A New Prosody-Assisted Mandarin ASR System

Sin-Horng Chen, *Senior Member, IEEE*, Jyh-Her Yang, Chen-Yu Chiang, *Member, IEEE*, Ming-Chieh Liu, and Yih-Ru Wang, *Member, IEEE*

***Abstract*—This paper presents a new prosody-assisted automatic speech recognition (ASR) system for Mandarin speech. It differs from the conventional approach of using simple prosodic cues on employing a sophisticated prosody modeling approach based on a four-layer prosody-hierarchy structure to automatically generate 12 prosodic models from a large unlabeled speech database by the joint prosody labeling and modeling (PLM) algorithm proposed previously. By incorporating these 12 prosodic models into a two-stage ASR system to rescore the word lattice generated in the first stage by the conventional hidden Markov model (HMM) recognizer, we can obtain a better recognized word string. Besides, some other information can also be decoded, including part of speech (POS), punctuation mark (PM), and two types of prosodic tags which can be used to construct the prosody-hierarchy structure of the testing speech. Experimental results on the TCC300 database, which consists of long paragraphic utterances, showed that the proposed system significantly outperformed the baseline scheme using an HMM recognizer with a factored language model which models word, POS, and PM. Performances of 20.7%, 14.4%, and 9.6% in word, character, and base-syllable error rates were obtained. They corresponded to 3.7%, 3.7%, and 2.4% absolute (or 15.2%, 20.4%, and 20% relative) error reductions. By an error analysis, we found that many word segmentation errors and tone recognition errors were corrected.**

***Index Terms*—Prosody modeling, prosody-assisted automatic speech recognition (ASR), prosody-hierarchy structure.**

I. Introduction

THE use of prosodic information in automatic speech recognition (ASR) is an attractive research topic in recent years. Prosody refers to the suprasegmental features of continuous speech, such as accentuation, prominence, tone, pause, intonation, and rhythm. Prosody is physically encoded in the variations of pitch contour, energy level, duration, and silence of spoken utterances. Prosody is known to closely correlate with the linguistic features of various levels, say from phone, syllable, word, phrase, to sentence or above. Owing to those correlations, prosody is potentially useful for ASR. Generally, the task of prosody-assisted ASR is to first exploit prosodic cues correlated to linguistic features, and to then model their relationships with linguistic features and prosodic-acoustic features, and to lastly incorporate these models into the ASR framework.

In the past, many studies on using prosodic information to assist in ASR have been reported [1]–[7] for American English [1]–[4], [6], [7] and Spanish [5]. Ananthakrishnan *et al.* [1]–[3] proposed to incorporate a prosodic language model and a prosodic acoustic model into the conventional hidden Markov model (HMM)-based ASR recognizer by rescoring the N-best word sequences or the word lattice. The prosodic acoustic model used Gaussian mixture model (GMM) or multilayer perceptrons (MLP) to model the relation of binary pitch accent label of word and the prosodic-acoustic features extracted from the F0 track, energy, and duration cues of context. The prosodic language model was a trigram language model (LM) with compound tokens of words and their binary pitch accent labels. Besides, an unsupervised adaptation approach to jointly refining the two categorical prosody models and bootstrapping prosodic labels was also proposed to assist in solving the problem of lacking large corpora annotated with relevant prosodic symbols [1]. Relative improvements of 1.2%–3.1% in word error rate (WER) were obtained on the Boston University Radio News Corpus (BU-RNC). Chen *et al.* [4] used two prosodic events, intonational phrase boundary and pitch accent, in ASR to construct prosody-dependent word and phoneme models. A relative improvement of 6.9% in WER was achieved on BU-RNC. Milone *et al.* [5] proposed a method to use the accentual information in ASR. The method first estimated a sequence of accentual structure of words from speech signal using F0 and energy by an HMM-based classifier or a neural tree networks classifier, and then incorporated it into the recognition process. An LM built to take into account the accentual structure of words in phrase was used. A relative improvement of 28.91% in WER was achieved on a medium-vocabulary Spanish continuous-speech recognition task. Vergyri *et al.* [6] proposed to integrate models of different prosodic knowledge sources into ASR. They included word duration model, pause language model, and prosodic model of hidden events (e.g., sentence boundaries and speech disfluencies). Relative improvements of 2.6–3.1% in WER were achieved on the Switchboard database. Ostendorf *et al.* [7] presented a statistical modeling framework for incorporating prosody in the speech recognition process. Several issues were discussed, including prosodic feature extraction in different time scales and normalization, prosody modeling using an intermediate symbol representation in contrast to directly conditioning on acoustic correlates, the use of questions about prosodic structure in acoustic model clustering, dynamic pronunciation modeling conditioned on acoustic-prosodic features, etc.

Besides, some other studies on using prosodic information to assist in Mandarin ASR can also be found [8]–[13]. In [8], a recurrent neural network (RNN) was used to detect word-boundary information from the input prosodic features with base-syllable boundary being predetermined by an

Manuscript received January 26, 2011; revised May 09, 2011; accepted January 02, 2012. Date of publication February 07, 2012; date of current version March 30, 2012. This work was supported by the NSC of Taiwan under Contract NSC98-2221-E-009-075-MY3 and NSC99-2221-E-009-009-MY3. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dimitra Vergyri.

S.-H. Chen, J.-H. Yang, C.-Y. Chiang, and Y.-R. Wang are with the Department of Electrical Engineering, National Chiao Tung University, Hsinchu 300, Taiwan (e-mail: neil.yang0204@gmail.com).

M.-C. Liu was with the Department of Electrical Engineering, National Chiao Tung University, Hsinchu 300, Taiwan He is now with the Taiwan Semiconductor Manufacturing Company, Hsinchu 300-78, Taiwan.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2012.2187192

HMM-based acoustic decoder. The word boundary information was then used to assist the linguistic decoder in solving word-boundary ambiguity as well as pruning unlikely paths. An absolute improvement of 1.1% in character error rate (CER) was achieved on a large-vocabulary speaker-dependent (SD) Mandarin continuous ASR task. Huang *et al.* [9], [10] utilized decision tree-based or GMM-based prosodic models of syllable- and word-level to generate the prosodic likelihood score for rescoring in a two-pass recognition process. Absolute CER improvements of 1.06% [9] and 1.45% [10] were reported on a large-vocabulary multi-speaker continuous ASR task. In [11], word-dependent tone modeling using prosodic features of syllable duration and three F0 values with two back-off schemes was proposed for Mandarin ASR. A minor improvement on CER was achieved on a Mandarin broadcast news corpus. Ni *et al.* [12] proposed an implicit tone model using F0 contour features and an explicit tone model using both prosodic and lexical features for assisting in Mandarin ASR. An improvement of 3.65% in CER was achieved on the Project-863 database. In [13], Ni *et al.* incorporated a GMM-based prosody-dependent tonal syllable duration model and a maximum entropy (ME)-based syntactical prosody model into a prosody-dependent acoustic model recognizer by rescoring the syllable lattice. Only tonal syllable recognition rate was reported on the Project-863 database.

Prosody modeling was also used in some other speech recognition tasks. Liu *et al.* [14] conducted enriching speech recognition to automatic detection of sentence boundaries and disfluencies on both conversational telephone speech and broadcast news tasks of NIST RT-04F evaluation using both prosodic and lexical features. Shriberg *et al.* [15] employed the decision tree method to model rhythmic and melodic features of speech for several applications including sentence segmentation and disfluency detection, topic segmentation in broadcast news, dialog act labeling and word recognition in conversational speech. Although prosody modeling was useful in those applications, only minor improvements on word recognition were achieved.

It can be found from above discussions that prosody modeling is the main concern in all those previous studies. The methods of prosody modeling in those studies can be classified into two classes: 1) direct modeling of target classes [8], [10]–[12], and 2) prosody modeling via intermediate abstract phonological categories [1]–[6], [9], [13], such as ToBI [16] and INTSINT [17]. In direct modeling of target classes, the relationship between prosodic acoustic features and target classes (usually, linguistic feature, e.g., lexical tone, lexical word, etc.) is directly modeled by a pattern classifier, such as GMM, decision tree, RNN, ME, etc. This approach is advantageous on bypassing manual labeling of prosodic tags and hence can avoid the inter-annotator inconsistency. Nevertheless, the variability or space of both prosodic-acoustic and linguistic features (target) may be too large when considering more features of various levels or wider time window. Therefore, only limited linguistic and prosodic-acoustic features are incorporated in this direct modeling approach [8], [10]–[12]. On the other hand, prosody modeling via intermediate abstract phonological categories [1]–[6], [9], [13] first explores important prosodic cues or events potentially useful for ASR and then builds prosodic models to describe the relations of these prosodic cues with linguistic features of various levels and prosodic-acoustic features using a prosody-annotated speech database. Fig. 1 shows a conceptual block diagram of the prosody modeling using intermediate abstract phonological categories. Usually, prosody annotation is based on the ToBI labeling system [16] and is performed manually. The variability of prosodic-acoustic features can be reduced by introducing a finite discrete set of prosody tags so as to make the construction of prosody-syntax relationship easier. The main drawback of this approach is the lack of a large well-labeled database due to time-consuming labeling work and inconsistent labeling between human annotators so that only few obvious prosodic cues, such as pitch accent and intonational phrase boundary, are used. This leads to the limitation of the effect of using prosodic information on improving the ASR performance. Although some studies [13], [18], [19] conducted automatic prosody labeling to enlarge the size of prosody-annotated corpus, their prosodic models were still trained with manually annotated speech corpora so that their performances were subject to the quality of human prosody labeling. Table I summarizes the primary features of prosody modeling and experiment setting for those previous studies on prosody-assisted ASR for comparison.

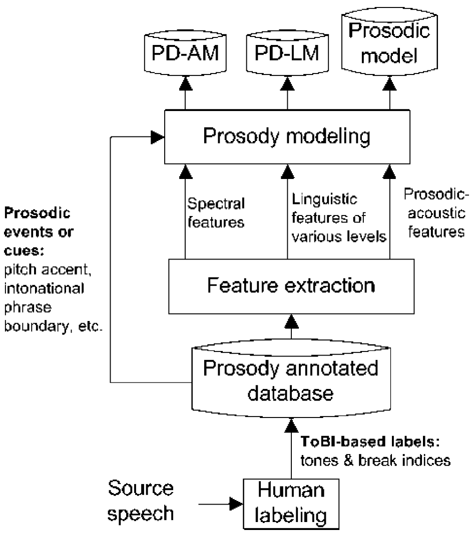


Fig. 1. Conceptual block diagram of the prosody modeling class using intermediate abstract phonological categories. PD-AM and PD-LM denote prosody-dependent acoustic model and prosody-dependent language model.

