VQ saturated when the number of prosodic states was greater than 16. The sequential optimization procedure took 69 iterations to reach a convergence. Following is the presentation of the analyses, discussion, and findings of our experiment, which is arranged in the order that an examination and interpretation of the parameters of the four prosodic models was introduced in Secs. IV A–IV D, then, to evaluate the performance of the models proposed, explorations in the relationships between prosodic breaks and linguistic features of texts, the length of prosodic constituents, and the general pitch patterns of prosodic constituents obtained in our method were described in Secs. IV E–IV G, and, finally, to further verify the labeling outcomes generated by our models, a comparison conducted between human labeling and our labeling was given in Secs. IV H and IV I.

A. The syllable pitch contour model

We first examined the parameters of the syllable pitch contour model $P(\mathbf{sp}_n|p_n,B_{n-1}^n,t_{n-1}^{n+1})$. The covariance matrices of the original and normalized syllable log $F0$ contour feature vectors are shown below:

$$\mathbf{R}_{\mathbf{sp}} = \begin{bmatrix} 883.7 & 23.9 & -25.6 & -0.5 \\ 23.9 & 90.5 & 9.7 & -8.2 \\ -25.6 & 9.7 & 17.8 & -0.9 \\ -0.5 & -8.2 & -0.9 & 5.0 \end{bmatrix} \times 10^{-4} \Rightarrow \mathbf{R}_{\mathbf{sp}^r}$$

$$= \begin{bmatrix} 3.5 & 0.2 & -0.2 & 0.0 \\ 0.2 & 31.9 & 2.6 & -1.5 \\ -0.2 & 2.6 & 11.1 & 0.6 \\ 0.0 & -1.5 & 0.6 & 3.7 \end{bmatrix} \times 10^{-4}.$$

Obviously, all elements of $\mathbf{R}_{\mathbf{sp}^r}$ were much smaller than those of $\mathbf{R}_{\mathbf{sp}}$. This showed that the influences of the affecting factors considered were indeed essential to the variation in **sp**.

Figure 3 displays the APs of five tones. We find from the figure that the APs of the first four tones conformed well to the standard tone patterns found by Chao.[62] As for tone 5, its low dipping pattern resembles the pattern of tone 3 to some degree. This also matched the finding in the previous study about tone 5.[63]

Table I displays the APs (log $F0$ levels) and the distribution of the 16 prosodic states. It can be seen from Table I that these log $F0$ levels spanned widely to cover the whole dynamic range of log $F0$ variation with lower indices of prosodic state corresponding to lower log $F0$ levels, and the prosodic states distributed normally with relatively few located at the two extremes of high and low prosodic states.

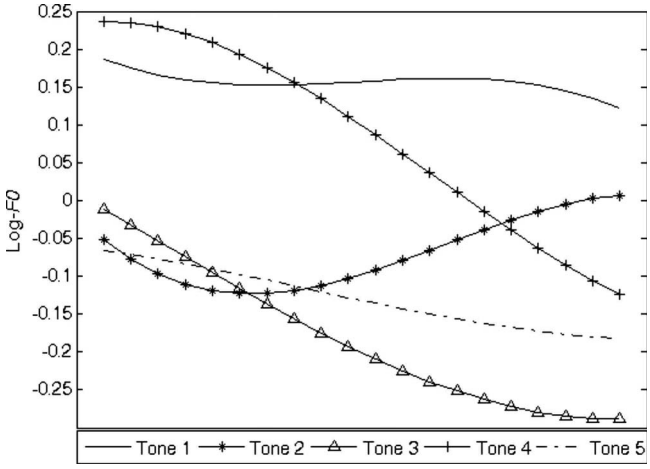

FIG. 3. The APs of five tones.

Figures 4(a) and 4(c) display the APs of forward and backward coarticulations, $\boldsymbol{\beta}_{B,tp}^f$ and $\boldsymbol{\beta}_{B,tp}^b$, for the three break types of $B0$, $B1$, and $B4$. These three break types were chosen on purpose to show extreme cases of intersyllable coarticulation: $B0$ for tightly coupling, $B1$ for normal coupling, and $B4$ for no coupling. Some interesting phenomena can be observed from the figure. First, it can be seen from Fig. 4(a) that most APs of forward coarticulation for $B0$ and $B1$, $\boldsymbol{\beta}_{B0,tp}^f$ and $\boldsymbol{\beta}_{B1,tp}^f$, were bended in their beginning parts. These bendings were to compensate the level mismatch between the beginning and ending parts of the log $F0$ contours of the tone pairs for highly coarticulated preceding and current syllables, so as to make their log $F0$ contours be concatenated more smoothly. For example, the upward bending at the beginning parts of $\{\boldsymbol{\beta}_{B,tp}^f|tp=(1,2),(1,3),(2,2),(2,3),(1,5)\}$ were due to *H-L* mismatches, while the downward bending at the beginning parts of $[\boldsymbol{\beta}_{B,tp}^f|tp=(3,1),(3,4),(5,1),(5,4),(4,1)(4,4)]$ corresponded to *L-H* mismatches. Similarly, it can be observed from Fig. 4(c) that the ending parts of the APs of backward coarticulation for $B0$ and $B1$, $\boldsymbol{\beta}_{B0,tp}^b$ and $\boldsymbol{\beta}_{B1,tp}^b$, were bended. But the degrees of their upward and downward bendings were generally smaller. This conformed to the observation reported in Ref. 64 that the carry-over effect on the syllable $F0$ contour influenced by the preceding syllable is much larger than the anticipation effect caused by the following syllable. Second, it can be found from Figs. 4(a) and 4(c) that most APs of forward and backward coarticulations for $B4$ with the same current tone looked similar and hence were nearly independent of their respective preceding and succeeding tones. This showed that the intersyllable coarticulation across a $B4$ break was relatively low as compared with those of $B0$ and $B1$. Moreover, many APs of forward and backward coarticulations for $B4$ were downward bended in their beginning and ending parts, respectively. They exhibited the onset and offset phenomena at the beginning and ending syllables of BG/PG. Furthermore, we find from Figs. 4(b) and 4(d) that most utterance initial and final patterns, $\boldsymbol{\beta}_{B_b,t}^f$ and $\boldsymbol{\beta}_{B_e,t}^b$, looked very similar to those of $\boldsymbol{\beta}_{B4,tp}^f$ and $\boldsymbol{\beta}_{B4,tp}^b$, respectively, to show the same onset and offset phenomena at the two types of utterance boundaries. We also find that $\boldsymbol{\beta}_{B_e,3}^b$ and $\boldsymbol{\beta}_{B_e,5}^b$ were two exceptional patterns which

TABLE I. The APs [log $F0$ levels, $\boldsymbol{\beta}_p(1)$] and the distribution [$P(p)$] of the 16 prosodic states.

State index p	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$\boldsymbol{\beta}_p(1)$	−0.77	−0.50	−0.37	−0.28	−0.22	−0.16	−0.10	−0.05	0.01	0.06	0.12	0.17	0.24	0.31	0.38	0.49
$P(p)$	0.00	0.01	0.02	0.04	0.07	0.10	0.11	0.12	0.12	0.10	0.09	0.08	0.06	0.05	0.03	0.01

had lower levels. These probably resulted from the total relaxation of pronunciation at the utterance ending for these two tones. Third, it can be found from Fig. 4(c) that the APs of $\boldsymbol{\beta}^b_{B0,(3,3)}$ and $\boldsymbol{\beta}^b_{B1,(3,3)}$ were upward bended drastically in their ending parts. As combining with the AP of tone 3 shown in Fig. 3, these bendings would make the integrated log $F0$ patterns of the first syllable in a (3,3) tone pair change from middle-falling tone-3 shape to middle-rising tone-2 shape to fulfill the well-known 3-3 tone *sandhi* rule which says that the first tone 3 of a 3-3 tone pair will change to a tone 2. On the contrary, we find that the pattern $\boldsymbol{\beta}^b_{B4,(3,3)}$ did not bend upward. This showed that the 3-3 tone sandhi rule did not apply when the syllable juncture was a $B4$. Lastly, we made some comments to the APs of forward and backward coarticulations for $B2$-1, $B2$-2, and $B3$. Basically, the APs of $B2$-1 and $B3$ resembled to those of $B4$ but with smaller upward and downward bendings, and $B2$-2 had similar patterns to those of $B1$ but with smaller upward and downward bendings.

From above analyses, we find that the inferred syllable pitch contour model provides a meaningful interpretation to the variation in syllable pitch contour controlled by several major affecting factors. With this capability, the model can be used in Mandarin TTS to generate pitch contour if all tags of prosodic-state and break type can be properly predicted from the input text. It can also be used in Mandarin ASR to manipulate pitch information for tone discrimination.

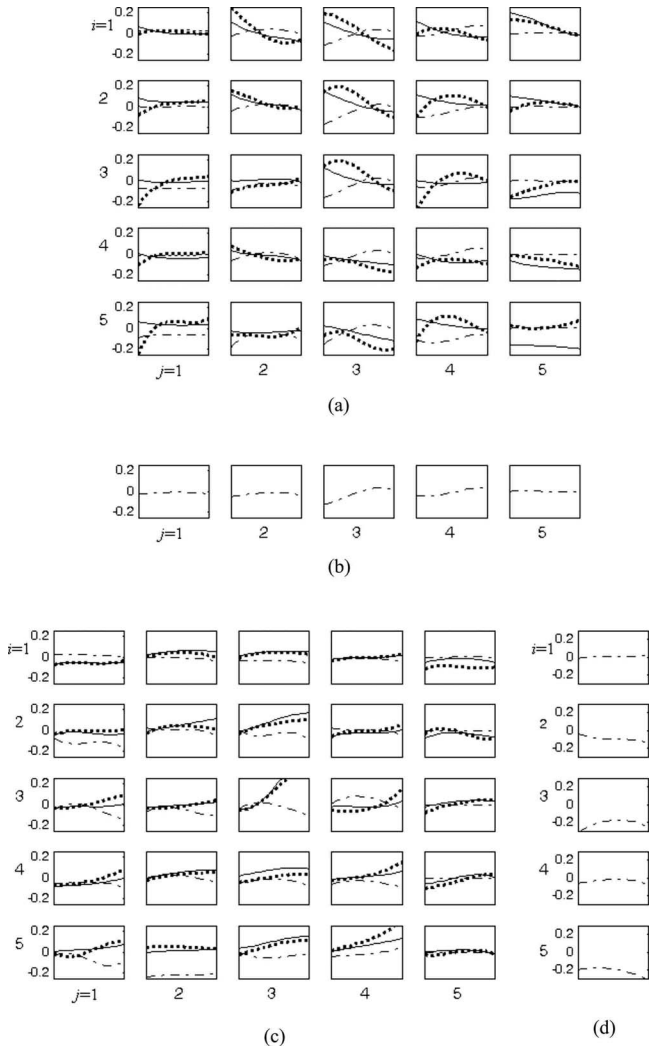

FIG. 4. The APs of (a) forward and (c) backward coarticulations, $\boldsymbol{\beta}^f_{B,ip}$ and $\boldsymbol{\beta}^b_{B,ip}$, for $B0$ (point line), $B1$ (solid line), and $B4$ (dashed line); and the APs of (b) utterance onset and (d) utterance offset, $\boldsymbol{\beta}^f_{B_b,i}$ and $\boldsymbol{\beta}^b_{B_e,i}$, for B_b and B_e. Here $tp=(i,j)$ and $t=j$ or i.

B. The break-acoustics model

The two break-acoustics models, $g(\mathrm{pd}_n;\alpha_{B_n,\mathbf{l}_n},\beta_{B_n,\mathbf{l}_n})$ and $N(\mathrm{ed}_n;\mu_{B_n,\mathbf{l}_n},\sigma^2_{B_n,\mathbf{l}_n})$, were built by the decision tree method using the question set Θ_1. One decision tree was constructed for each break type. Figure 5 displays the distributions of pause duration and energy-dip level for the root nodes of these six break types. It can be found from the

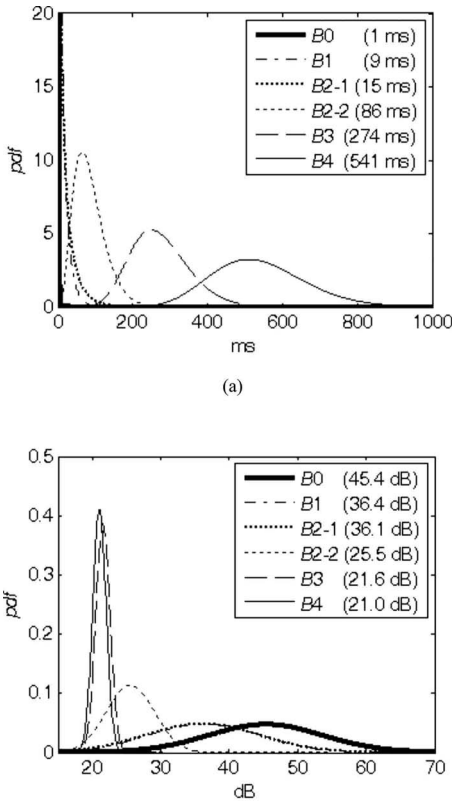

FIG. 5. The pdfs of (a) pause duration and (b) energy-dip level for the root nodes of these six break types. Numbers in ( ) denote the mean values.

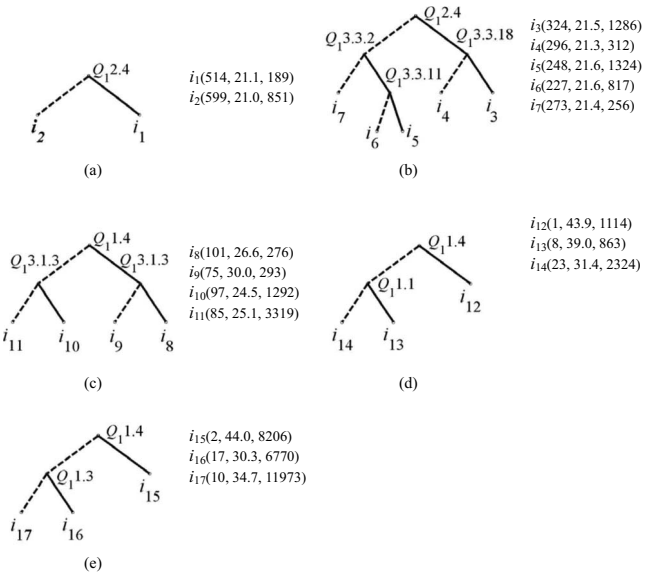

FIG. 6. The decision trees of the break-acoustics model for (a) *B*4, (b) *B*3, (c) *B*2-2, (d) *B*2-1, and (e) *B*1. The numbers in a bracket denote average pause duration in milliseconds (left), energy-dip level in decibels (middle), and sample count (right) of the associated node. Solid line indicates positive answer to the question and dashed line indicates negative answer.

figure that the break types of higher level were generally associated with longer pause duration and lower energy-dip level. *B*0 had very short pause duration and widespread energy-dip level with very high mean value. *B*1 and *B*2-1 had similar distributions of short pause durations and widespread high energy-dip level. *B*2-2 had medium long pause duration and medium high energy-dip level. Both *B*3 and *B*4 had widespread long pause duration and low energy-dip level. These conformed to the prior knowledge about break types.[4–7]

To further examine the model, we show its decision trees for the five break types of *B*4, *B*3, *B*2-2, *B*2-1 and *B*1 in Fig. 6. It is noted here that no tree split for *B*0 due to the relative uniformity on the acoustic prosodic features of its samples. Generally, the questions used to split trees of higher-level break types (*B*4 and *B*3) tended to be related to higher-level syntactic features, such as PM ($Q_1$2.4) and syntactic phrase size ($Q_1$3.1.3, $Q_1$3.3.2, $Q_1$3.3.11, and $Q_1$3.3.18). On the contrary, the questions of lower-level phonetic features ($Q_1$1.1, $Q_1$1.3, and $Q_1$1.4) tended to split trees of lower-level break types (*B*1 and *B*2-1).

From above discussions, we find that the inferred break-acoustics model describes the relationship of the break type of syllable juncture with the two intersyllable acoustic features and some contextual linguistic features very well. So it seems that the model can be used to predict major and minor breaks from acoustic and linguistic cues for some applications, such as segmenting speech into sentences and generation of punctuations from speech.

C. The prosodic-state model

We then examined the prosodic-state model. Figure 7 displays some most significant transitions of $P(p_n|p_{n-1},B_{n-1})$ for six break types. For *B*0 and *B*1, the general high-to-low, nearby-state transitions showed that the syllable log *F*0 level

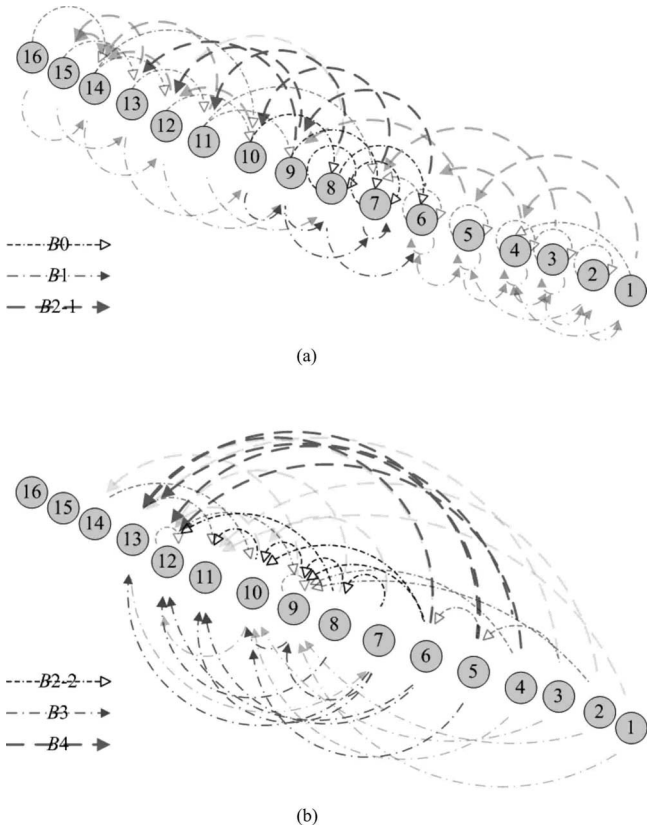

FIG. 7. The most significant prosodic-state transitions for (a) *B*0, *B*1, and *B*2-1, and (b) *B*2-2, *B*3, and *B*4. Here, the number in each node represents the index of the prosodic state. Note that bold and thin lines denote the primary and secondary state transitions, respectively.

declined slowly within PWs. We also find that some low-to-high, nearby-state transitions occurred within PWs of low pitch level. This demonstrated the sustaining phenomenon of the log *F*0 trajectory at the ending part of some PPhs. For *B*2-2, it had both high-to-low and low-to-high state transitions. For *B*2-1, *B*3, and *B*4, their low-to-high state transitions showed clearly the phenomena of syllable log *F*0 level resets across PWs, PPhs, and BG/PGs. Compared with these clear log *F*0 level resets, the resets of *B*2-2 were insignificant. Combining the results shown in Figs. 5 and 7, we find that *B*2-1 and *B*2-2 had different acoustic characteristics: *B*2-1 had significant log *F*0 reset with very short pause duration, while *B*2-2 had longer pause duration with low or no log *F*0 reset.

From above findings, since the prosodic states defined in our study mainly carry the full information of pitch-level variation in the upper three layers of prosodic structure (PW, PPh, or BG/PG), the prosodic-state model can roughly represent dynamic patterns of PW, PPh, and BG/PG and may be applied to pitch contour generation in Mandarin TTS.

D. The break-syntax model

The break-syntax model $P(B_n|\mathbf{l}_n)$ was built by the decision tree method using the question set Θ_2. Figure 8 displays the decision tree of the break-syntax model. The tree was divided into four subtrees, *T*3–*T*6, by the three questions of $Q_2$2.1.1 (PM?), $Q_2$2.1.3 (minor PM?), and $Q_2$1.3 (intra-

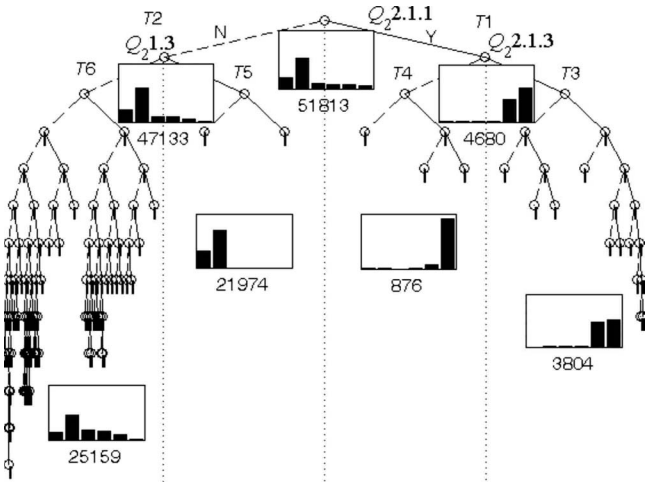

FIG. 8. The decision tree of the break-syntax model. The bar plot associated with a node denotes the distributions of these six break types ($B0$, $B1$, $B2$-1, $B2$-2, $B3$, and $B4$, from left to right) and the number is the total sample count of the node.

word?). It can be seen from the figure that the root node of subtree $T3$, which corresponded to syllable juncture with minor PM, was mainly composed of $B3$ and $B4$. Similarly, the root nodes of subtrees $T4$ and $T5$, corresponding to major PM and intraword syllable juncture, were mainly composed of $B4$ and $B0/B1$, respectively. Due to the fact that the break-type constituents of both $T4$ and $T5$ were pure, they had very simple tree structures. On the contrary, subtree $T6$ was a miscellaneous collection of all other types of syllable juncture without PM. So, it had the most complex tree structure.