In this paper, a new prosody-assisted ASR system is proposed for Mandarin speech. It differs from the conventional prosody-assisted ASR system with prosody modeling shown in Fig. 1 mainly on adopting a systematic way to perform prosody modeling. Fig. 2 shows a conceptual block diagram of the proposed approach of prosody modeling. It is an extension of our previous study on the joint prosody labeling and modeling using an unlabeled speech database [20]. A four-layer model of prosody hierarchy of Mandarin speech defined based on two types of prosodic tags, break type and prosodic state, is first chosen. Several prosodic models are then designed to describe various relationships of these two types of tags with both the linguistic features of texts and the prosodic-acoustic features of speech signals. Lastly, the joint prosody labeling and modeling (PLM) algorithm proposed previously [20] is used to train those prosodic models from a large unlabeled speech database. The new approach is advantageous on involving abundant prosodic cues in the prosody modeling for assisting in ASR. We can therefore expect that it performs better on improving the word recognition performance. Besides, more information other than the

TABLE I
COMPARISON BETWEEN PROSODY-ASSISTED ASR STUDIES

	Prosody modeling					Experiment setting				
Literature	**PE**	**PH**	**PL**	**PAF**	**LF**	**LNG**	**STL**	**VSZ**	**SPK**	**IMP (%)**
Ni [13]	2B+2S	1-L	SS	F0/d	t	M	R	TSR	SI	9.82/24.4(tonal syllable)
Huang [9]	2B	2-L	R	F0*/d*/e*/pd	t/WB	M	B	100K	SD	1.06/5.5(character)
Ana.[1]	2A	-	UA	F0*/d*/e*	W	E	R	-	-	1/3.1
Chen [4]	2B+2A	1-L*	S	F0/d	ph/W/POS	E	R	-	SI	1.73/6.9
Vergyri [6]	3P+5HE	-	-	F0*/d*/pd	ph/W	E	C	8K	SI	1.1,0.7,0.9/3.9,2.6, 3.1
Milone [5]	AS	-	-	F0/e/d	W	S	R	<500	SI	2.18/28.91
Huang [10]	Dir	-	-	F0*/d*/e*/pd	t/WB	M	B	100K	SD	1.45/7.5(character)
Ni [12]	Dir	-	-	F0/d/e/pd	t/WB	M	R	48188	SI	3.65/21.5(character)
Lei [11]	Dir	-	-	F0/d	t/ts/W	M	B	49k	SI	0.7, 1/6, 5.2(character)
Wang [8]	Dir	-	-	F0*/d/e*/pd	SJ	M	R	110K	SD	1.1/4.2(character)
proposed	7B+PS	4-L	U	F0*/d*/e/pd/ed	t/s/f/WL/WB/POS/PM	M	R	60K	SI	9.82/24.4(tonal syllable)

PE: prosodic event = {B: break type | PS: prosodic state | S: phrase stress | A: binary pitch accent | HE: hidden events | AS: accentual structure of words | P: level of pause duration | Dir: direct prosody modeling}; **PH: prosody hierarchy** = {L: layer}; **PL: prosody labeling** = {U: unsupervised | UA: unsupervised adaptation | SS: semi-supervised | S: supervised | R: taking lexical word as potential PW}; **PAF: prosodic-acoustic feature** = {F0: fundamental frequency | d: duration | e: energy | pd: pause duration | ed: energy dip level |*: with differential}; **LF: linguistic feature** = {t: tone | ph: phone | s: base-syllable type | ts: tonal syllable | f: final type | SJ: syllable juncture in a word | W: word | WL: word length | WB: word boundary | POS: part of speech | PM: punctuation mark}; **LNG: language** = {M: Mandarin | E: English | S: Spanish}; **STL: style** = {R: read | B: broadcasting | C: conversational}; **VSZ: vocabulary size in word**, TSR: tonal syllable recognition; **SPK: speaker** = {SI: speaker independent | SD: speaker dependent}; **IMP: improvement in absolute/relative accuracy.**

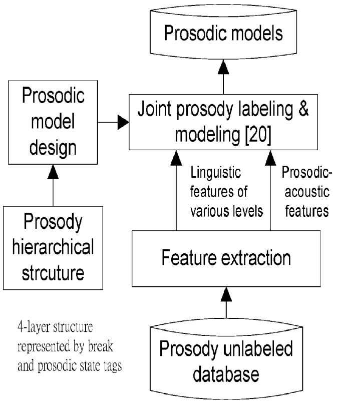

Prosodic models
Prosodic model design
Joint prosody labeling & modeling [20]
Linguistic features of various levels
Prosodic-acoustic features
Prosody hierarchical strcuture
Feature extraction
4-layer structure represented by break and prosodic state tags
Prosody unlabeled database

Fig. 2. Prosody modeling approach in the proposed prosody-assisted ASR system.

word string can be decoded. It includes prosodic tags which implicitly represent the prosody-hierarchy structure of the testing utterance, and some linguistic features such as part-of-speech (POS) and punctuation mark (PM).

The paper is organized as follows. In Section II, the proposed prosody-assisted ASR system for Mandarin speech is presented in detail. Experimental results are discussed in Section III. Some conclusions are given in the last section.

II. PROPOSED PROSODY-ASSISTED ASR SYSTEM

In the proposed prosody-assisted Mandarin-speech ASR system, prosody modeling is first performed and then incorporated into a two-stage speech recognizer by rescoring the word lattice generated by the first-stage speech decoding using the conventional HMM-based ASR recognizer. In the following subsections, we present the design of the proposed prosodic models for ASR, the training of prosodic models by the PLM algorithm [20], and the two-stage speech recognizer in detail.

A. Design of Prosodic Models for ASR

A most commonly agreed and used prosody-hierarchy structure consists of four layers including syllable layer, prosodic word layer, prosodic phrase layer (or intermediate phrase), and intonation phrase layer. Basically, the four-layer structure interprets the pitch and duration variations of syllable well for short sentential utterances. To interpret the contributions of higher-level discourse information to the wider-range and larger variations on the prosodic-acoustic features of long utterances beyond just sentential utterances, Tseng *et al.* [21] proposed a hierarchical prosodic phrase grouping (HPG) model of Mandarin speech. The HPG model consists of five layers, listed in bottom-up order: syllable (SYL), prosodic word (PW), prosodic phrase (PPh), breath group (BG), and prosodic phrase group (PG). The first three layers in the hierarchy are the same as those of the four-layer prosodic structure mentioned above. The fourth BG layer is formed by combining a sequence of PPhs, and a sequence of BGs, in turn, constitutes the fifth PG layer. The above five prosodic constituents are delimited by six break types denoted as $B0$, $B1$, $B2$, $B3$, $B4$, and $B5$ [21]. First, $B0$ and $B1$ represent respectively non-breaks of reduced syllable boundary (or tightly-coupling syllable juncture) and normal syllable boundary, within a PW, which have no identifiable pauses between SYLs. Second, PW boundary $B2$ is perceived as a minor-break boundary where a slight tone of voice change usually follows, while PPh boundary $B3$ is perceived as a clear pause. Third, $B4$ and $B5$ are defined for BG and PG boundaries, respectively. $B4$ is a breathing pause and $B5$ is a complete speech paragraph end characterized by final lengthening coupled with weakening of speech sounds.

In this paper, we adopt a four-layer hierarchy structure, which is a modified version of the HPG model, in the prosody modeling for assisting in ASR to consider the recognition of long Mandarin utterances of paragraphs. The motivation of using the four-layer hierarchy model is owing to its suitability for describing the prosody of long paragraphic utterances of Mandarin. The model employs two types of prosodic tags to repre-

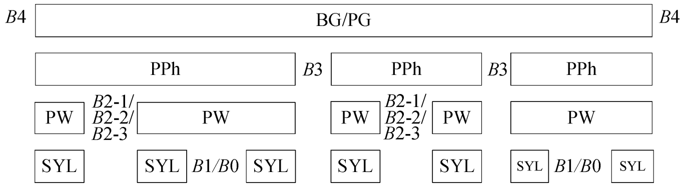


Fig. 3. Prosody-hierarchy model of Mandarin speech used in this study [20], [21].

sent the four-layer prosody-hierarchy structure. One is the break tag used to separate two consecutive prosodic constituents. We modify the break type labeling scheme of the HPG model by dividing $B2$ into three types, $B2-1$, $B2-2$, and $B2-3$, and combining $B4$ and $B5$ into one denoted simply by $B4$. Here, $B2-1$, $B2-2$, and $B2-3$ represent PW boundaries with F0 reset, short pause and pre-boundary syllable duration lengthening, respectively. The reason of refining $B2$ into three types is to consider the difference of their prosodic boundary correlates (i.e., prosodic-acoustic features) to be modeled. On the contrary, the combination of $B4$ and $B5$ is owing to the similarity of their prosodic-acoustic characteristics. Therefore, the break-type tag set used is $\{B0, B1, B2{-}1, B2{-}2, B2{-}3, B3, B4\}$. As shown in Fig. 3, these seven break-type tags can be used to delimit an utterance into four types of prosodic units, namely SYL, PW, PPh, and BG/PG.

Another type of prosodic tag is prosodic state which is conceptually defined as the state in a prosodic phrase to account for the prosodic-acoustic feature variations imposed on higher-level prosodic constituents (i.e., PW, PPh, and BG/PG). The consecutive prosodic state sequence of a prosodic constituent hence forms a prosodic-acoustic feature pattern to characterize it. In practice, prosodic state serves as an intermediate discrete representation of the effects on the variation of a syllable's prosodic-acoustic feature from linguistic features of word-level or above. In this study, three types of prosodic states are used, respectively, for syllable pitch level, syllable duration, and syllable energy level.

Based on the four-layer prosody-hierarchy model, several prosodic models are designed to describe the various relationships of the three types of features: the two types of prosodic tags, the linguistic features of various levels, and the prosodic-acoustic features. The prosodic model design is based on the following maximum *a posterior* (MAP) formulation to find the best linguistic transcriptions $\Lambda_l = \{\mathbf{W}, \mathbf{POS}, \mathbf{PM}\}$, prosodic tags $\Lambda_p = \{\mathbf{B}, \mathbf{P}\}$, and acoustic segmentation Υ_s for the given input acoustic features $\Lambda_a = \{\mathbf{X}_a, \mathbf{X}_p\}$:

$$\begin{aligned}\Lambda_l{}^*, \Lambda_p{}^*, \Upsilon_s{}^* &= \arg\max_{\Lambda_l,\Lambda_p,\Upsilon_s} P(\mathbf{W}, \mathbf{POS}, \mathbf{PM}, \mathbf{B}, \mathbf{P}, \Upsilon_s | \mathbf{X}_a, \mathbf{X}_p)\\ &= \arg\max_{\Lambda_l,\Lambda_p,\Upsilon_s} P(\mathbf{W}, \mathbf{POS}, \mathbf{PM}, \mathbf{B}, \mathbf{P}, \Upsilon_s, \mathbf{X}_a, \mathbf{X}_p) \quad (1)\end{aligned}$$

where $\mathbf{W} = \{w_1^M\}$ is a word sequence; $\mathbf{POS} = \{pos_1^M\}$ is a POS sequence associated with $\mathbf{W}$; $\mathbf{PM} = \{pm_1^M\}$ is a PM sequence; M is the total number of words; $\mathbf{B} = \{B_1^N\}$ is a break type sequence with $B_n \in \{B0, B1, B2-1, B2-2, B2-3, B3, B4\}$; $\mathbf{P} = \{\mathbf{p}, \mathbf{q}, \mathbf{r}\}$ with $\mathbf{p} = \{p_1^N\}$, $\mathbf{q} = \{q_1^N\}$, and $\mathbf{r} = \{r_1^N\}$ representing prosodic state sequences for syllable pitch level, duration, and energy level, respectively; N is the total number of syllables; $\mathbf{X}_a$ is a frame-based spectral feature sequence (i.e., MFCCs and their first-order and second-order derivatives); and $\mathbf{X}_p = \{\mathbf{X}, \mathbf{Y}, \mathbf{Z}\}$ is a prosodic-acoustic feature sequence with $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ representing sequences of syllable-based features, syllable-juncture features, and inter-syllable differential features, respectively. More detailed prosodic-acoustic features are given as: syllable pitch contour (**sp**), syllable energy level (**se**), and syllable duration (**sd**) for $\mathbf{X}$; syllable-juncture pause duration (**pd**) and energy-dip level (**ed**) for $\mathbf{Y}$; and normalized pitch-level jump (**pj**) and two normalized duration lengthening factors (**dl** and **df**) for $\mathbf{Z}$. Notations of tags and features are summarized in Table II.