Figure 9 displays the more detailed structures of these four subtrees up to the fourth layer. From Figs. 9(a) and 9(b), we find that nodes in $T3$ and $T4$ were mainly split by questions related to high-level linguistic features such as $Q_2 3.3.19$ (Is the length of the following syntactic phrase/sentence greater than 6?) and $Q_2 3.3.29$ (Is the length of the preceding syntactic phrase/sentence greater than 7?). As shown in Fig. 9(c), $T5$ had two leaf nodes split by $Q_2 1.1$ (Does the following syllable have a null initial or initial in $\{m,n,l,r\}$?). The set associated with positive answer was mainly composed of $B0$, while another set was mainly composed of $B1$. As shown in Fig. 9(d), $T6$ was constructed by questions related to features of various levels, including $Q_2 1.1$, $Q_2 2.4.18$ (Is the preceding word "DE"?), $Q_2 2.3.2$ (Is the preceding word a function word?), $Q_2 3.3.24$ (Is the length of the preceding syntactic phrase greater than 2?), and so on. We also find from Fig. 9 that the purities of the break-type constituents were high for leaf nodes of $T4$ and $T5$,

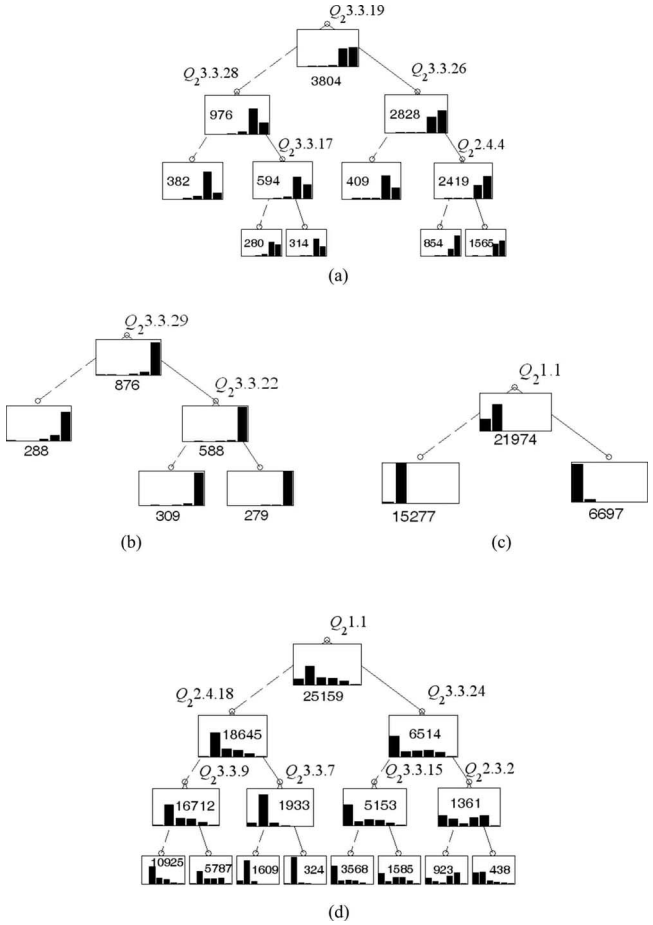

FIG. 9. The more detailed structures of subtrees of (a) $T3$, (b) $T4$, (c) $T5$, and (d) $T6$. Solid line indicates positive answer to the question and dashed line indicates negative answer.

medium high for nodes of $T3$, and relatively low for most nodes of $T6$. This implies that it is difficult to correctly label (or predict) the break types of syllable junctures other than intraword and those with major PM by the break-syntax model using only linguistic features without the help of acoustic cues.

E. Analyses of the labeled break types

Since the purpose of announcers' broadcasting is to propagate information accurately to the audience relying exclusively on their audio perception, our well-trained informant skillfully manipulated as many segmental and prosodic cues as possible, such as clear and precise articulation, strategic variations in the fundamental frequency, volume, syllable length, and types of breaks. These prosodic information carried in the utterance speech, in turn, reflects the infor-

TABLE II. Statistics of break types labeled for 121 prefixes and 195 suffixes.

Labeled break type		$B0$	$B1$	$B2$-1	$B2$-2	$B3$	$B4$	Total count
Prefix	Preboundary	94	1289	460	545	193	5	2586
	Postboundary	584	1475	344	178	5	0	2586
Suffix	Preboundary	1046	2466	31	20	3	0	3566
	Postboundary	307	1479	272	482	568	458	3566

TABLE III. Statistics of break types labeled for the DE words.

Labeled break type	*B*0	*B*1	*B*2-1	*B*2-2	*B*3	*B*4	Total count
Preboundary	168	1600	146	1	0	0	1915
Postboundary	210	1035	331	294	41	4	1915

mant's mental grammar, her Mandarin linguistic competence that determines when to form a semantically appropriate word chunk, a PPh, or a larger unit, and hence where and how long a break in an utterance should be so that the informant's speech would sound natural, informative, and attention attracting to the audience.

As a research based on our informant's speech data rich in the Mandarin prosodic cues, our break-type-labeling model also can generate appropriate break types consistent with native speakers' psychological reality. To verify this point, we examined the relationship between some special groups of words/morphemes and their concurring break types that both our break-type-labeling model and the ordinary Mandarin native speakers would consistently produce. These special groups of words/morphemes include (1) affix morpheme; (2) DE; (3) Ng, Di, and T; (4) VE; (5) Caa and Cb; and (6) P.[65] The results are discussed in more detail as follows.

1. Set of affix morpheme

It is well known that prefixes and suffixes are bound morphemes that attach to their preceding or following heads to form units of complex words. Since the resultant form after combining the head and the affix is a unit, it is reasonable to predict that the breaks at the boundaries between the head and the affix tend to fall in *B*0 or *B*1 types. These phenomena were observed in our corpus. We found that some Mandarin Chinese monosyllabic prefixes, such as *bu-* "un-, dis-, in-," *ke-* "-able," *wu-* "un-, -less, without," etc.,[62,66] tend to join the following heads to form legitimate words as in *bu-li* "unfavorable," *bu-fang-bian* "inconvenient," *ke-wu* "detestable," *ke-sing* "feasible," *wu-sian* "limitless," and *wu-shuang* "unparalleled." Similarly, by attaching monosyllabic suffixes, such as *-bian* "side," *-zhe* "-er, -or," *-hua* "-ize," etc., to the preceding roots, we can derive complex words as in *lu-bian* "roadside, curb," *he-bian* "riverside," *zuo-zhe* "author, writer," *sing-zhe* "religious practicer," *gung-yie-hua* "industrialize," and *min-zhu-hua* "democratize."

Table II lists the statistics of the break types labeled for the syllable boundaries of 121 prefixes and 195 suffixes. It can be seen from the table that 79.6% of the postsyllable boundaries of these 121 prefixes and 98.5% of the presyllable boundaries of these 195 suffixes were labeled as *B*0 or *B*1. These prosodic findings reflect the fact that morphologically the combination of head and affix generates a lexical unit, and thus the break between them is determined to be the break type of intra-PW category by our method. The results were also consistent with some rules found in Refs. 59, 60, and 63.

2. Word set of DE

The words in the DE set particularly refer to *de*, *zhe*, and *di*, which serve multifunctions including a possessive marker, an adjective marker, and an adverbial marker.[65] They are characterized by the fact that a DE word can combine with a wide range of preceding syntactic constituents to form a possessive adjective as in a noun phrase (NP)-de structure: *xue-sheng-de quan-li* "students' right" to derive an adjective phrase as in a verb phrase (VP)-de structure: *se-siang-zhe qin* "nostalgia" or to function as an adverbial phrase as in a DM-de structure: *ke-ren yi-bo-bo-di yong-jin-dien-lai* "guest were flocking to the shop." Despite the variety of the preceding constituent, a DE word, similar to a suffix, builds closer connection with its preceding constituent to form a larger syntactic unit; consequently, it is predictable that the break at the DE words' preboundary position tends to fall into *B*0 and *B*1, which means a pause is hardly to be perceived at this juncture. It is also reasonable to infer that due to a looser connection between the DE words and the following constituent, less *B*0 and *B*1 would occur at the postboundary position.

The statistics in Table III indicates that the distribution of the break types labeled by our model just conformed with our anticipation; while 92.3% preboundary breaks of the DE words were *B*0 and *B*1, only 65% postsyllable boundaries of the DE words fell into the same types, which suggests that for the DE words, the majority of the neighboring breaks are unperceivable, and in most cases only at the postboundary

TABLE IV. Statistics of break types labeled for the word sets of Ng, Di, and T.

Labeled break type		*B*0	*B*1	*B*2-1	*B*2-2	*B*3	*B*4	Total count
Ng	Preboundary	97	420	19	12	0	2	550
	Postboundary	26	81	17	58	245	123	550
Di	Preboundary	107	83	12	1	0	2	205
	Postboundary	30	68	36	41	11	19	205
T	Preboundary	89	84	14	11	0	0	198
	Postboundary	0	5	1	2	22	168	198

TABLE V. Statistics of break types labeled for word set of VE.

Labeled break type	*B*0	*B*1	*B*2-1	*B*2-2	*B*3	*B*4	Total count
Postboundary	63	177	99	108	159	234	840

position can perceivable breaks be sensed. This result also matched the findings in Refs. 59, 60, and 63.

3. Word sets of Ng, Di, and T

Ng, Di, and T represent the word sets of Mandarin Chinese localizers, aspectual adverbs, and particles,[65] respectively. The distinctive shared feature of these sets of words is that almost all the words are no longer than two syllables in length and that when combining with other syntactic constituent to form a larger phrase, they are all positioned at the end of the derived phrase, such as *san-tian ho*$_{\mathrm{Ng}}$ "three days latter," *kai-hui dang-zhung*$_{\mathrm{Di}}$ "while the meeting is being held," and *bu-qu le-ma*$_{\mathrm{T}}$ "not going?." Due to the characteristic of being postpositioned in a phrase, these words are inclined to be incorporated with their preceding constituents, and predictably barely any pauses can be perceived at the preboundary position. The statistic results listed in Table IV indicate that our model's break-labeling performance just exactly met our expectation. As high as 94%, 93%, and 87% of the presyllable boundaries of the words in this category were labeled as *B*0 or *B*1.

On the other hand, it is also interesting to find that for the breaks at the postsyllable boundaries, 67% and 96% of them were labeled as *B*3/*B*4 especially for the Ng-set and T-set words, respectively. Further investigation reveals that most of the longer breaks were caused by a following PM, an index representing the occurrence of a detectable pause. Besides, because the T-set words are phrasal or sentential final particles and hence are highly likely to be followed by a PM, a much higher ratio of *B*3/*B*4 could be found.

4. Word set of VE

VE represents a class of transitive verbs that take a sentence as the object, such as *ren-wei* "to suppose/think/believe (that)," *gan-dao* "to feel (that)," *biao-she* "to show/indicate/mean/suggest (that)," etc.[65] It is evident that since the message carried in a sentential object, compared to a NP object for example, demands longer time to process mentally before being accurately expressed, a longer pause is reasonably anticipated to occur after a VE verb for information operation. Based on the statistic results listed in Table V, on the whole 72% postword boundaries of the VE verbs were labeled as breaks with distinctly audible pauses, namely, *B*2-1, *B*2-2, *B*3, or even *B*4, another quite favorable evidence that the break types labeled by our model were consistent with the pause duration people usually take in their utterances.

However, it cannot be neglected that no less than 28% postword boundaries of the VE verbs were labeled as *B*0 or *B*1, implying that seemingly our model still generated quite a few unexpected break types for the VE verbs. Further observation of the data, nevertheless, found two main reasons to account for this discrepancy of labeling. First, besides a sentential object, part of the VE verbs could also take a NP object, so the breaks occurring before a NP object were predictably shorter than before a sentential object. The other reason for the occurrence of *B*0/*B*1 after a VE verb is that to express attitudinal, temporal, spatial, or manner information about a VE verb, a small word from the DE, Di, Ng, or T sets (such as *de*, *zhe*, *le*, *guo*, etc.) was attached to the verb, and this attachment and the close connection between the small word and the VE verb caused no need to pause at the juncture. However, the originally expected long pause (*B*3/*B*4) after the VE verb did not actually disappear; it was retained and only lagged behind to occur after the VE verb.

5. Word sets of Caa and Cb

Caa and Cb are two subcategories of Mandarin conjunctions, representing conjunctive conjunctions and correlative conjunctions,[65] respectively. In the case of Caa, the arguments linked by the Caa conjunctions are words or phrases of identical syntactic categories and are usually associated in their meaning as in *feng*$_{\mathrm{N}}$ he$_{\mathrm{Caa}}$ *yu*$_{\mathrm{N}}$ "wind and rain," *re*$_{\mathrm{VH}}$ *hia-shi*$_{\mathrm{Caa}}$ *leng*$_{\mathrm{VH}}$ "hot or cold," *si*$_{\mathrm{Neu}}$ *zhi*$_{\mathrm{Caa}}$ *shi*$_{\mathrm{Neu}}$ *sui*$_{\mathrm{Nf}}$ "from four to ten years old," and the like. Upon observation, we found that people usually tend to take a longer pause at preword boundary than at the postword context, forming a sensible rhythmic variation and hence facilitating message delivery. The statistics of the labeling results in Table VI informs us that 90% of the Caa preboundary breaks were not shorter than *B*2-2, while, on the contrary, 98% of the postword breaks were not longer than *B*2-2, a labeling outcome verifying our observation of the Caa words' neighboring breaks; that is, longer pauses tended to occur at the boundary between the preceding argument and the conjunction. The results matched some findings in Ref. 38.

On the other hand, the Cb conjunctions function to join

TABLE VI. Statistics of break types labeled for word sets of Caa and Cb.

Labeled break type		*B*0	*B*1	*B*2-1	*B*2-2	*B*3	*B*4	Total count
Caa	Preboundary	5	32	1	127	214	26	405
	Postboundary	52	104	157	85	7	0	405
Cb	Preboundary	61	46	23	39	168	512	849
	Postboundary	135	284	166	95	150	19	849

TABLE VII. Statistics of break types labeled for word sets of P07 and P21.

Labeled break type		*B*0	*B*1	*B*2-1	*B*2-2	*B*3	*B*4	Total count
P07	Preboundary	0	39	0	9	32	9	89
	Postboundary	8	38	34	9	0	0	89
P21	Preboundary	1	168	12	24	88	53	346
	Postboundary	27	79	208	28	4	0	346

two clauses—a syntactic unit much larger than Caa's arguments—into a compound sentence, and therefore have higher potential to be preceded or followed by a PM in written texts to delimit the domain of a clause or a sentence; in read speech the occurrence of a PM elicits the announcer to take a longer pause to index a message transition or a piece of new message is coming. Our statistic results show that in the case of Cb conjunctions 80% of the preword boundaries and 20% of the postword boundaries were labeled as *B*3/*B*4, which means much more PMs occurred before Cb conjunctions than afterward.

6. Word set of P

P represents the class of Chinese prepositions, which precede a required argument and together play several semantic roles and indicate various relationships such as time, location, tool, purpose, etc. Although Chinese Knowledge and information Processing (CKIP) categorizes prepositions into 65 types,[65] only 13 types are active in the Sinica Treebank corpus. As for the adjacent pause of a preposition, it is reasonable to expect that due to the close connection of a preposition and its following argument, the pause at the postword boundary tends to be short. For convenience of illustration, only *ba/jiang* (labeled as P07) and *zai* (labeled as P21), two typical and most frequently used prepositions, are selected out as the representative examples for discussion.

The statistic results in Table VII show that on the whole for both *ba/jiang* and *zai* about 90% of the postword boundaries were labeled as breaks no longer than *B*2-1 (a break type caused by a pitch jump instead of lengthened pause duration), which indicates that the pauses at this juncture were either unperceivable or tending to be very short, again another confirmation of our model's sound labeling job. Besides, a closer look at the distribution of break-type percentages reveals that as high as 49% and 69% of the postword breaks were *B*2-1 for *ba/jiang* and *zai*, respectively. This statistics reflected our informant's idiosyncratic style of articulating prepositional phrase; namely, besides leaving no pauses, she often made a pitch jump between a preposition and the following argument to cause a sensible short pause.

On the other hand, as far as the labeling at the preword boundary is concerned, most labels were either *B*1 or *B*3/*B*4; that is, 46% and 41% of the labels were *B*3/*B*4 and 44% and 49% of them were *B*1 for *ba/jiang* and *zai*, respectively, which suggests that our informant either took quite a long pause or just no pause at the preword position. To explain this phenomenon, further examination on the data containing these two prepositions revealed that the informant's long breaks (*B*3/*B*4) before a preposition were contributed by a left PM, and in the remained cases she usually took no pause at this position.

F. Analyses of prosodic constituents

Based on the break-type labeling, we can divide the syllable sequence of each utterance into three types of prosodic constituents (i.e., PW, PPh, and BG/PG) to form a four-layer prosodic structure. Statistics in Table VIII shows that the average lengths for these three types of prosodic constituents are, respectively, 3.17 syllables or 1.85 lexical words (LWs) for PWs; 6.98 syllables, 4.02 LWs, or 1.69 PWs for PPhs; and 16.69 syllables, 9.62 LWs, 4.07 PWs, or 1.94 PPhs for BG/PGs.

According to the histograms displayed in Fig. 10, the length of each of these three prosodic constituents spans, respectively, from 1 to 12 syllables for PWs, from 1 to 33 syllables for PPhs, and from 1 to 99 syllables for BG/PGs. Besides, the histograms also reveal that quite a few PPhs and BG/PGs, whose average lengths are supposed to be about 6.98 and 16.69 syllables, respectively, are nevertheless no longer than three syllables in length. Further investigation into these oddly short PPhs and BG/PGs indicates that the main reason lies in several special structure patterns of these constituents that require a long pause to highlight their prominence for successful information processing. First of all, in the case of short BG/PGs, defined as a sequence of syllables bounded by a *B*4 on both sides, many of the particularly short BG/PGs, actually consisted of a monosyllabic subject and VE verb, which, as discussed in Sec. IV E 4, due to its sentential object was tending to be followed by a long break up to *B*4; accordingly, bounded by a *B*4 on both sides, the structure pattern of a subject plus a VE verb, both monosyllabic in length, could generate as many short BG/PGs, as possible.

As for the cases of short PPhs, defined as a sequence of syllables delimited by (1) a *B*3 at both sides or (2) a *B*3 and a *B*4 at each side, respectively, most of the *B*3s or *B*4s bounding the very short PPhs were actually caused by the

TABLE VIII. Statistics of three types of prosodic constituents. Value in parentheses denotes standard deviation.