TABLE II
NOTATIONS OF PROSODIC TAGS, PROSODIC-ACOUSTIC FEATURES AND LINGUISTIC FEATURES

Λ_p : prosodic tags	**B**: break types	
	P: prosodic states	**p**: pitch prosodic states **q**: duration prosodic states **r**: energy prosodic states
$\mathbf{X}_p$: prosodic-acoustic features	**X**: syllable prosodic-acoustic features	**sp**: syllable pitch contours **sd**: syllable duration **se**: syllable energy levels
	Y: syllabe-juncture prosodic-acoustic features	**pd**: pause duration **ed**: energy-dip levels
	Z: inter-syllable differential prosodic-acoustic features	**pj**: normalized pitch-level jumps **dl**: normalized duration lengthening factor 1 **dl**: normalized duration lengthening factor 2
Λ_l : linguistic features	**W**: words **POS**: part-of-speeches **PM**: punctuations marks	
	t: tones **s**: base-syllable types **f**: final types	

To make (1) mathematically tractable, we adopt the following assumptions: 1) Like the conventional acoustic model (AM), spectral feature sequence $\mathbf{X}_a$ depends only on word sequence $\mathbf{W}$; 2) Prosodic-acoustic feature sequence $\mathbf{X}_p$ depends on both prosodic tag sequence Λ_p and linguistic feature sequence Λ_l; 3) Syllable prosodic-acoustic feature sequence $\mathbf{X}$ is independent of syllable-juncture and inter-syllable differential prosodic-acoustic feature sequences, $\mathbf{Y}$ and $\mathbf{Z}$; 4) Break tag sequence $\mathbf{B}$ depends mainly on contextual linguistic feature sequence Λ_l; and 5) Prosodic state sequence $\mathbf{P}$ depends $\mathbf{B}$ only. The reason is that $\mathbf{P}$ is used to characterize the prosodic constituents' patterns which are mainly determined by the prosody hierarchy specified by the break type sequence $\mathbf{B}$. The relation between linguistic features and prosody hierarchy is built through the modeling of $\mathbf{B}$. In other words, the linguistic feature Λ_l can influence the prosodic state through $\mathbf{B}$. We therefore ignore the direct dependency of $\mathbf{P}$ on Λ_l for simplicity. Based on these assumptions, (1) is rewritten as

$$\begin{aligned}\Lambda_l{}^*, \Lambda_p{}^*, \Upsilon_s{}^* \approx \arg\max_{\Lambda_l,\Lambda_p,\Upsilon_s} &\{P(\mathbf{X}_a, \Upsilon_s | \mathbf{W}) P(\mathbf{W}, \mathbf{POS}, \mathbf{PM})\\ &\cdot P(\mathbf{B}|\Lambda_l) P(\mathbf{P}|\mathbf{B}) P(\mathbf{X}|\Upsilon_s, \Lambda_p, \Lambda_l) P(\mathbf{Y}, \mathbf{Z}|\Upsilon_s, \Lambda_p, \Lambda_l)\} \quad (2)\end{aligned}$$

where $P(\mathbf{X}_a, \Upsilon_s|\mathbf{W})$ is an AM; $P(\mathbf{W}, \mathbf{POS}, \mathbf{PM})$ is an LM which describes the relations among $\mathbf{W}$, $\mathbf{POS}$, and $\mathbf{PM}$; $P(\mathbf{B}|\Lambda_l)$ is the break-syntax model which describes how a syllable-juncture break is influenced by the contextual linguistic features of all levels;$P(\mathbf{P}|\mathbf{B})$ is the prosodic state

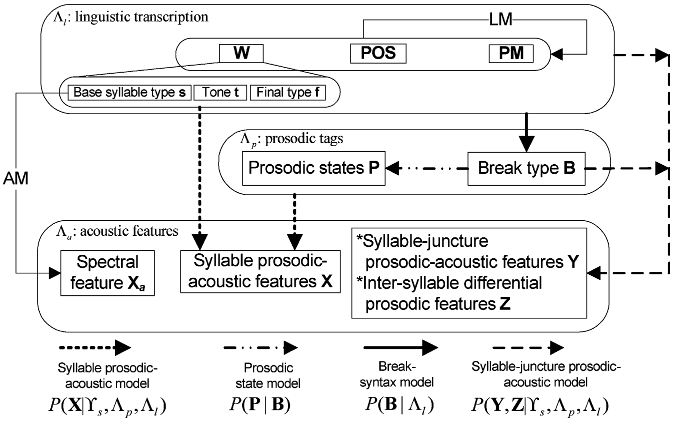

Λ_l: linguistic transcription
LM
W
POS
PM
Base syllable type **s**
Tone **t**
Final type **f**
AM
Λ_p: prosodic tags
Prosodic states **P**
Break type **B**
Λ_a: acoustic features
Spectral feature $\mathbf{X}_a$
Syllable prosodic-acoustic features **X**
*Syllable-juncture prosodic-acoustic features **Y**
*Inter-syllable differential prosodic features **Z**
Syllable prosodic-acoustic model
$P(\mathbf{X}|\Upsilon_s,\Lambda_p,\Lambda_l)$
Prosodic state model
$P(\mathbf{P}|\mathbf{B})$
Break-syntax model
$P(\mathbf{B}|\Lambda_l)$
Syllable-juncture prosodic-acoustic model
$P(\mathbf{Y},\mathbf{Z}|\Upsilon_s,\Lambda_p,\Lambda_l)$

Fig. 4. Relationships of AM, LM, and four prosodic models with prosodic tags, linguistic features, and prosodic-acoustic features.

model describing the variation of prosodic state conditioned on the neighboring break type; $P(\mathbf{X}|\Upsilon_s,\Lambda_p,\Lambda_l)$ is the syllable prosodic-acoustic model which describes the influences of the two types of prosodic tags and the contextual syllable-level linguistic features on the variations of syllable F0 contour, duration and energy level; and $P(\mathbf{Y},\mathbf{Z}|\Upsilon_s,\Lambda_p,\Lambda_l)$ is the syllable-juncture prosodic-acoustic model which describes how the prosodic-acoustic features at or across a syllable juncture are influenced by both the break type of the juncture and the contextual linguistic features. Fig. 4 shows the relationships of features involved in the four prosodic models, LM, and AM.

In implementation, we need to further elaborate these four prosodic models. First, the break-syntax model $P(\mathbf{B}|\Lambda_l)$ is approximated by

$$P(\mathbf{B}|\Lambda_l) \approx \prod_{n=1}^{N-1} P(B_n|\Lambda_{l,n}) \tag{3}$$

where $P(B_n|\Lambda_{l,n})$ is the break type model for the juncture following syllable n, and $\Lambda_{l,n}$ is the contextual linguistic features surrounding syllable n. Since the space of linguistic features $\Lambda_{l,n}$ is large, we partition it into several classes $C(\Lambda_{l,n})$ by the CART decision tree algorithm [22] using the maximum-likelihood gain criterion. The question set used in the CART consists of 216 questions considering the following linguistic features around the juncture: 1) the initial type of the following syllable; 2) interword/intraword indicator; 3) lengths and 4) POSs of the words before and after the juncture if it is an interword; and 5) PM type for an interword juncture.

Second, the prosodic state model $P(\mathbf{P}|\mathbf{B})$ is further divided into three sub-models and approximated as

$$P(\mathbf{P}|\mathbf{B}) = P(\mathbf{p}|\mathbf{B})P(\mathbf{q}|\mathbf{B})P(\mathbf{r}|\mathbf{B}) \approx P(p_1)P(q_1)P(r_1) \cdot\left[\prod_{n=2}^{N} P(p_n|p_{n-1},B_{n-1})P(q_n|q_{n-1},B_{n-1})P(r_n|r_{n-1},B_{n-1})\right] \tag{4}$$

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

Third, the syllable prosodic-acoustic model $P(\mathbf{X}|\Upsilon_s,\Lambda_p,\Lambda_l)$ is further divided into three sub-models and approximated as

$$\begin{aligned}&P(\mathbf{X}|\Upsilon_s,\Lambda_p,\Lambda_l)\\ &\approx P(\mathbf{sp}|\Upsilon_s,\mathbf{B},\mathbf{p},\mathbf{t})P(\mathbf{sd}|\Upsilon_s,\mathbf{B},\mathbf{q},\mathbf{t},\mathbf{s})P(\mathbf{se}|\Upsilon_s,\mathbf{B},\mathbf{r},\mathbf{t},\mathbf{f})\\ &\approx\prod_{n=1}^{N} P(sp_n|p_n,B_{n-1}^n,t_{n-1}^{n+1})P(sd_n|q_n,s_n,t_n)P(se_n|r_n,f_n,t_n)\end{aligned} \tag{5}$$

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

$$sp_n = sp_n^r + \beta_{t_n} + \beta_{p_n} + \beta^f_{B_{n-1},t_{n-1}^n} + \beta^b_{B_n,t_n^{n+1}} + \mu_{sp} \tag{6}$$

where sp_n is a vector of four orthogonally-transformed parameters representing the observed log-F0 contour of syllable n [23]; sp_n^r is the modeling residue; β_{t_n} and β_{p_n} are the affecting patterns (APs) for t_n and p_n, respectively; $\beta^f_{B_{n-1},t_{n-1}^n}$ and $\beta^b_{B_n,t_n^{n+1}}$ are the forward and backward coarticulation APs contributed from syllable $n-1$ and syllable $n+1$, respectively; and μ_{sp} is the global mean of pitch vector. In this study, β_{p_n} is set to have nonzero value only in its first dimension in order to restrict the influence of prosodic state merely on the log-F0 level of the current syllable. By assuming that sp_n^r is zero-mean and normally distributed, i.e., $N(sp_n^r;0,R_{sp})$, we have

$$\begin{aligned}&P(sp_n|p_n,B_{n-1}^n,t_{n-1}^{n+1})\\ &= N(sp_n;\beta_{t_n}+\beta_{p_n}+\beta^f_{B_{n-1},t_{n-1}^n}+\beta^b_{B_n,t_n^{n+1}}+\mu_{sp},R_{sp}).\end{aligned} \tag{7}$$

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

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

$$P(sd_n|q_n,s_n,t_n)=N(sd_n;\gamma_{t_n}+\gamma_{s_n}+\gamma_{q_n}+\mu_{sd},R_{sd}) \tag{8}$$

$$P(se_n|r_n,f_n,t_n)=N(se_n;\omega_{t_n}+\omega_{f_n}+\omega_{r_n}+\mu_{se},R_{se}). \tag{9}$$

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

Lastly, the syllable-juncture prosodic-acoustic model is further divided into five sub-models and approximated as

$$\begin{aligned}&P(\mathbf{Y},\mathbf{Z}|\Upsilon_s,\Lambda_p,\Lambda_l)\\ &\approx P(\mathbf{pd},\mathbf{ed},\mathbf{pj},\mathbf{dl},\mathbf{df}|\Upsilon_s,\Lambda_p,\Lambda_l)\\ &\approx\prod_{n=1}^{N-1} P\left(pd_n,ed_n,pj_n,dl_n,df_n\,|\Upsilon_s,B_n,\Lambda_{l,n}\right)\end{aligned}$$

$$\approx \prod_{n=1}^{N-1} \Big\{ g(pd_n; \alpha_{B_n,\Lambda_{l,n}}, \eta_{B_n,\Lambda_{l,n}}) N(ed_n; \mu_{ed,B_n,\Lambda_{l,n}}, \sigma^2_{ed,B_n,\Lambda_{l,n}}) \\ \cdot N(pj_n; \mu_{pj,B_n,\Lambda_{l,n}}, \sigma^2_{pj,B_n,\Lambda_{l,n}}) N(dl_n; \mu_{dl,B_n,\Lambda_{l,n}}, \sigma^2_{dl,B_n,\Lambda_{l,n}}) \\ \cdot N(df_n; \mu_{df,B_n,\Lambda_{l,n}}, \sigma^2_{df,B_n,\Lambda_{l,n}}) \Big\} \quad (10)$$

where $g(pd_n; \alpha_{B_n,\Lambda_{l,n}}, \eta_{B_n,\Lambda_{l,n}})$ is a Gamma distribution for pause duration pd_n of the juncture following syllable n (referred to as juncture n hereafter); ed_n is the energy-dip level of juncture n and is modeled by a normal distribution:

$$pj_n = (sp_{n+1}(1) - \beta_{t_{n+1}}(1)) - (sp_n(1) - \beta_{t_n}(1)) \quad (11)$$

is the normalized pitch-level jump across juncture n; $sp_n(1)$ is the first dimension of syllable pitch contour sp_n (i.e., syllable pitch level); $\beta_{t_n}(1)$ is the first dimension of the tone AP;

$$dl_n = (sd_n - \gamma_{t_n} - \gamma_{s_n}) - (sd_{n-1} - \gamma_{t_{n-1}} - \gamma_{s_{n-1}}) \quad (12)$$
$$df_n = (sd_n - \gamma_{t_n} - \gamma_{s_n}) - (sd_{n+1} - \gamma_{t_{n+1}} - \gamma_{s_{n+1}}) \quad (13)$$

are two normalized duration lengthening factors before and across juncture n. Both dl_n and df_n are modeled as normal distribution. Since the space of $\Lambda_{l,n}$ is large, the CART algorithm with the node splitting criterion of maximum-likelihood (ML) gain is adopted to concurrently classify the five features of pd_n, ed_n, pj_n, dl_n, and df_n for each break type according to the same question set used in the training of the break-syntax model. Each leaf node represents the product of the five sub-models. So, seven decision trees are constructed for the syllable-juncture prosodic-acoustic model. It is noted that normal distribution is used to model ed_n, pj_n, dl_n, and df_n because of its simplicity and fit to the real data distribution. As for pd_n, normal distribution is not suitable because pd_n is distributed unsymmetrically due to the restriction of nonnegative and the tendency of small value for some break types such as B0 and B1. Like the state duration of phone HMM model, Gamma distribution is suitable for this kind of data.