Average length in	Prosodic constituent		
	PW	PPh	BG/PG
Syllable	3.17(1.74)	6.98(3.48)	16.69(9.49)
Lexical word	1.85(1.03)	4.01(2.17)	9.62(5.43)
PW	1.00	1.69(1.55)	4.07(2.90)
PPh	X	1.00	1.94(1.75)

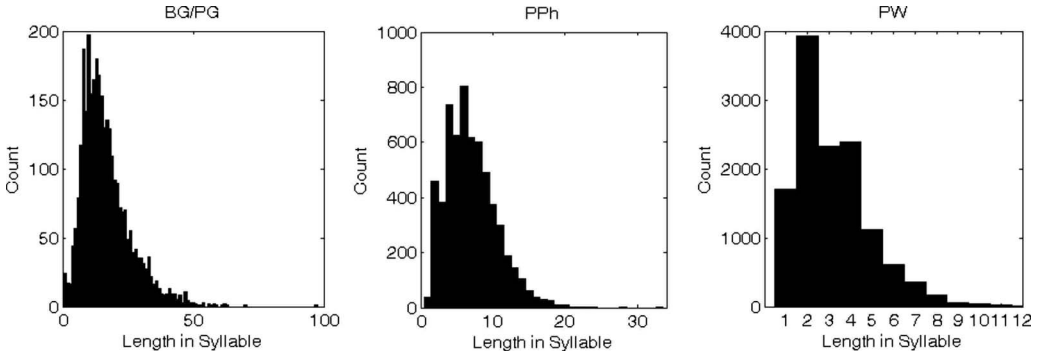

FIG. 10. Histograms of lengths for BG/PG, PPh, and PW.

existence of PMs that cued long pause duration. Table IX shows the statistic results of the short PPh instances with respect to the existence of PMs at their two endings. As shown in the table, 66% of one-syllable PPhs were bounded by PMs on both sides, and most of them were numbers that were used to enumerate events. On the other hand, in the case of two- or three-syllable PPhs, on the whole about 84% of them were delimited at least by a left-sided PM, which means that the majority of these PPhs occurred at the beginning of a sentence. In terms of the internal structure of the two-syllable PPhs, 91% of them were bisyllabic LWs functioned to express transitional relationships like contrast, comparison, reinforcement, or addition. As for the three-syllable PPhs, their structures were either a topicalized trisyllabic noun or any phrasal structure composed of two smaller syntactic elements as in a subject-VE structure (*wo ren-wui* "I suppose"), a preposition-noun structure (*you Chung-qing* "from Chung-qing"), a noun-localizer structure (*hun-zhan zhong* "in the scuffle"), etc., and the long pauses adjacent to these PPhs were, on the informant's part, strategies to cause prominent stress on these short phrases, and on the audience's part, offered the listeners longer time to process and catch the information with least distortion.

G. Pitch patterns of prosodic constituents

We then explored the log $F0$ patterns of the three prosodic constituents of PW, PPh, and BG/PGs,. First, we extracted the prosodic-state patterns from the observed pitch contour, $\mathbf{sp}_n$, by eliminating the influence of the current tone, the coarticulations from the two nearest neighboring tones, and the global mean, i.e.,

TABLE IX. Count of short PPh instances with respect to the existence of PM at their two endings.

	PPh length in syllable		
Count of PPh instances	1	2	3
No PMs on both sides	5	38	56
PM on right side only	1	8	28
PM on left side only	6	254	178
PMs on both sides	23	159	121
Total	35	459	383

$$\mathrm{pm}_n = \mathbf{sp}_n(1) - \boldsymbol{\beta}_{t_n}(1) - \boldsymbol{\beta}^f_{B_{n-1},tp_{n-1}}(1) - \boldsymbol{\beta}^b_{B_n,tp_n}(1) - \boldsymbol{\mu}(1) \text{ for } 1 \leqslant n \leqslant N, \tag{13}$$

where $\mathbf{x}(1)$ denotes the first dimension of vector $\mathbf{x}$. A sequence of pm_n delimited by $B2\text{-}1/B2\text{-}2/B3/B4$ at both sides is regarded as a prosodic-state pattern formed by integrating the log $F0$ mean patterns of the three prosodic constituents we considered. A model of prosodic-state pattern is therefore defined by

$$\mathrm{pm}_n = \mathrm{pm}^r_n + \beta_{\mathrm{PW}_n} + \beta_{\mathrm{PPh}_n} + \beta_{\mathrm{BG/PG}_n}, \tag{14}$$

where pm^r_n is the residual of log $F0$ mean at syllable n and β_{PW_n}, β_{PPh_n}, and $\beta_{\mathrm{BG/PG}_n}$ are the log $F0$ patterns of PW, PPh, and BG/PGs, with $\mathrm{PW}_n=(i,j)$, $\mathrm{PPh}_n=(i,j)$, and $\mathrm{BG/PG}_n=(i,j)$ denoting that syllable n is located at the jth place of an i-syllable PW, PPh, and BG/PGs, respectively. The model was trained by a sequential optimization procedure. After well training, the variances of $\mathbf{sp}_n(1)$, pm_n, and pm^r_n were 883.7×10^{-4}, 359.1×10^{-4}, and 191.2×10^{-4}, respectively. Hence, the total residual error (TRE), which is the percentage of sum-squared residue over the observed sum-squared

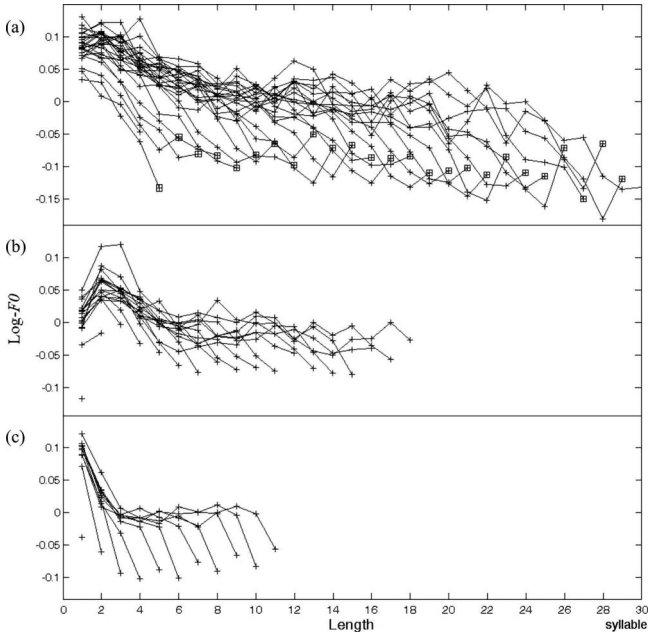

FIG. 11. The log $F0$ patterns of (a) BG/PG, (b) PPh, and (c) PW. The special symbol "□" in (a) indicates the ending syllable of a log $F0$ pattern.

TABLE X. Correlations between unsupervised and human-labeled breaks.

Human unsupervised	*b*1	*b*2	*b*3	*b*4	Total
*B*0	836	207	9	0	1052
*B*1	1970	726	70	0	2766
*B*2-1	81	313	53	1	448
*B*2-2	20	93	227	12	352
*B*3	0	0	137	260	397
*B*4	0	0	4	265	269
B_e	0	0	0	42	42
Total	2907	1339	500	580	5326

log *F*0 mean, is about 21.6% by the current representation.

Figure 11 displays the patterns of β_{PW_n}, β_{PPh_n}, and $\beta_{\mathrm{BG/PG}_n}$ with different lengths. It is noted that only the patterns calculated using more than 20 instances of prosodic-state patterns are displayed because we want to know their general log *F*0 patterns. It can be found from Fig. 11(a) that all $\beta_{\mathrm{BG/PG}}$ had declining patterns with dynamic range spanning approximately from −0.1 to 0.1. Moreover, most of them had short ending resets. From Fig. 11(b), we find that short β_{PPh} had rising-falling patterns, while long β_{PPh} had rising-falling-sustaining-falling patterns. Moreover, they had smaller dynamic range spanning approximately in [−0.07,007]. Lastly, we find from Fig. 11(c) that short β_{PW} showed high-falling patterns, while long β_{PW} showed falling-sustaining-falling patterns. Their dynamic range spanned approximately from −0.1 to 0.1.

From above analyses, we find that the prosodic-state tags possess rich information to represent the high-level prosodic constituents of the four-layer prosodic structure defined in this study. All these three types of log *F*0 patterns generally agree with the findings of previous studies on intonation patterns of Mandarin speech.[55,63,67,68] The superposition patterns $\beta_{\mathrm{PPh}}+\beta_{\mathrm{BG/PG}}$, and all these three patterns (β_{PW}, β_{PPh}, and $\beta_{\mathrm{BG/PG}}$), resembled the intonation patterns reported in the studies Tseng and co-workers[69–72] and the study of Chen *et al.*,[34] respectively. Furthermore, with this prosodically meaningful finding, these quantitative prosodic constituent patterns combining with the APs of tone and coarticulation (i.e., $\boldsymbol{\beta}_t$ and $\boldsymbol{\beta}^f_{B,tp}/\boldsymbol{\beta}^b_{B,tp}$) can be used in Mandarin TTS to generate pitch contour if all break type can be properly predicted from the input text. However, due to the fact that the errors of the current representation are still high, a further study to explore a more efficient representation is worthwhile doing in the future.

H. A comparison with human labeling

To further evaluate the performance of break labeling of the proposed method, a part of the Sinica Treebank corpus used in this study was labeled cooperatively by two experienced labelers working in the Phonetics Laboratory, Department of Foreign Languages and Literatures of National Chiao Tung University. The annotated dataset consisted of 42 utterances with 5326 syllables. The labeling system used was a ToBI-like one developed by the laboratory, which represents the Mandarin speech prosody by a four-layer structure containing syllable, PW, intermediate phrase, and intonation phrase. These four prosodic constituents are delimited by four break types of *b*1, *b*2, *b*3, and *b*4, respectively. Here *b*1 represents an implicit nonbreak index, *b*2 is a perceivable break index for PW boundary, *b*3 is a minor-break index, and *b*4 is a major-break index.

Table X displays the correlation matrix of the break indices labeled by the two methods. It can be found from Table X that 97.8% of human-labeled *b*4s, i.e., major breaks, were labeled as break indices of phrase or utterance boundaries (i.e., *B*3, *B*4, or B_e) in our method, and 96.5% of *b*1s, i.e., nonbreaks, were labeled as indices of SYL boundaries within PW (i.e., *B*0 or *B*1). This indicates that the two labeling methods were consistent for the two extreme cases of nonbreak and major break. It is also observed from the table that *b*3s mainly (73.6%) corresponded to break indices $\geq$*B*2-2, suggesting that the intermediate phrase boundaries in manual labeling, defined and perceived by the labelers as a minor break, were, to quite a certain extent, consistently judged as a clearly perceived short pause (*B*2-2) or medium pause (*B*3) in our labeling. However, in the cases of *b*2, 69.7% of them, defined as perceivable breaks, inconsistently corresponded to nonbreaks (*B*0 or *B*1) in our scheme. To account for such inconsistency, a statistics on the internal morphological and syntactic structures of the PWs delimited by *B*2 and *b*2 shows that (1) while as high as nearly 69.3% of PW-LW correspondence occurred in the human labeling, 40.0% of such correspondence was found in our method, and (2) while 41.2% of the PWs labeled by our method was cases of compound words or long phrases composed of at least four syllables, only 2.2% of the PWs in the similar types was judged by the labelers. This significant discrepancy in the demarcation of PWs between these two methods suggests that labelers, though trained to listen to the prosodic cues with visual aids of graphic user interface to label the breaks, tended to subjectively treat LWs as PWs or as pronunciation units rather than objectively and exclusively relied on the actual prosodic features in prosodic labeling. This inclination obviously resulted in shorter average lengths of prosodic constituents in human labeling. Figure 12 displays the histograms of length of the prosodic constituents formed by the two labeling methods. It can be found from the figure that the average lengths of PWs, PPhs, and BG/PGs labeled by our method were indeed longer than human-labeled PWs, intermediate phrases, and intonational phrases, respectively.

From the perspective of prosodic features, it can be

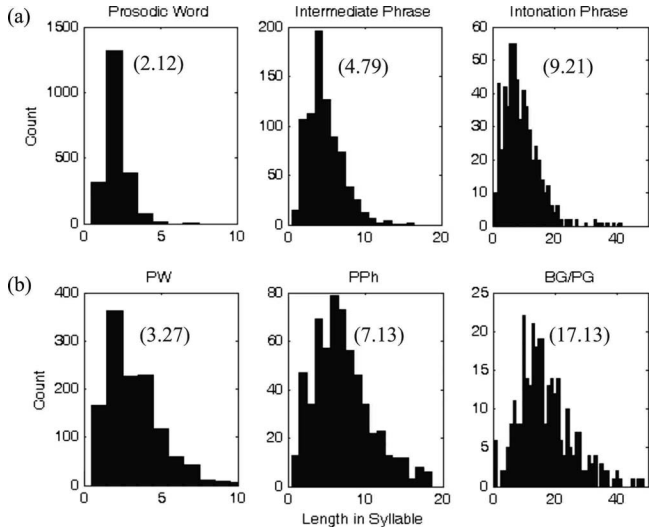

FIG. 12. The histograms of length of the prosodic constituents formed by (a) the human labelers and (b) the proposed methods. The numbers in ( ) represent the average length of prosodic constituents.

found from Figs. 13(a) and 13(b) that the similar histograms of pause duration and normalized pitch jump in the same rows represented labeling consistency in our method, while the distinguishable histograms in the same columns expressed labeling inconsistency in human labeling. Furthermore, Table XI displays symmetric Kullback–Leibler[73] (KL2) distances for the two break labeling methods to measure the difference between two acoustic feature distributions that belong to different break indices labeled by the same method. It can be found from Table XI that the KL2 distances for the proposed unsupervised method were generally greater than those of human labeling. Moreover, we find from Table XI(a) that the KL2 distances of pause duration were relatively large for all break index pairs of the proposed method except $(B1, B2\text{-}1)$; nevertheless the KL2 distances of normalized pitch jump for $(B1, B2\text{-}1)$ were large. On the contrary, we find from Table XI(b) that the KL2 distances of both acoustic features were low for $(b1, b2)$ of human labeling. This confirms that the six break types $B0-B4$ in our labeling have distinct characteristics of acoustic features but the break types in human labeling have less discriminated ones. Specifically, $B4$ has very large pause duration and significant pitch reset, $B3$ has large pause duration and pitch reset, $B2$-2 has medium pause duration, $B2$-1 and $B1$ have small pause duration but $B2$-1 has significant pitch reset and $B0$ has almost no pause duration. This property will be advantageous to our labeling method on those prosody modeling applications using acoustic features.

I. A labeling example

A typical example displaying the labeling results of the beginning part of a long utterance by the two methods is given in Fig. 14. We first examined the labeling results of our method. From Fig. 14(a), we find that the three PMs were labeled as two $B3$ and $B4$. One other $B3$ without PM appeared at the right boundary of a nine-syllable NP. Besides, there existed five $B2$-1 and four $B2$-2. They all appeared at interword junctures. We also find from Fig. 14(b) that all

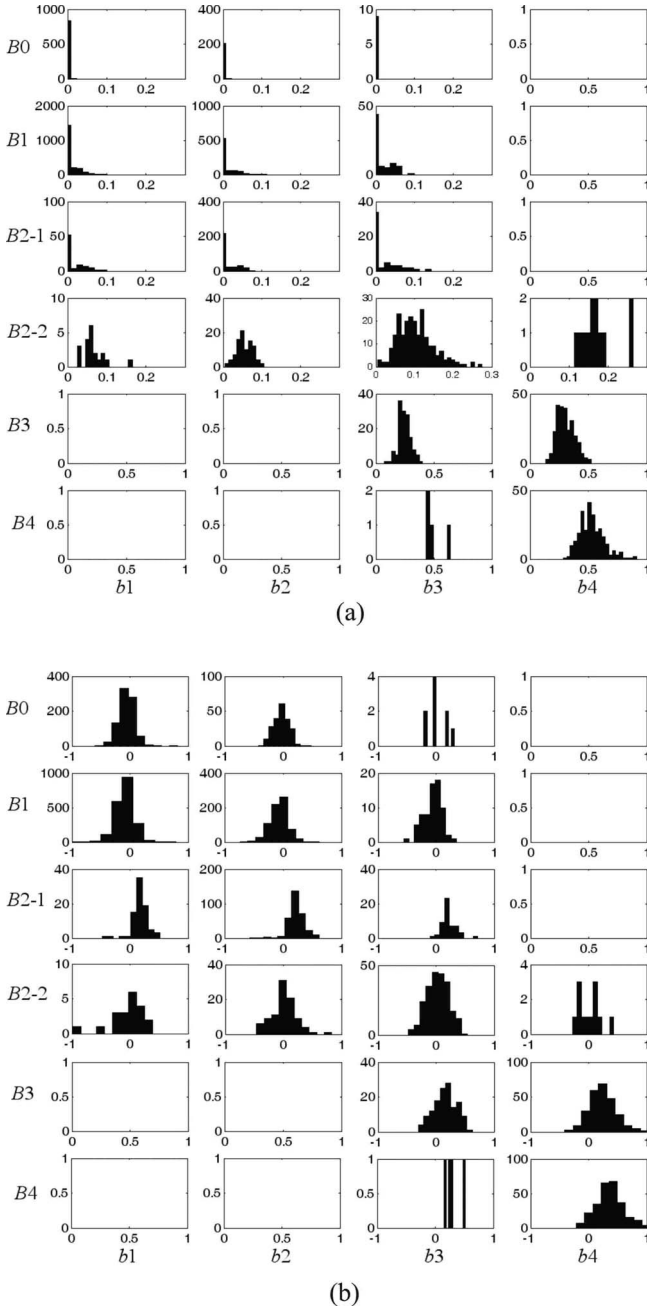

FIG. 13. The histograms of (a) pause duration (in seconds) and (b) normalized pitch jump (in log $F0$) for syllable-juncture instances belonging to subgroups with different break-index pairs labeled by the two methods.

three $B3$ and five $B2$-1 had clear normalized log $F0$ reset. Moreover, the curve of integrating APs of prosodic state and the global mean of pitch level showed smoother PW patterns derived via removing the tone and coarticulation effects from the observed zigzag curve of log $F0$ mean. We then compared the results of the two labeling methods. It can be found from Fig. 14(a) that aside from giving indices of breaks to all the above-mentioned breaks labeled by our method, human labelers gave four additional breaks to divide the nine-syllable-NP (*xing-zheng-yuan zhu-ji-chu de tong-ji*) PW into three PWs, and the two four-syllable compound-word PWs, "*jin-kou* (import) *jin-e* (the amount of money)" and "*qu-nian* (last year) *tong-qi* (the same period)," into four two-syllable words. To justify whether the deletions of these four human-

TABLE XI. KL2 distances measuring the difference between two acoustic feature distributions that belong to different break indices labeled by the same method: (a) the proposed method and (b) human labeling. Upper and lower triangular matrices represent KL2 distances for pause duration and normalized pitch jump, respectively.

		*B*0	*B*1	*B*2-1	*B*2-2	*B*3	*B*4
(a)	*B*0		2.63	3.39	23.59	23.42	22.77
	*B*1	0.19		0.16	14.21	23.28	22.66
	*B*2-1	4.59	4.87		11.92	21.17	20.62
	*B*2-2	0.52	0.72	2.79		13.84	18.85
	*B*3	1.66	2.12	1.25	1.43		12.71
	*B*4	3.69	4.18	0.36	2.50	0.88	
			*b*1	*b*2	*b*3	*b*4	
(b)		*b*1		0.12	6.83	23.16	
		*b*2	0.24		6.07	22.10	
		*b*3	0.60	0.36		10.56	
		*b*4	2.05	1.20	0.82		

labeled breaks were reasonable, we examined the pause durations of these four word junctures and the normalized pitch patterns of the three integrated PWs. The pause durations were 12, 40, 22, and 1 ms. Obviously, they were all not significant. Besides, as seen in Fig. 14(b) all the three normalized pitch patterns of nine-syllable-NP PW and two four-syllable compound-word PWs were smooth. So the deletions of these four breaks by our method seemed reasonable.

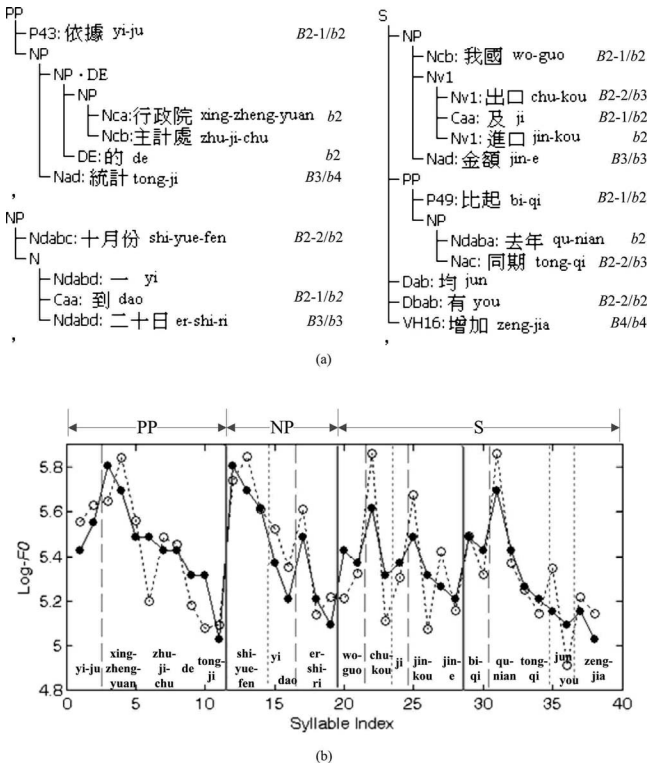

FIG. 14. An example of the automatic prosody labeling. (a) Syntactic trees with prosodic tags: uppercase *B* and lowercase *b* for break index labeled by our method and the human labeler, respectively, and (b) syllable log *F*0 means: observed (open circle) and prosodic state+global mean (close circle). Solid/dashed/dotted lines represent *B*3/*B*2-1/*B*2-2, respectively. The utterance is "*yi-ju* (according to) *xing-zheng-yuan* (the Executive Yuan) *zhu-ji-chu* (Directorate-General of Budget, Accounting and Statistics) *de* (DE) *tong-ji* (statistics), *shi-yue-fen* (October) *yi* (1st) *dao* (to) *er-shi-ri* (20th), *wo-guo* (our country) *chu-kou* (export) *ji* (and) *jin-kou* (import) *jin-e* (the amount of money) *bi-qi* (in comparison with) *qu-nian* (last year) *tong-qi* (the same period) *jun* (both) *you* (to have some) *zeng-jia* (increase).