B. Training of the Proposed Prosodic Models

The joint prosody labeling and modeling (PLM) algorithm proposed previously [20] is adopted to train all these 12 models from an unlabeled speech database. The PLM algorithm is a sequential optimization procedure based on the ML criterion to jointly label the prosodic tags for all utterances of the training corpus and estimate the parameters of all 12 prosodic models. It is composed of two parts: initialization and iteration. The initialization part first determines initial prosodic tags of all utterances, and then estimates initial parameters of the prosodic models by a specially designed procedure. The iteration part first defines an objective likelihood function for each utterance by

$$Q = \\ \left(\prod_{n=1}^{N-1} P(B_n|\Lambda_{l,n}) \right) \\ \cdot \left(P(p_1)P(q_1)P(r_1) \left[\prod_{n=2}^{N} \begin{matrix} P(p_n|p_{n-1},B_{n-1})P(q_n|q_{n-1},B_{n-1}) \\ P(r_n|r_{n-1},B_{n-1}) \end{matrix} \right] \right) \\ \cdot \left(\prod_{n=1}^{N} P(sp_n|p_n,B_{n-1}^{n},t_{n-1}^{n+1})P(sd_n|q_n,s_n,t_n)P(se_n|r_n,f_n,t_n) \right) \\ \cdot \left(\prod_{n=1}^{N-1} \begin{matrix} g(pd_n;\alpha_{B_n,\Lambda_{l,n}},\beta_{B_n,\Lambda_{l,n}})N(ed_n;\mu_{ed,B_n,\Lambda_{l,n}},\sigma^2_{ed,B_n,\Lambda_{l,n}}) \\ N(pj_n;\mu_{pj,B_n,\Lambda_{l,n}},\sigma^2_{pj,B_n,\Lambda_{l,n}})N(dl_n;\mu_{dl,B_n,\Lambda_{l,n}},\sigma^2_{dl,B_n,\Lambda_{l,n}}) \\ N(df_n;\mu_{df,B_n,\Lambda_{l,n}},\sigma^2_{df,B_n,\Lambda_{l,n}}) \end{matrix} \right) \quad (14)$$

It then performs a multistep iterative procedure to relabel the prosodic tags of each utterance with the goal of maximizing Q and update the parameters of all prosodic models sequentially and iteratively. In the following, we describe the sequential optimization procedure in more detail.

1) Initialization:

Initially Labeling of Break Indices: The initial break index of each syllable juncture is determined by a decision tree shown in Fig. 5. The decision tree is designed based on the general knowledge of the break types obtained in our previous prosody labeling and modeling study on a single-speaker database [20]. First, a juncture is labeled as $B4$ if its pause duration is longer than a large threshold $Th1$. Then, it is assigned as $B3$ if its pause duration is longer than $Th2$. Then, all intrawords are labeled as $B0/B1$. We then mark interwords with medium pause duration ($\geq Th3$) as $B2-2$, with medium pitch jump ($\geq Th4$) as $B2-1$, and with medium pre- or post-syllable lengthening ($\geq Th5$ and $Th6$) as $B2-3$. All remaining interwords are labeled as $B0/B1$. Lastly, $B0/B1$ are refined as $B0$ if the syllable juncture has continuous F0 trajectory; otherwise, it is labeled as $B1$. All these six thresholds are determined in a systematic way by an algorithm to avoid determining them by trial-and-error. The algorithm is discussed in detail as follows.

The algorithm is designed using both linguistic and acoustic cues to determine these six thresholds. First, we consider that PMs are usually associated with long breaks and assigned to $B3$ or $B4$. We hence collect the pause durations of all word junctures with PM and use scalar quantization to divide them into two clusters. Two gamma distributions are accordingly constructed to stand for pause duration distributions of $B4$ and $B3$, i.e., $f_{B3}(pd)$ and $f_{B4}(pd)$, respectively. The threshold $Th1$ is then set to be the equal probability intersection between the two distributions. Then, we construct a Gamma distribution $f_{B0/1}(pd)$ for $B0/B1$ by using the pause durations of all intrawords. Another Gamma distribution $f_{B2-2}(pd)$ for $B2-2$ is then constructed by using the pause durations of all non-PM interword junctures with apparent pause durations defined based on the criterion of $f_{B3}(pd) > f_{B0/1}(pd)$. This can exclude non-PM interwords with pause duration similar to those of $B0/B1$. The thresholds $Th2$ and $Th3$ are then set to be the equal probability intersections of $f_{B2-2}(pd)/f_{B3}(pd)$ and $f_{B2-2}(pd)/f_{B0/1}(pd)$.

We then determine the three thresholds, $Th4$, $Th5$, and $Th6$, which are used to label initial $B2-1$ and $B2-3$. First, six Gaussian distributions of the normalized F0 jump and the two duration lengthening factors, i.e., $f_{\text{PM}}(pj)$, $f_{\text{intra}}(pj)$, $f_{\text{PM}}(dl)$, $f_{\text{intra}}(dl)$, $f_{\text{PM}}(df)$, and $f_{\text{intra}}(df)$, for both PM and intraword are constructed using data of interwords with PM and of intrawords, respectively. Then, a Gaussian distribution of pj for $B2-1$, i.e., $f_{B2-1}(pj)$, is constructed using non-PM interwords with apparent pitch jump defined based on the criterion of $f_{\text{PM}}(pj) > f_{\text{intra}}(pj)$. Similarly, two Gaussian distributions of dl and df for $B2-3$, i.e., $f_{B2-3}(dl)$ and $f_{B2-3}(df)$, are constructed using non-PM interwords with apparent duration lengthening defined based on the criteria of $f_{\text{PM}}(dl) > f_{\text{intra}}(dl)$ and $f_{\text{PM}}(df) > f_{\text{intra}}(df)$. Lastly, $Th4$, $Th5$ and $Th6$ are set to be the equal probability inter-

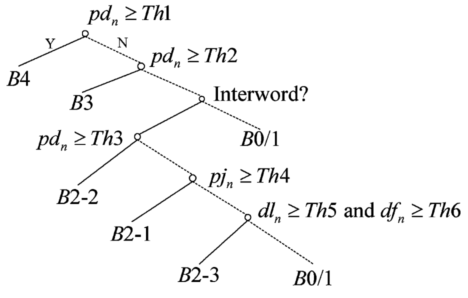


Fig. 5. Decision tree for initial break type labeling.

sections of $f_{\text{intra}}(pj)/f_{B2-1}(pj)$, $f_{\text{intra}}(dl)/f_{B2-3}(dl)$ and $f_{\text{intra}}(df)/f_{B2-3}(df)$.

Initialization of 12 Prosodic Models: The initializations of the break-syntax model and the syllable-juncture prosodic-acoustic model can be done independently with initial break indices of all syllable junctures being given. We realize them by the CART algorithm [22]. Then, the initializations of the three syllable prosodic-acoustic models are considered. Since they are multi-parametric representation models to superimpose several APs of major affecting factors to form the observed syllable prosodic-acoustic features, 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 effects from the surface prosodic-acoustic features for the estimations of the remaining APs. Based on this idea, we determine the order of initial AP estimation according to the availability of affecting factor and the size of AP. The resulting ordering is listed as follows: global means $\mu_{sp}/\mu_{sd}/\mu_{se}$, tone $\beta_t/\gamma_t/\omega_t$, coarticulation $\beta_{B,t}^f/\beta_{B,t}^b$, base-syllable/final type γ_s/ω_f, and prosodic states $\beta_p/\gamma_q/\omega_r$. It is noted that an improper ordering AP estimation may result in poor AP estimates. For example, if we reverse the order of initial estimation of tone and base-syllable APs (i.e., γ_t and γ_s) of syllable duration, then the value of γ_s for base-syllable "de" will decrease significantly while the value of γ_t for Tone 5 will increase accordingly. This is due to the high-frequency character "的" which dominates both distributions of Tone 5 and base-syllable "de". We also note that the initial pitch, duration and energy prosodic-state indices are assigned by applying vector quantization (VQ) to the residues of syllable F0 level, duration, and energy level, respectively; and their APs are set to be the corresponding codewords. Lastly, the initializations of the three prosodic state transition models are done using the labeled prosodic-state indices and break indices.

2) Iteration: The iteration is a multistep procedure listed as follows:

Step 1) Update the APs of tones $\beta_t/\gamma_t/\omega_t$ with all other APs being fixed.

Step 2) Update the APs of coarticulation $\beta_{B,t}^f/\beta_{B,t}^b$ with all other APs being fixed.

Step 3) Update the APs of base-syllable/final type, γ_s/ω_f, with all other APs being fixed.

Step 4) Relabel the prosodic state sequence of each utterance by the Viterbi algorithm so as to maximize Q defined in (14).

Step 5) Update the APs of prosodic state, $\beta_p/\gamma_q/\omega_r$, variances, $R_{sp}/R_{sd}/R_{se}$, and the prosodic state transition model.

Step 6) Relabel the break type sequence of each utterance by the Viterbi algorithm so as to maximize Q defined in (14).

Step 7) Update the decision trees of the break-syntax model and of the syllable-juncture prosodic-acoustic model.

Step 8) Repeat Steps 1 to 7 until a convergence is reached.

C. Two-Stage Prosody-Assisted ASR System

Fig. 6 displays a block diagram of the proposed two-stage prosody-assisted ASR system. It first uses the conventional HMM-based word recognizer with a syllable-based AM and a word-bigram LM in the first stage to generate a word lattice. It then employs a factored LM (FLM) [24] and the 12 prosodic models discussed above in the second stage to rescore the word lattice and find the best recognition result. Here the FLM is an extension of the conventional word-based LM to jointly describe the relations of the word sequence $\mathbf{W}$, the part-of-speech sequence $\mathbf{POS}$, and the punctuation mark sequence $\mathbf{PM}$. The FLM is composed of a word-trigram model, a factored POS model and a factored PM model, and is formulated as

$$P(\mathbf{W}, \mathbf{PM}, \mathbf{POS}) \approx \prod_{i=1}^{M} \left\{ \underbrace{P(w_i|w_{i-2}^{i-1})}_{\text{word-trigram LM}} \underbrace{\cdot P(pos_i|pos_{i-1}, w_i)}_{\text{factored POS model}} \underbrace{P(pm_{i-1}|pos_{i-1}^{i}, w_{i-1})}_{\text{factored PM model}} \right\} \tag{15}$$

Here, the FLM approach [24] is applied to the modeling of the two factored models of POS and PM. The SRILM toolkit [25] with Witten–Bell smoothing is used to train these three models.

In the second-stage rescoring process, a product of 16 probabilities from three types of models (i.e., AM, FLM, and prosodic models) is computed as we completely expand the speech decoding equation shown in (2). For considering the relative importance of each individual model to ASR, a log-linear combination scheme to integrate these 16 probabilities is adopted in this study:

$$L(S, \Lambda_\alpha) = \log C\left(\Lambda_\alpha\right) + \sum_{j=1}^{16} \alpha_j \log p_j \tag{16}$$

where $S = [p_1 \cdots p_{16}]$ is a 16-dimensional vector formed by these 16 probabilities; $\Lambda_\alpha = [\alpha_1 \cdots \alpha_{16}]$ is a weighting vector; and $C\left(\Lambda_\alpha\right)$ is a normalization factor. The discriminative model combination (DMC) method [26] is employed to find the optimal weighting vector for minimizing the word error rate on a development set. The DMC method uses the well-known Generalized Probabilistic Descent (GPD) algorithm [27] to iteratively minimize a smoothed empirical word error rate on the development set.

III. Experimental Results

A. Database and Experiment Setting

The proposed ASR method was tested on a large Mandarin read speech database TCC300 [28]. The database consists of two sets: 103-speaker short sentential utterances (Set A) and 200-speaker long paragraphic utterances (Set B). The database was collected for Mandarin ASR. Set A was designed to consider the phonetic balance of Mandarin speech, while Set B was designed to additionally consider the usage for prosody study.