V. CONCLUSIONS

In this paper, a new approach of joint prosody labeling and modeling for Mandarin speech has been proposed. It first employed four prosodic models to describe the relationship of two types of prosodic tags to be labeled with the input acoustic prosodic features and linguistic features, and then used a sequential optimization procedure to determine all prosodic tags and estimate the parameters of the four prosodic models jointly using the Sinica Treebank speech corpus. Experimental results showed that the estimated parameters of the four prosodic models were able to penetratingly explore and appropriately describe the hierarchy of Mandarin prosody. First, the syllable pitch contour model was able to interpret the variation in syllable pitch contour controlled by such affecting factors as lexical tones, adjacent breaks, and prosodic state. Next, the prosodic-state model was developed to clearly describe the declination effect of log *F*0 level within PW and the resets across PW, PPh, and BG/PG, and hence to extract the pitch patterns of each prosodic constituent. Then, the break-acoustics model could demonstrate the distinct acoustic characteristics for each of the six break types. The last model, the break-syntax model, was built to express the general relationship between the break type and the linguistic features of various levels. Besides, the performance of our models was further confirmed by the corresponding relationships found between the break indices labeled and their associated words which served as evidences to manifest the connections between prosodic and linguistic parameters, and it was also verified by our more consistent and discriminative prosodic feature distributions than those in human labeling by a quantitative comparison. In conclusion, the method we proposed to develop the joint prosody labeling and modeling for Mandarin speech was able to construct interpretive prosodic models and generate prosodic tags that were automatically and consistently labeled.

Some future works are worth doing. First, the syllable pitch contour model can be extended to jointly model syl-

lable pitch contour, syllable duration, and syllable energy level simultaneously. Second, the database with prosodic tags being properly labeled can be used to exploit the hierarchical structure of Mandarin prosody in more detail. Third, the break-syntax model can be extended to consider more linguistic features and applied to the problem of break-type prediction from linguistic features. Fourth, the break-acoustics model can be extended to include more acoustic and linguistic features and applied to the problems of speech segmentation and punctuation generation. Lastly, the four prosodic models can be used to provide useful prosodic information to assist in ASR.

ACKNOWLEDGMENTS

This work was supported by the NSC of Taiwan under Contract Nos. NSC95-2218-E-002-027 and NSC95-2752-E009-014-PAE. The authors would like to thank Academia Sinica, Taiwan for providing the Tree-Bank text corpus, and Dr. Ho-Hsien Pan of Phonetics Laboratory, Department of Foreign Languages and Literatures of National Chiao Tung University, Taiwan for her generous and helpful assistance in manually labeling our experimental database.

APPENDIX: THE ALGORITHM TO DETERMINE ALL THRESHOLDS OF THE DECISION TREE FOR INITIAL BREAK LABELING

1. Determinations of Th1, Th2, and Th3

Th1, Th2, and Th3 are pause-duration thresholds set to sequentially distinguish $B4$, $B3$, and $B2\text{-}2/B1$ with significant pause duration from other break types. First, the two gamma distributions for $B3$ and $B4$ are estimated using two clusters of pause duration samples of syllable juncture with PM clustered by VQ. The one with larger mean is regarded as the distribution for $B4$, and another is for $B3$. We then construct an empirical gamma distribution of pause duration $f_{B0/B1}(\mathrm{pd})$ for $B0/B1$ by using all samples of intraword juncture. An empirical distribution of pause duration $f_{B2\text{-}2}(\mathrm{pd})$ for $B2$-2 is then constructed by using all samples of interword juncture without PM but with apparent pause. Here, the condition of apparent pause is evaluated based on the criterion of $f_{B3}(\mathrm{pd}_n) > f_{B0/B1}(\mathrm{pd}_n)$ which can exclude non-PM interword samples with pause duration similar to those of $B0/B1$. Lastly, the thresholds Th3, Th2, and Th1 are set as the equal-probability intersections of $f_{B0/B1}(\mathrm{pd})$, $f_{B2\text{-}2}(\mathrm{pd})$, $f_{B3}(\mathrm{pd})$, and $f_{B4}(\mathrm{pd})$.

2. Determination of Th5

The pitch jump threshold Th5 is set to distinguish between $B2$-1 and $B0/B1$. We first define the normalized log $F0$ level jump by

$$\xi_n = (\mathbf{sp}_{n+1}(1) - \boldsymbol{\beta}_{t_{n+1}}(1)) - (\mathbf{sp}_n(1) - \boldsymbol{\beta}_{t_n}(1)), \qquad \text{(A1)}$$

where $\mathbf{x}(1)$ denotes the first dimension of vector $\mathbf{x}$. It is noted that the APs of five tones, $\boldsymbol{\beta}_t$, can be estimated in advance before break-type labeling by simply averaging all samples of each tone. Then two empirical Gaussian distributions of normalized log $F0$ level jump, $f_{\mathrm{intra}}(\xi)$ and $f_{\mathrm{PM}}(\xi)$, for intraword and PM junctures are constructed using all samples of intraword syllable junctures and all PM junctures, respectively. We then construct an empirical Gaussian distribution of normalized log $F0$ level jump $f_{B2\text{-}1}(\xi)$ for $B2$-1 by using all samples of interword junctures without PM but with apparent normalized log $F0$ level jump. The condition of apparent normalized log $F0$ level jump is evaluated based on the criterion of $f_{\mathrm{PM}}(\xi_n) > f_{\mathrm{intra}}(\xi_n)$ which can exclude non-PM interword junctures with normalized log $F0$ level jump similar to intraword juncture. Lastly, the threshold Th5 is set as the equal-probability intersection of $f_{\mathrm{intra}}(\xi)$ and $f_{B2\text{-}1}(\xi)$.

3. Determinations of Th4 and Th6

The $F0$ pause duration threshold Th4 and the energy-dip level threshold Th6 are set to distinguish between $B0$ and $B1$. Basically, $B1$ should have very short $F0$ pause duration and large energy-dip level because it represents tightly coupling syllable juncture. So, we simply set Th4 to be 1 frame (=10 ms). For Th6, the two Gaussian distributions for $B0$ and $B1$ are estimated using two clusters of energy-dip level samples of intraword juncture clustered by VQ. Then, the threshold Th6 is set as the equal-probability intersection of the two Gaussian distributions.

[1] E. Selkirk, "On prosodic structure and its relation to syntactic structure," *Nordic Prosody* (Tapir, Trondheim, Norway), Vol. **2**, pp. 111–140.
[2] E. Selkirk, *Phonology and Syntax: The Relation Between Sound and Structure* (MIT Press, Cambridge, MA, 1984).
[3] M. Beckman and J. Pierrehumbert, "Intonational structure in Japanese and English," *Phonology Yearbook 3* (Cambridge University Press, UK, 1986), pp. 255–309.
[4] A.-J. Li, Y.-Q. Zu, and Z.-Q. Li, "A national database design and prosodic labeling for speech synthesis," Proceedings of the Oriental COCOSDA Workshop 1999, pp. 13–16.
[5] A.-J. Li and M.-C. Lin, "Speech corpus of Chinese discourse and the phonetic research," Proceedings of the ICSLP 2000, Vol. **4**, pp. 13–18.
[6] J.-F. Cao, "Rhythm of spoken Chinese—Linguistic and paralinguistic evidences," Proceedings of the ICSLP 2000, Vol. **2**, pp. 357–360.
[7] C.-Y. Tseng, S.-H. Pin, Y.-L. Lee, H.-M. Wang, and Y.-C. Chen, "Fluent speech prosody: Framework and modeling," Speech Commun. special issue on quantitative prosody modeling for natural speech description and generation, **46**, 284–309 (2005).
[8] S.-H. Pin, Y.-L. Lee, Y.-C. Chen, H.-M. Wang, and C.-Y. Tseng, "A Mandarin TTS system with an integrated prosodic model," Proceedings of the ISCSLP 2004, pp. 169–172.
[9] 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," Proceedings of the ICME 2000, Vol. **1**, pp. 448–4511.
[10] 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**, 226–239 (1998).
[11] Y. Liu, E. Shriberg, S. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," IEEE Trans. Audio, Speech, Lang. Process. **14**, 526–1540 (2006).
[12] Y. Gotoh and S. Renals, "Sentence boundary detection in broadcast speech transcripts," Proceedings of the ISCA Workshop: Automatic Speech Recognition: Challenges for the New Millennium ASR 2000, pp. 228–235.
[13] E. Shriberg, A. Stolcke, D. Hakkani-Tur, and G. Tur, "Prosody-based automatic segmentation of speech into sentences and topics," Speech Commun. **32**, 127–154 (2000).
[14] J.-H. Kim and P. C. Woodland, "A combined punctuation generation and speech recognition system and its performance enhancement using prosody," Speech Commun. **41**, 563–577 (2003).
[15] J.-H. Kim and P. C. Woodland, "The use of prosody in a combined system for punctuation generation and speech recognition," Proceedings of the