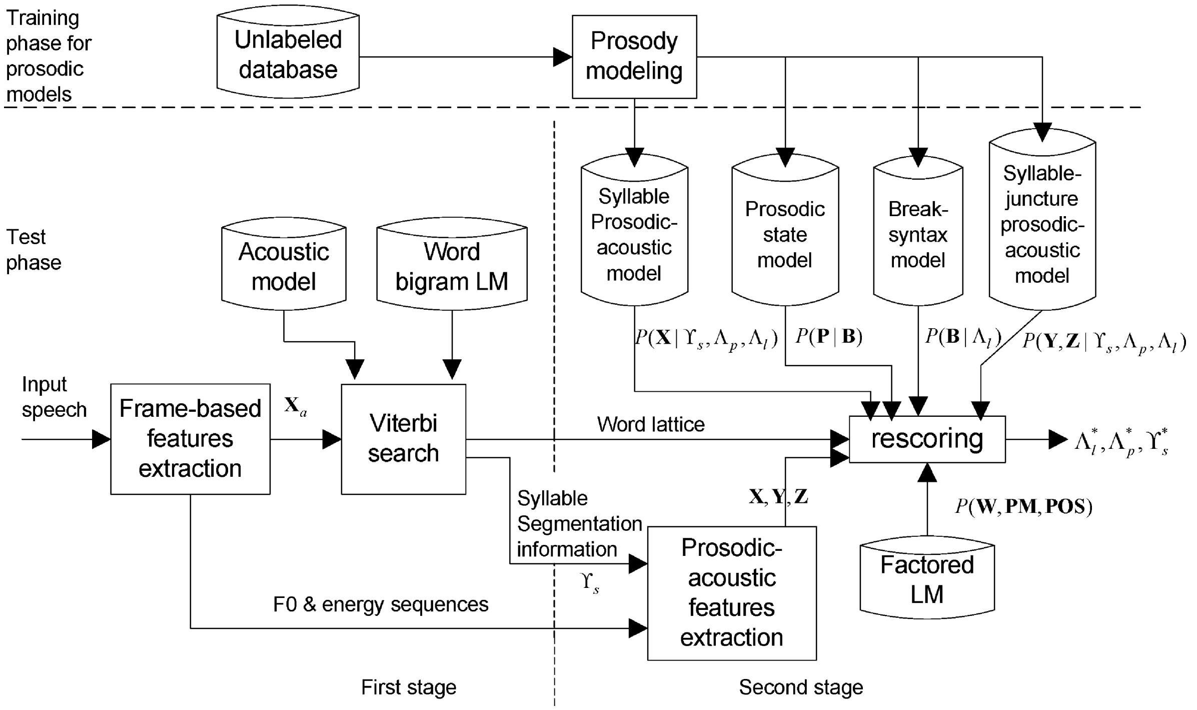

Training phase for prosodic models
Unlabeled database
Prosody modeling
Test phase
Acoustic model
Word bigram LM
Syllable Prosodic-acoustic model
Prosodic state model
Break-syntax model
Syllable-juncture prosodic-acoustic model
$P(\mathbf{X}|\Upsilon_s,\Lambda_p,\Lambda_l)$
$P(\mathbf{P}|\mathbf{B})$
$P(\mathbf{B}|\Lambda_l)$
$P(\mathbf{Y},\mathbf{Z}|\Upsilon_s,\Lambda_p,\Lambda_l)$
Input speech
Frame-based features extraction
$\mathbf{X}_a$
Viterbi search
Word lattice
rescoring
$\Lambda_l^*,\Lambda_p^*,\Upsilon_s^*$
Syllable Segmentation information
Υ_s
X,Y,Z
$P(\mathbf{W},\mathbf{PM},\mathbf{POS})$
Prosodic-acoustic features extraction
Factored LM
F0 & energy sequences
First stage
Second stage

Fig. 6. Block diagram of the two-stage prosody-assisted ASR system.

The database was divided into a training set (about 90%, 274 speakers, 23 hours) and a test set (about 10%, 29 speakers, 2.43 hours). A set of 411 eight-state base-syllable HMM models was generated from the training set by HTK 3.4 [29] with the MMIE criterion [30]. The acoustic feature vector is composed of 12 MFCCs and their delta and delta-delta terms, 1 delta energy and 1 delta-delta energy. For testing the proposed prosody-assisted ASR system, the Set B part of the test set was used. The test subset contained 226 utterances of 19 speakers with length about 2 hours. The total number of words in the test subset is 14 993. All testing data were long utterances with average length of 117.2 syllables.

A text corpus was employed to train both the word-bigram LM and the FLM which were used, respectively, in the first- and second-stage speech decodings. The corpus contained in total about 139 million words and was formed by combining the following three corpora: 1) Sinorama: a news magazine with 9.87 million words; 2) NTCIR: an information retrieval (IR) test bench consisting of several domains with 124.4 million words; and 3) Sinica Corpus: a general text corpus comprising 4.8 million words with manually POS tagging. The POS tags used in this study are the same as those used in the syntactic parsing of the Sinica Treebank [31]. There are in total 46 types of POS. A conditional random field CRF-based tagger was employed to segment all texts in the corpus into word-POS sequences. The tagger was trained on the Sinica Corpus. For simplicity, PMs were categorized into four classes: comma, period, major PM (including dot, exclamation mark, question mark, semicolon, and colon), and non-PM. A 60 000-word lexicon was also constructed based on word frequency.

B. Prosody Modeling

A training subset containing utterances of 164 speakers was used for prosody modeling. It was selected from the training set and consisted of long paragraphic utterances with prosody being properly pronounced. A subjective judgment based on the rhythm and melody of an utterance was applied to determine whether it was properly pronounced. Two major types of ill-pronounced utterances were found: 1) bad rhythm—read each character isolatedly to insert a pause after every character; and 2) bad melody—read each character with almost the same pitch level to result in a flat intonation. The excluding of those ill-pronounced training utterances could avoid polluting the generated prosodic models so as to degrade their effectiveness on assisting in ASR. The total length of the training subset was about 8.3 hours. All speech signals were time-aligned using the 411 base-syllable HMM models mentioned above. Five prosodic-acoustic features were then extracted, including syllable pitch contour vector, syllable duration, syllable energy level, and syllable-juncture pause duration and energy-dip level. It is noted that syllable pitch contour vectors were extracted from the frame-based F0 values normalized by speaker-level mean and variance; while both syllable duration and syllable energy level were normalized by their corresponding speaker-level means and variances. It is also noted that the three inter-syllable differential prosodic-acoustic features [i.e., pj_n, dl_n, and df_n defined in (11)–(13)] were obtained automatically in the prosodic model training by the PLM algorithm [20]. The texts of the training subset were processed by the CRF-based tagger mentioned previously to extract all linguistic features needed in the prosody modeling. The PLM algorithm [20] was then applied to automatically generate the 12 prosodic models from the training subset. In realizing the PLM algorithm, the numbers of pitch, duration and energy prosodic states were all set to be 16. For avoiding over-fitting the decision trees of the break-syntax model and the syllable-juncture prosodic-acoustic model, the following two stop criteria were used: 1) The size of a leaf node must be larger than 700 syllables; and 2) The relative improvement of likelihood must be larger than 0.0065 in node splitting. These two values were determined empirically. Finally, the total numbers of nodes (leaf nodes) obtained were 63(31) and 46(27) for these two models, respectively.

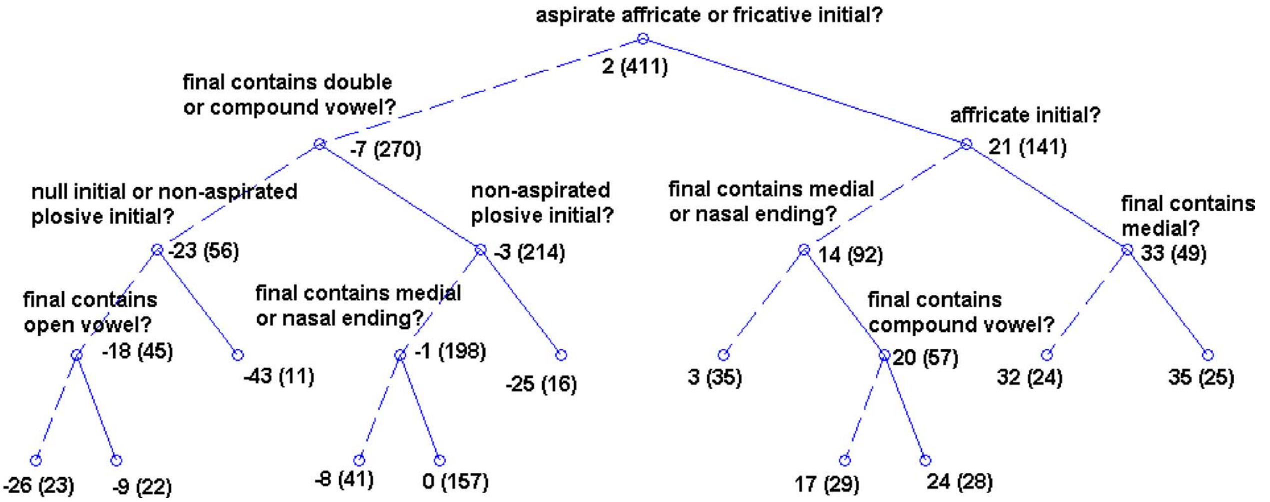

aspirate affricate or fricative initial?
2 (411)
final contains double or compound vowel?
-7 (270)
affricate initial?
21 (141)
null initial or non-aspirated plosive initial?
-23 (56)
non-aspirated plosive initial?
-3 (214)
final contains medial or nasal ending?
14 (92)
final contains medial?
33 (49)
final contains open vowel?
-18 (45)
-43 (11)
final contains medial or nasal ending?
-1 (198)
-25 (16)
3 (35)
final contains compound vowel?
20 (57)
32 (24)
35 (25)
-26 (23)
-9 (22)
-8 (41)
0 (157)
17 (29)
24 (28)

Fig. 7. Decision tree analysis of duration APs of all 411 base-syllables. Numbers associated with each leaf node represents the average length (ms) of the APs and the data count (in the bracket). Solid line indicates positive answer to the question and dashed line indicates negative answer.

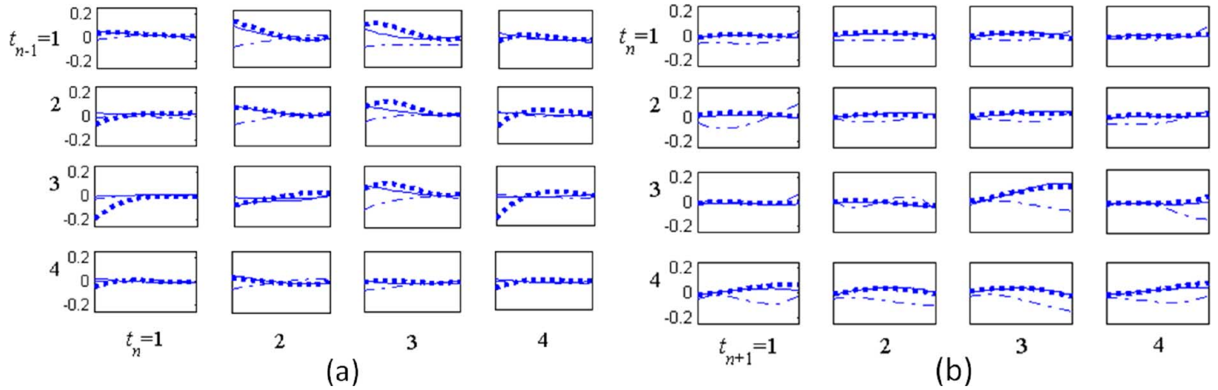


Fig. 8. (a) Forward and (b) backward coarticulation patterns, $\boldsymbol{\beta}^f_{B_{n-1},t^n_{n-1}}$ and $\boldsymbol{\beta}^b_{B_n,t^{n+1}_n}$, for $B0$ (point line), B1(solid line), and B4(dashed line).

TABLE III
APs of Five Tones

Tone	1	2	3	4	5
Pitch mean (log-Hz)	0.097	-0.050	-0.110	0.065	-0.069
Duration (ms)	9	5	-5	5	-54
Energy level (dB)	0.874	-0.623	-0.785	0.840	-1.567

A quantitative analysis of the prosody modeling result is given as follows. Table III shows the APs of five tones. As shown in the table, Tone 1 and Tone 4 had high pitch mean, long duration and high energy level; while Tone 3 and Tone 5 had low pitch mean, short duration and low energy level. It is noted that a negative value of tone AP of syllable duration means the length of a syllable with this tone type is smaller than the average length of all syllables with the same base-syllable type regardless of their tone type. These agreed with the prior linguistic knowledge and generally matched with those of other previous studies [32], [33].

Fig. 7 displays the decision-tree analysis of the duration APs of all 411 base-syllables. It can be found from the figure that the base-syllables with aspirated affricate (q, ch, c) or fricative (f, h, x, sh, s) initials were much longer in average than all other base-syllables. On the other hand, base-syllables with more vowel components (double/compound vowel), medial, or nasal ending in final were generally longer. These results were also confirmed in the previous study [33].

Fig. 8 depicts the forward and backward coarticulation patterns for the three extreme cases of break types, i.e., $B0$ (tightly coupling), $B1$ (normal) and $B4$ (major break). Several characteristics of these APs can be found. First, the forward coarticulations mainly affected the beginning parts of syllable pitch contours, while the backward coarticulations affected the ending parts. Secondly, we find from the dynamic ranges of these APs that the coarticulation effect was the most serious for $B0$ junctures and the least for $B4$ junctures. Third, for tightly coupling $B0$ junctures, most coarticulation APs demonstrated well the effect to compensate for tone concatenation mismatch of their pitch contours. For example, the upward bending at the beginning parts of $\{\beta^f_{B,t}|t^n_{n-1} = (1,2),(1,3),(2,2),(2,3)\}$ were due to H-L mismatches, while the downward bending at the beginning parts of $\{\beta^f_{B,t}|t^n_{n-1} = (3,1),(3,4)\}$ corresponded to L-H mismatches. Fig. 9(a) illustrates the effect of the forward coarticulation AP of Tone 1 in the 1–3 tone pair on raising the beginning part of the following Tone 3 pitch pattern in order to be better matched with the high ending level of the preceding Tone 1 pitch pattern. Fourth, the well-known *sandhi* rule that Tone 3-Tone 3 will change to Tone 2-Tone 3 has been learned in the backward coarticulation AP of 3–3 tone pair. Fig. 9(b) illustrates this effect. Lastly, the forward coarticulations are generally larger than the backward coarticulations. The above mentioned characteristics generally conformed well to the observation found by Xu [34].

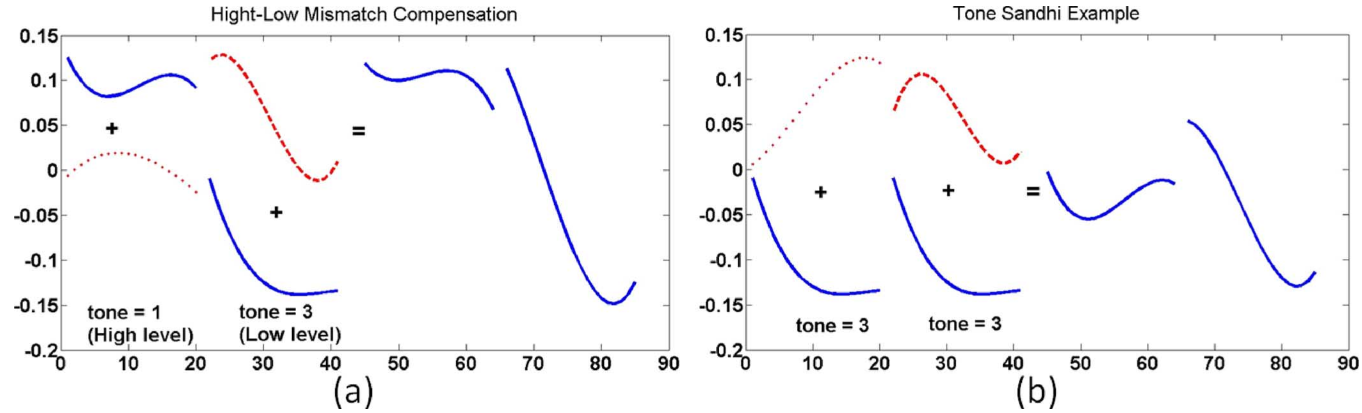


Fig. 9. Two examples demonstrate the effects of coarticulation APs. (a) Tone 1-Tone 3 and (b) the *sandhi* rule of Tone 3-Tone 3. Solid lines (left): basic tone pitch patterns; point lines: backward APs; dashed lines: forward APs; and solid lines (right): the resulting pitch patterns.

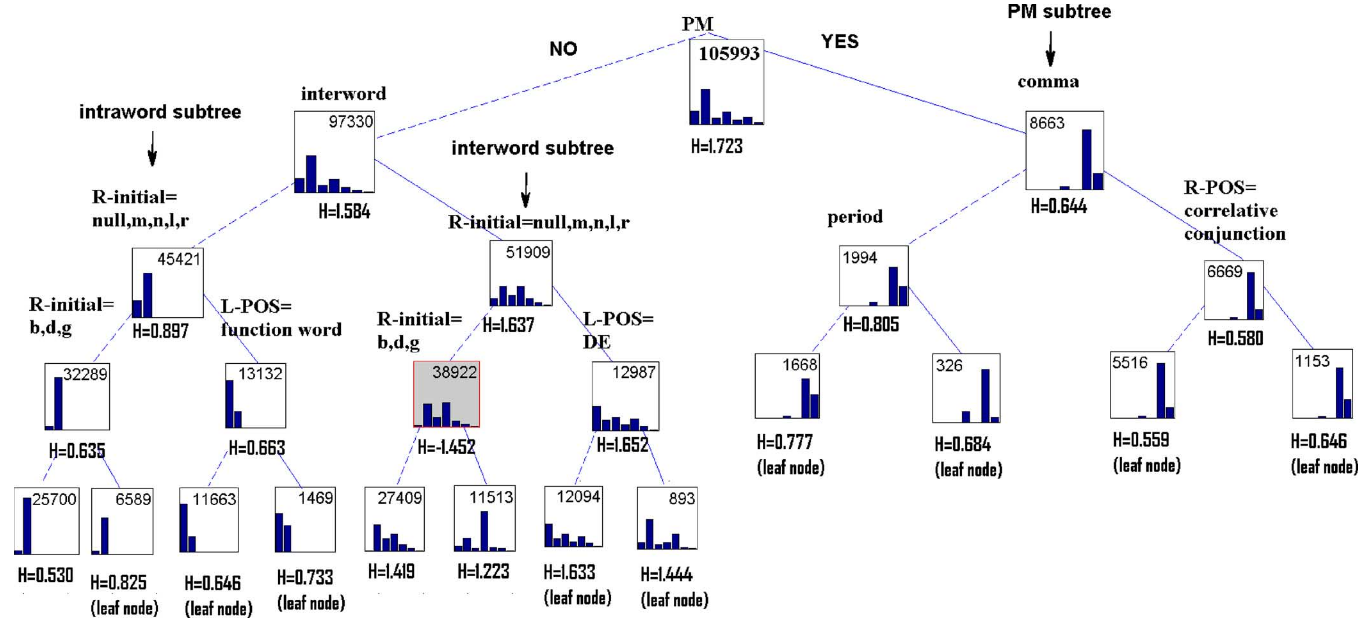


Fig. 10. Decision tree for the break-syntax model. The bar plot associated with a node denotes the distribution of these seven break types ($B0$, $B1$, $B2-1$, $B2-2$, $B2-3$, $B3$, $B4$, from left to right) and the number is the total data count of the node. H is the Shannon entropy to measure the uncertainty of break type distribution.

Fig. 10 displays the major part of the decision tree of the break-syntax model. As shown in the figure, the entropy of the break type distribution decreased as we traced down the decision tree with more linguistic features being involved. The most important linguistic features used in the decision tree were PM and interword/intraword. The two sub-trees corresponding to PM and intraword were relatively simpler with the entropy of the break type distribution decreasing fast, while the sub-tree of interword was very complicated with the entropy decreasing slowly. Besides, the break type distributions of the nodes in the PM sub-tree concentrated mainly on $B3$ and $B4$, while they were on $B0$ and $B1$ for nodes in the intraword sub-tree. Moreover, phonetic information was important for the intraword sub-tree to further discriminate between $B0$ and $B1$. For the PM sub-tree, the type of PM was important. Fig. 11 displays a deeper part of the interword sub-tree. Major linguistic features used were: "stop" initial in the following syllable, content/function word, the word "DE," and various types of POS.

Fig. 12 shows the major parts of decision trees of the break-acoustic model for the seven break types. We can find from the statistics of root notes that the break types of higher level were generally associated with longer pause duration, lower energy-dip level, larger normalized pitch-level jump, and larger duration lengthening factors. Besides, $B2-3$ was similar to $B1$ and $B2-1$ in the distributions of pause duration, and energy-dip level. $B2-1$, $B3$, and $B4$ had positive normalized pitch jumps in average, while $B0$, $B1$, and $B2-3$ had negative ones. These results illustrated the declination and reset effects of log-F0 at intra-PW and inter-PW syllable boundaries, respectively. The two normalized duration lengthening factors for $B2-2$, $B2-3$, $B3$, and $B4$ were relatively larger than those of $B0$, $B1$, and $B2-1$. These distributions showed the lengthening effect for the last syllable of PW, PPh, and PG/BG.

For each break type, the likelihood of the syllable-juncture prosodic-acoustic modeling increases as we traced down these decision trees with more linguistic features being involved. This means the use of linguistic features can improve the modeling of syllable-juncture prosodic-acoustic features. It is noted here that no tree-splitting occurred for $B4$ due to the relative uniformity on the prosodic-acoustic prosodic features of its data. The questions used to split trees of pause-related break types (i.e., $B3$ and $B2-2$) tended to be related to higher-level syntactic

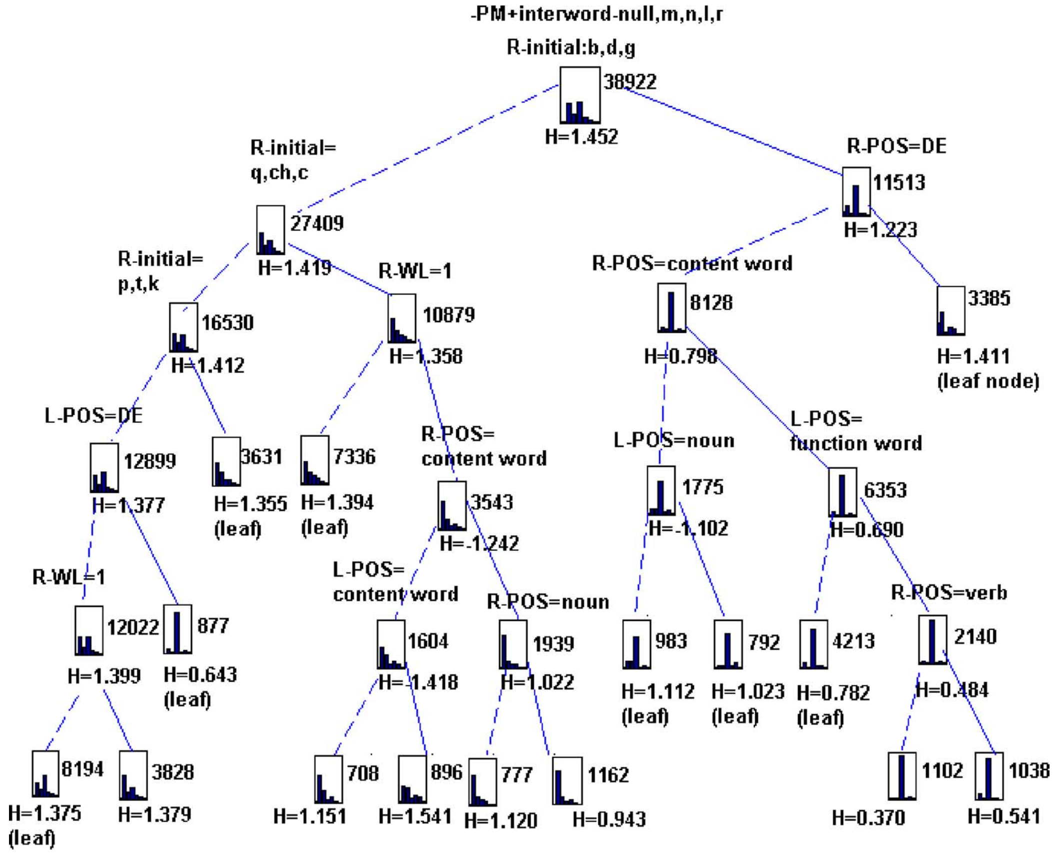


Fig. 11. Deeper part of the decision tree for the break-syntax model. It is the sub-tree starting from the shaded node shown in Fig. 10.