Eurospeech 2001, pp. 2757–2760.
[16]H. Christensen, Y. Gotoh, and S. Renals, "Punctuation annotation using statistical prosody models," Proceedings of the ISCA Workshop on Prosody in Speech Recognition and Understanding 2001, pp. 35–40.
[17]J.-F. Yeh and C.-H. Wu, "Edit disfluency detection and correction using a cleanup language model and an alignment model," IEEE Trans. Audio, Speech, Lang. Process. **14**, 1574–1583 (2006).
[18]M. Lease, M. Johnson, and E. Charniak, "Recognizing disfluencies in conversational speech," IEEE Trans. Audio, Speech, Lang. Process. **14**, 1566–1573 (2006).
[19]C.-K. Lin and L.-S. Lee, "Improved spontaneous Mandarin speech recognition by disfluency interruption point (IP) detection using prosodic features," Proceedings of the Eurospeech 2005, pp. 1621–1624.
[20]K. Chen and M. Hasegawa-Johnson, "How prosody improves word recognition," Proceedings of the ISCA International Conference on Speech Prosody 2004, pp. 583–586.
[21]K. Chen, M. Hasegawa-Johnson, A. Cohen, S. Borys, S.-S. Kim, J. Cole, and J.-Y. Choi, "Prosody dependent speech recognition on radio news corpus of American English," IEEE Trans. Audio, Speech, Lang. Process. **14**, 232–245 (2006).
[22]K. Chen and M. Hasegawa-Johnson, "Improving the robustness of prosody dependent language modeling based on prosody syntax dependence," Proceedings of the IEEE ASRU 2003, pp. 435–440.
[23]E. Shriberg and A. Stolcke, "Direct modeling of prosody: An overview of applications in automatic speech processing," Proceedings of the ISCA International Conference on Speech Prosody 2004, pp. 575–582.
[24]J.-H. Yang, Y.-F. Liao, Y.-R. Wang, and S.-H. Chan, "A new approach of using temporal information in Mandarin speech recognition," Proceedings of the ISCA International Conference on Speech Prosody 2006, Vol. **SPS4-3**.
[25]X. Lei and M. Ostendorf, "Word-level tone modeling for Mandarin speech recognition," Proceedings of the IEEE ICASSP 2007, Vol. **4**, pp. 665–668.
[26]C.-Y. Tseng, "Recognizing Mandarin Chinese fluent speech using prosody information—An initial investigation," Proceedings of the ISCA International Conference on Speech Prosody 2006.
[27]K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling English prosody," Proceedings of the ICSLP 1992, Vol. **2**, pp. 867–870.
[28]A. Batliner, J. Buckow, H. Niemann, E. Noth, and V. Warnke, "The prosody module," in *Verbmobil: Foundations of Speech-to-Speech Translation*, edited by W. Wahlster (Springer, New York, 2000).
[29]M. Selting, *Prosody in Conversation* (Max Niemeyer, Tuebingen, Germany, 1995), in German.
[30]D. J. Hirst, "The symbolic coding of fundamental frequency curves: From acoustics to phonology," Proceedings of the International Symposium on Prosody 1994.
[31]P. A. Taylor, "The tilt intonation model," Proceedings of the ICSLP 1998, Vol. **4**, pp. 1383–1386.
[32]S.-H. Peng, M. K. M. Chan, C.-Y. Tseng, T. Huang, O.-J. Lee, and M. Beckman, "Towards a Pan-Mandarin system for prosodic transcription," in *Prosodic Typoloxgy: The Phonology of Intonation and Phrasing*, edited by S.-A. Jun (Oxford University Press, Oxford, 2005), pp. 230–270.
[33]A.-J. Li, "Chinese prosody and prosodic labeling of spontaneous speech," Proceedings of the ISCA International Conference on Speech Prosody 2002, pp. 39–46.
[34]G.-P. Chen, G. Bailly, Q.-F. Liu, and R.-H. Wang, "A superposed prosodic model for Chinese text-to-speech synthesis," Proceedings of the ISCSLP 2004, pp. 177–180.
[35]M.-S. Yu, N.-H. Pan, and M.-J. Wu, "A statistical model with hierarchical structure for predicting prosody in a Mandarin text-to-speech system," Proceedings of the ISCSLP 2002, pp. 21–24.
[36]G. Bailly and B. Holm, "SFC: A trainable prosodic model," Speech Commun. **46**, 348–364 (2005).
[37]M. Ostendorf and N. Veilleux, "A hierarchical stochastic model for automatic prediction of prosodic boundary location," Comput. Linguist. **20**, 27–52 (1994).
[38]X. Shen and B. Xu, "A CART-based hierarchical stochastic model for prosodic phrasing in Chinese," Proceedings of the ISCSLP 2000, pp. 105–109.
[39]H.-J. Peng, C.-C. Chen, C.-Y. Tseng, and K.-J. Chen, "Predicting prosodic words from lexical words—A first step towards predicting prosody from text," Proceedings of the ISCSLP 2004, pp. 173–176.
[40]J. Hirschberg and P. Prieto, "Training intonational phrasing rules automatically for English and Spanish text-to-speech," Speech Commun. **18**, 281–290 (1996).
[41]D.-W. Xu, H.-F. Wang, G.-H. Li, and T. Kagoshima, "Parsing hierarchical prosodic structure for Mandarin speech synthesis," Proceedings of the IEEE ICASSP 2006, Vol. **1**, pp. 14–19.
[42]X. Sun and T. H. Applebaum, "Intonational phrase break prediction using decision tree and n-gram model," Proceedings of the Eurospeech 2001, pp. 537–540.
[43]A. W. Black and P. Taylor, "Assigning phrase breaks from part-of-speech sequences," Proceedings of the Eurospeech 1997, pp. 995–998.
[44]Z. Sheng, J.-H. Tao, and D.-L. Jiang, "Chinese prosodic phrasing with extended features," Proceedings of the IEEE ICASSP 2003, Vol. **1**, pp. 492–495.
[45]J.-F. Li, G.-P. Hu, and R.-H. Wang, "Chinese prosody phrase break prediction based on maximum entropy model," Proceedings of the Interspeech 2004, pp. 729–732.
[46]Y.-Q. Shao, Y.-Z. Zhao, J.-Q. Han, and T. Liu, "Using different models to label the break indices for mandarin speech synthesis," Proceedings of the ICMLC 2005, Vol. **6**, pp. 3802–3807.
[47]J.-F. Li, G.-P. Hu, R.-H. Wang, and L.-R. Dai, "Sliding window smoothing for maximum entropy based intonational phrase prediction in Chinese," Proceedings of the IEEE ICASSP 2005, Vol. **1**, pp. 285–288.
[48]Z.-P. Zhao, T.-J. Zhao, and Y.-T. Zhu, "A maximum entropy Markov model for prediction of prosodic phrase boundaries in Chinese TTS," Proceedings of the IEEE GrC 2007, pp. 498–498.
[49]C. W. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," IEEE Trans. Speech Audio Process. **2**, 469–481 (1994).
[50]K. Chen, M. Hasegawa-Johnson, and A. Cohen, "An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model," Proceedings of the IEEE ICASSP 2004, Vol. **1**, pp. 509–512.
[51]V. Rangarajan, S. Narayanan, and S. Bangalore, "Acoustic-syntactic maximum entropy model for automatic prosody labeling," Proceedings of the IEEE Spoken Language Technology Workshop 2006, pp. 74–77.
[52]X.-J. Ma, W. Zhang, Q. Shi, W.-B. Zhu, and L.-Q. Shen, "Automatic prosody labeling using both text and acoustic information," Proceedings of the IEEE ICASSP 2003, Vol. **1**, pp. 516–519.
[53]A. F. Muller, H. G. Zimmermann, and R. Neuneier, "Robust generation of symbolic prosody by a neural classifier based on autoassociators," Proceedings of the IEEE ICASSP 2000, Vol. **3**, pp. 1285–1288.
[54]J.-H. Tao, "Acoustic and linguistic information based Chinese prosodic boundary labeling," Proceedings of the TAL 2004, pp. 181–184.
[55]S.-H. Chen, W.-H. Lai, and Y.-R. Wang, "A statistics-based pitch contour model for Mandarin speech," J. Acoust. Soc. Am. **117**, pp. 908–925 (2005).
[56]S.-H. Chen, W.-H. Lai, and Y.-R. Wang, "A new duration modeling approach for Mandarin speech," IEEE Trans. Speech Audio Process. **11**, 308–320 (2003).
[57]S.-H. Chen and Y.-R. Wang, "Vector quantization of pitch information in Mandarin speech," IEEE Trans. Commun. **38**, 1317–1320 (1990).
[58]L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees* (Wadsworth, Belmont, CA, 1984).
[59]Y. Qian and W.-Y. Pan, "Prosodic word: The lowest constituent in the Mandarin prosody processing," Proceedings of the ISCA International Conference on Speech Prosody 2002, pp. 591–594.
[60]J.-H. Tao, H.-G. Dong, and S. Zhao, "Rule learning based Chinese prosodic phrase prediction," Proceedings of the IEEE NLP-KE 2003, pp. 425–432.
[61]C.-R. Huang, K.-J. Chen, F.-Y. Chen, Z.-M. Gao, and K.-Y. Chen, "Sinica Treebank: Design criteria, annotation guidelines, and pn-line interface," Proceedings of the Second Chinese Language Processing Workshop 2000, pp. 29–37.
[62]Y.-R. Chao, *A Grammar of Spoken Chinese* (Berkeley Press, Berkeley, CA, 1968).
[63]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**, 1309–1320 (1989).
[64]Y. Xu, "Contextual tonal variations in Mandarin," J. Phonetics **25**, 61–83 (1997).
[65]K.-J. Chen and C.-R. Huang, "Part of speech (POS) analysis on Chinese language," CKIP Technical Report No. 93-05, Institute of Information Science, Academia Sinica, Taiwan, R.O.C., 1993 (in Chinese).
[66]Chinese knowledge Information Processing (CKIP), Academia Sinica, "An introduction to Academia Sinica balanced corpus for modern Mandarin Chinese," CKIP Technical Report No. 95-02, Institute of Information

Science, Academia Sinica, Taiwan, R.O.C. 1995 (in Chinese).
[67]C. Shih, "Declination in Mandarin," Proceedings of the ESCA Workshop on Intonation: Theory, Models and Applications 1997, pp. 293–296.
[68]Y. Yufang and W. Bei, "Acoustic correlates of hierarchical prosodic boundary in Mandarin," Proceedings of the ISCA International Conference on Speech Prosody 2002, pp. 707–710.
[69]C.-Y. Tseng and S.-H. Pin, "Mandarin Chinese prosodic phrase grouping and modeling: Method and implications," Proceedings of the TAL 2004, pp. 193–196.
[70]C.-Y. Tseng and S.-H. Pin, "Modeling prosody of Mandarin Chinese fluent speech via phrase grouping," Proceedings of the Speech and Language Systems for Human Communication (SPLASH-2004/Oriental-COCOSDA2004), 2004, pp. 53–57.
[71]C.-Y. Tseng and Z.-Y. Su, "Corpus approach to phonetic investigation—Methods, quantitative evidence and findings of Mandarin speech prosody," Proceedings of the Oriental COCOSDA Workshop 2006, pp. 123–138.
[72]C.-Y. Tseng, "Higher level organization and discourse prosody," Proceedings of the TAL 2006, pp. 23–34.
[73]S. Theodoridis and K. Koutroumbas, *Pattern Recognition* 2nd ed. (Elsevier, London, UK, 2003).