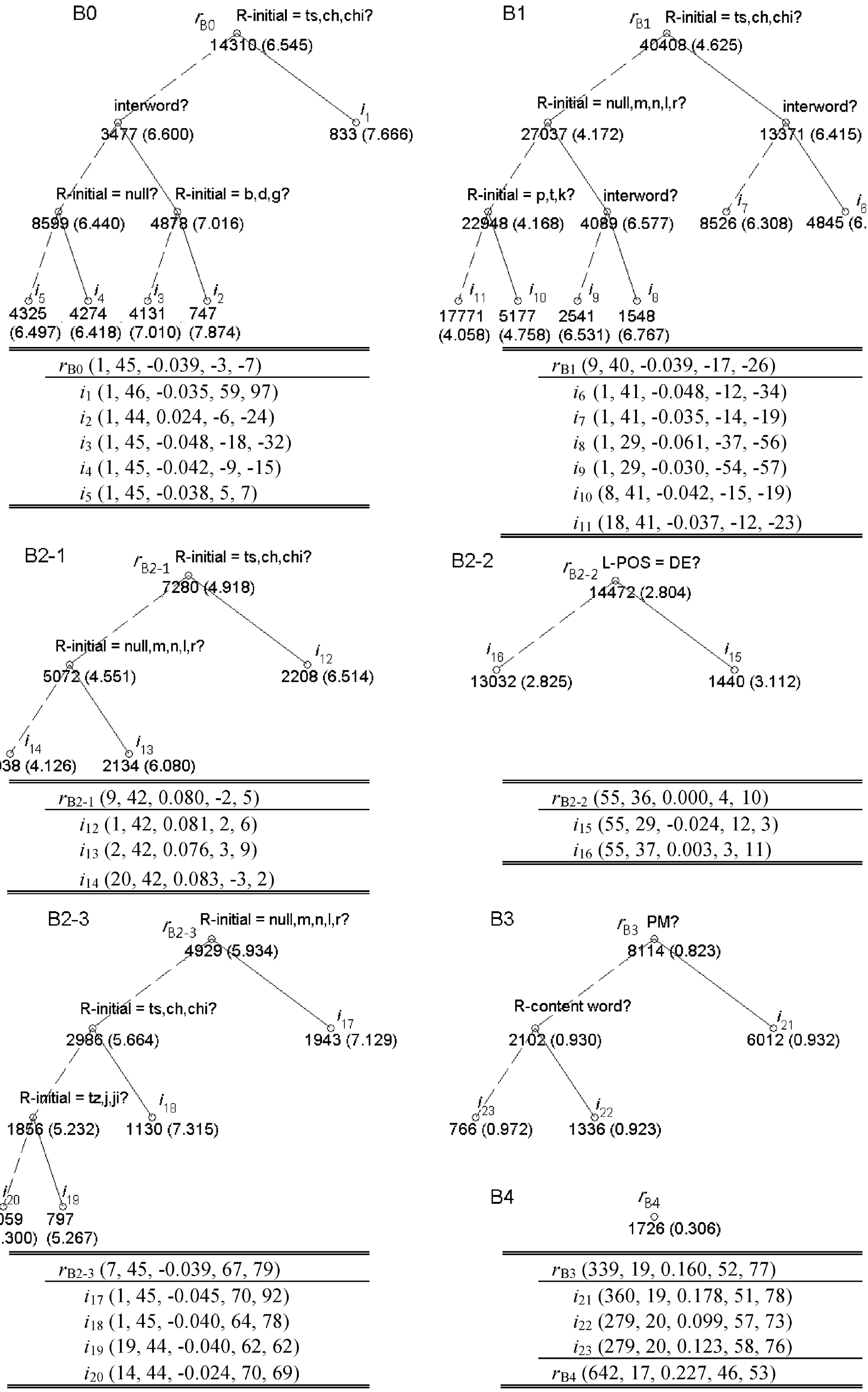
features, such as PM and POS. On the contrary, the questions of lower-level linguistic features, such as interword/intraword and phonetic features, were used to split trees of lower-level non-pause break types (i.e., $B0$, $B1$, $B2-1$, and $B2-3$).

Fig. 13 illustrates the transitions of pitch prosodic state $P(p_n|p_{n-1}, B_{n-1})$ for seven break types. For $B0$ and $B1$, the general high-to-low, nearby-state transitions showed that the syllable log-F0 level declined slowly within PWs. For $B2-2$, it had both high-to-low and low-to-high state transitions. For $B2-1$, $B3$, and $B4$, their low-to-high state transitions showed clearly the phenomena of syllable log-F0 level resets across PWs, PPhs, and BG/PGs. Comparing with these clear log-F0 level resets, the resets of $B2-2$ were insignificant. The transition of $B2-3$ is similar to those of $B0$ and $B1$. This implies no apparent pitch reset exists at the duration-lengthening juncture of $B2-3$. These phenomena were similar to those found in our previous study on the database of a single female speaker [20]. Table IV lists a summary of the parameter numbers (#para) of these 12 prosodic models.

C. Recognition Performance Evaluation

We then examined the recognition performance of the proposed prosody-assisted ASR system. We first performed the first-stage decoding by HTK using the 411 base-syllable HMM models and the word-bigram LM to generate a word lattice. We note that the beam-width of the first-stage recognition was set to a large value to make the resulting word lattice have a high cover rate of the correct words. This was to let the study focus mainly on the performance comparison between the scheme with and without using the prosodic models in the second-stage recognition. The WER, CER, and base-syllable error rate (SER) of the first-stage decoding were 29.6%, 21.4%, and 13.7%, respectively. Moreover, the oracle performance (i.e., the cover rate) of the word lattice, which corresponds to the best word string that can be decoded from the lattice, was 9.6%, 9.3%, and 7% for WER, CER, and SER, respectively. The oracle performance approached the upbound as we considered the high out-of-vocabulary (OOV) rate of 4.3% of the test data set. The use of the syllable-based HMM approach was justified by comparing its performance with those of 30.7%, 21.8%, and 13.7% in WER, CER, and SER achieved by the tri-phone HMM recognizer using similar size of total number of states. The syllable-based HMM recognizer we used was slightly better.

We then performed the second-stage decoding. A baseline scheme was first tested using only the FLM in the second-stage rescoring process without involving any prosodic model. Here, we kept the AM scores and replaced the word-bigram LM scores with the FLM scores. In implementation, we needed to expand the first-stage word lattice to consider the applicability of the word-trigram LM, all possible POSs for every candidate word, and four types of PM for every interword location. Besides, the log-linear combination of the scores of AM and the three FLM sub-models was considered. The DMC algorithm [26] was applied to find a set of four weights from a development set selected from the Set B part of the training set. The development set contained 18-minute speech of 33 speakers. For each utterance in the development set, a list of top-100 sequences was found and used in the DMC algorithm. Since the number of weights to be estimated is small, the data of the development set were sufficient. Table V shows the performance of the baseline scheme. The WER, CER, and SER were 24.4%, 18.1%, and 12%. This performance was much better than that of 29.6%,



r_{B0} (1, 45, -0.039, -3, -7)
i_1 (1, 46, -0.035, 59, 97)
i_2 (1, 44, 0.024, -6, -24)
i_3 (1, 45, -0.048, -18, -32)
i_4 (1, 45, -0.042, -9, -15)
i_5 (1, 45, -0.038, 5, 7)

r_{B1} (9, 40, -0.039, -17, -26)
i_6 (1, 41, -0.048, -12, -34)
i_7 (1, 41, -0.035, -14, -19)
i_8 (1, 29, -0.061, -37, -56)
i_9 (1, 29, -0.030, -54, -57)
i_{10} (8, 41, -0.042, -15, -19)
i_{11} (18, 41, -0.037, -12, -23)

r_{B2-1} (9, 42, 0.080, -2, 5)
i_{12} (1, 42, 0.081, 2, 6)
i_{13} (2, 42, 0.076, 3, 9)
i_{14} (20, 42, 0.083, -3, 2)

r_{B2-2} (55, 36, 0.000, 4, 10)
i_{15} (55, 29, -0.024, 12, 3)
i_{16} (55, 37, 0.003, 3, 11)

r_{B2-3} (7, 45, -0.039, 67, 79)
i_{17} (1, 45, -0.045, 70, 92)
i_{18} (1, 45, -0.040, 64, 78)
i_{19} (19, 44, -0.040, 62, 62)
i_{20} (14, 44, -0.024, 70, 69)

r_{B3} (339, 19, 0.160, 52, 77)
i_{21} (360, 19, 0.178, 51, 78)
i_{22} (279, 20, 0.099, 57, 73)
i_{23} (279, 20, 0.123, 58, 76)
r_{B4} (642, 17, 0.227, 46, 53)

Fig. 12. Decision trees of the break-acoustics model for seven break types. Solid (dash) line indicates positive (negative) answer to the question. Numbers in a node are data count and average likelihood (in a bracket). The statistics for each node are shown in the bracket of the tables below the trees. Note that r's represent root node of each break type. Numbers in the bracket, from left to right, denote average pause duration in ms, energy-dip level in dB, normalized pitch jump in log-Hz, and duration lengthening factors 1 and 2 in ms.

21.4%, and 13.7% reached by the ASR using the word-bigram LM.

Lastly, we evaluated the performance of adding prosodic models to the baseline scheme. We first categorized these 12 prosodic models into two classes: juncture-based and syllable-based. The former modeled acoustic cues or phenomena related to different types of juncture and hence was expected to be useful for distinguishing word boundary ambiguity. The latter modeled prosodic-acoustic feature patterns of different types of prosodic constituent so that they were expected to be useful for tone/word discrimination. We hence designed and tested two schemes of incorporating prosodic models. Scheme 1 incorporated the six juncture-based prosodic models, i.e., the break-syntax model and the five syllable-juncture

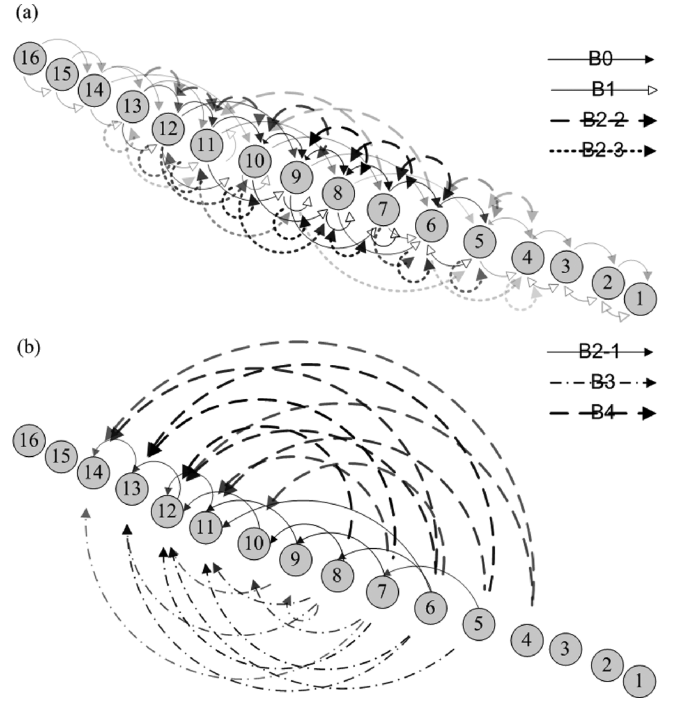


Fig. 13. Most significant prosodic state transitions for (a) $\mathrm{B}0, \mathrm{B}1, \mathrm{B}2-2$, and $\mathrm{B}2-3$, and (b) $\mathrm{B}2-1, \mathrm{B}3$, and $\mathrm{B}4$. Here, the number in each node represents the index of the prosodic state. Note that larger state index represents higher log-$\mathrm{F}0$ value and darker lines represent more important state transitions.

TABLE IV
SUMMARY OF PARAMETER NUMBERS OF 12 PROSODIC MODELS

Model	#para	Description
Break-syntax model	217	31 leaf nodes × 7 break probabilities
Syllable-juncture prosodic-acoustic model	270	27 leaf nodes × 2 parameters for 5 sub-models
Prosodic state model	5424	(16×16×7 + 16 initial probabilities) × 3
Syllable prosodic-acoustic model (APs)	1597	(5 tones+16 states)×3, 40 final types 82 base-syllables, 1400 coarticulations 12 means & variances

TABLE V
RECOGNITION PERFORMANCES OF THE BASELINE SCHEME, SCHEME 1, AND SCHEME 2 (%)

	WER	CER	SER
Baseline scheme	24.4	18.1	12.0
Scheme 1	21.3	15.0	10.2
Scheme 2	20.7	14.4	9.6

prosodic-acoustic sub-models, into the baseline FLM scheme, while Scheme 2 added all 12 prosodic models. In implementation, all values of frame-based F0, syllable duration, and energy level of the testing utterance were normalized by their corresponding utterance-level mean and variance. Here, the syllable segmentation corresponded to the best path of the first-stage decoding. Word lattice expansions were also realized to consider not only the applicability of the FLM like the case of realizing the baseline scheme, but also the incorporation of prosodic models. Two sets of 10 and 16 weights for model combination were respectively found for the two schemes by the DMC algorithm using the same development set. The recognition results are displayed in Table V. As shown in the table, WER, CER, and SER of 21.3%, 15.0%, and 10.2% for Scheme 1, and of 20.7%, 14.4%, and 9.6% for Scheme 2 were obtained. They represented 3.1%, 3.1%, and 1.8% absolute (or 12.7%, 17.1%, and 15% relative) error reductions over the baseline FLM scheme for Scheme 1, and 3.7%, 3.7%, and 2.4% absolute (or 15.2%, 20.4%, and 20% relative) error reductions for Scheme 2. Obviously, Scheme 1 outperformed the baseline scheme significantly, and Scheme 2 was even better. This showed that the word recognition performance could be greatly improved via correcting word segmentation errors by properly using juncture-based break-related information. Moreover, the recognition performance could be further improved slightly via correcting tone errors by modeling tone patterns of prosodic constituents. We can therefore conclude that the prosodic information are useful in ASR.

TABLE VI
EXPERIMENTAL RESULTS OF POS DECODING (%)

	Precision	Recall	F-measure
Baseline scheme	93.4	76.4	84.0
Scheme 2	93.4	80.0	86.2

TABLE VII
EXPERIMENTAL RESULTS OF PM DECODING (%)

	Precision	Recall	F-Measure
Baseline scheme	55.2	37.8	44.8
Scheme 2	61.2	53.0	56.8

Aside from generating the recognized word sequence, the system also produced some other linguistic and prosodic information of the testing utterance, including POS, PM, syllable prosodic state, and syllable-juncture break type. Table VI shows the recognition results of POS. Precision, recall and F-measure were computed as metrics for performance evaluation. Here, precision is defined as the ratio of the number of correctly recognized words with correct POS, $N_{\mathrm{corretW,corretPOS}}$, to the total number of correctly recognized words; while recall is defined as the ratio of $N_{\mathrm{corretW,corretPOS}}$ to the total number of words. As shown in the table, the performances of precision, recall, and F-measure were 93.4%, 76.4%, and 84% for the baseline scheme, and were improved to 93.4%, 80%, and 86.2% by Scheme 2. Since a correct decoding of POS was only meaningful when the word was correctly decoded, the recalls were bounded by the word correct rates which were 78.9% and 82.15% for the baseline scheme and Scheme 2, respectively.

Table VII shows the recognition results of PM. As shown in the table, the performances of precision, recall, and F-measure were 55.2%, 37.8%, and 44.8% for the baseline FLM scheme, and were improved to 61.2%, 53%, and 56.8%, respectively, by Scheme 2. Notice that the syllable-based alignment between the recognition result and the reference transcription was performed for the evaluation. By error analysis, we found that many major PMs were misrecognized as commas. Since this type of error was not serious, we therefore reevaluated the performance of PM recognition by collapsing all PMs (i.e., comma, dot, and major PMs) into a single PM class. The resulting precision, recall, and F-measure were 76.1%, 65.9%, and 70.6% for Scheme 2 verse 66.1%, 45.3%, and 53.8% for the baseline scheme.

Table VIII shows the results of tone recognition. The performances of precision, recall, and F-measure were 87.9%, 87.5%, and 87.7% for the baseline FLM scheme, and were improved to 91.9%, 91.6%, and 91.7% by Scheme 2. Obviously, the significant improvement of tone recognition mainly resulted from the

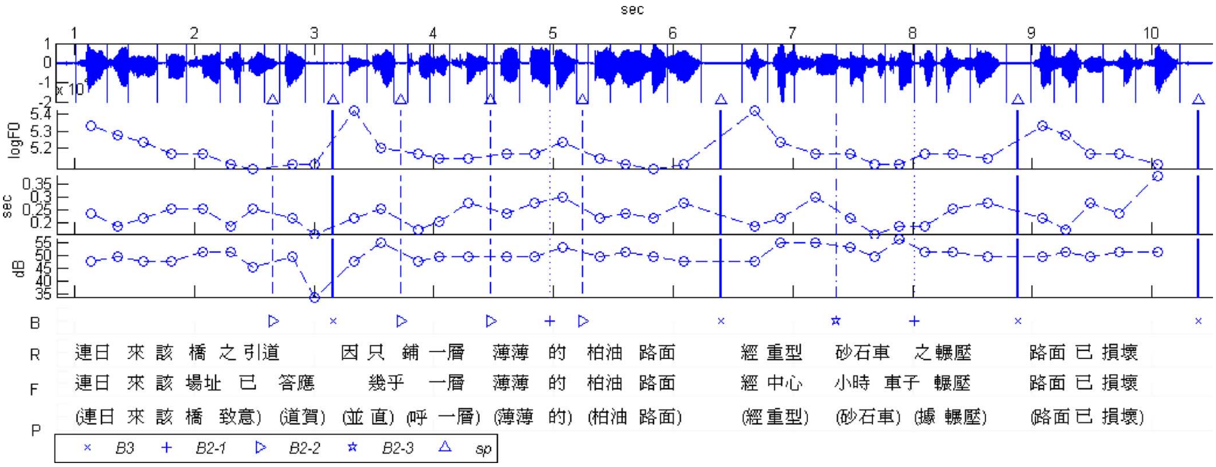


Fig. 14. Example of recognition results for a partial paragraph. Eight panels represent, respectively, waveform, prosodic state AP+global mean of syllable log-F0 level, syllable duration, and syllable energy level, break type (B), reference transcription (R), result of baseline scheme (F) and proposed system (P). The utterance is "lian-ri lai (Day by day) gai-qiao (the bridge) zhi (DE) yin-dao (road), yin (because) zhi (only) pu (pave) yi-ceng (one layer) de (DE) bo-you (asphalt) lu-mian (surface), jing (by) zhong-xing (heavy-duty) sha-sh-che (trunk) zhi (DE) nian-ya (rolling), lu-main (surface) yi (already) sun-huai (broken).

(a) …牙醫師(dentist) 公會(association) 理事長(council chairman) 郭振興(Zhen-Xing Guo)…

(b)…牙醫師(dentist) 公會(association) 理事(council member) 張國政(Guo-Zheng Zhang) 新(new)…

(c) …牙醫師(dentist) $B2$-2 公會(association) $B0$ 理事長(council chairman) $B3$ 或(or) $B2$-2 真心(true heart) $B3$…

Fig. 15. Example of the negative effect of OOV on word error correction: (a) reference transcription, and the recognition results of (b) the baseline scheme and (c) the proposed system.

TABLE VIII
EXPERIMENTAL RESULTS OF TONE DECODING (%)

	Precision	Recall	F-Measure
Baseline scheme	87.9	87.5	87.7
Scheme 2	91.9	91.6	91.7

proper use of tone information in the prosody modeling for syllable pitch contour and syllable duration.

An error analysis was conducted to examine the recognition results in more detail. First, we found that the WER improvement of the proposed system mainly lay in the corrections of word segmentation errors and tone recognition errors. This conformed to our expectation because both syllable-juncture breaks and syllable tones were properly modeled in the prosody modeling. Fig. 14 illustrates an example. As shown in the figure, there were four prosodic phrases (PPh's) separated by $B3$. In the 3rd PPh, the text "經 (jing, by) 重型 (zhong-xing, heavy) 砂石車 (sha-sh-che, trunk) 之 (zhi, DE) 輾壓 (nian-ya, rolling)" were recognized as "經 (jing, by) 中心 (zhong-xin, center) 小時 (xiao-shi, hour) 車子 (che-zi, car) 輾壓 (nian-ya,rolling)" by the baseline scheme. There were three word recognition errors (i.e., 中 心 (zhong-xin), 小 時 (xiao-shi) and 車 子 (che-zi)) and one segmentation error (between 時 "shi" and 車 "che"). The proposed system corrected two word recognition errors. One is the correction of "中心 (zhong-xin)" to "重型 (zhong-xing, heavy)". Tone modeling is the key factor for this correction. Another is the correction of "小時 (xiao-shi) 車子 (che-zi)" to "砂石車 (sha-sh-che)." This word recognition error correction is through the correction of the segmentation error via labeling a $B2-1$ break after the corrected word.

Second, we found that many segmentation error corrections did not lead to word recognition error corrections. The existence of OOV was one of the major factors to hamper the improvement. Fig. 15 illustrates an example. As shown in (b), the two words "理 事 長 (council chairman) 郭 振 興 (Zhen-Xing Guo)" were erroneously recognized as " 理 事 (council member) 張國政 (Guo-Zheng Zhang) 新 (new)" by the baseline scheme. Both words were not correctly recognized and there existed two word segmentation errors. As shown in (c), the proposed system corrected the first word segmentation error and decoded its boundary as a $B3$ break. This led to the correct recognition of the first word, but not the second word because it is an OOV. Moreover, the OOV caused one word substitution error and one word insertion error. Actually, the OOV rate of the test set was only 4.3%, but OOVs caused extra errors of word insertions and deletions to result in total about 8.1% word errors.

Third, we also found that some syllable segmentation errors were corrected by the proposed system. The sum of syllable insertion and deletion error rates was reduced from 1.79% of the baseline FLM scheme to 1.2% of Scheme 2. One major factor to contribute to the improvement was the use of the syllable duration model $P(sd_n|q_n, s_n, t_n)$ shown in (9). Actually, the use of the syllable duration model and break tags in the prosody modeling also contributed to the reduction of the sum of word insertion and deletion error rates from 6.1% of the baseline FLM scheme to 5.5% of Scheme 2.

An additional advantage of the proposed system was the decoding of the two types of prosodic tags. As mentioned before, they were closely correlated with the four-layer prosody-hierarchy model. We could therefore use them to construct a hierarchical structure of prosody for the testing utterance. Taking the recognition results shown in Fig. 14 as an example, we can describe the prosody structure of the utterance as follows. On the top level, there are four prosodic phrases (PPhs) separated by three $B3$ breaks. From the first two panels of Fig. 14, we find that all three $B3$ breaks were associated with long pauses and large pitch resets. So, these three $B3$ breaks were all labeled well. On the next level, there are 2, 5, 3, and 1 prosodic words

TABLE IX
COMPLEXITY OF THE EXPANDED LATTICE FOR RESCORING

	NEL	AEL	DEL	RTF
Baseline scheme	584.6	21650	326.3	2.64
Scheme 2	1192.7	43837	660.8	6.57

(PWs) in these four PPhs, respectively. Within these four PPhs, PWs were separated by $(B2-2)$, $(B2-2, B2-2, B2-1, B2-2)$, $(B2-3, B2-1)$, and $(-)$. As shown in the first three panels of Fig. 14, all four $B2-2$ breaks were associated with short pauses, the $B2-3$ break was associated with a pre-boundary lengthening, and the two $B2-1$ breaks were associated with medium pitch resets. So, they were all properly labeled. Lastly, the bottom level is composed of syllables separated by $B0$ or $B1$ breaks. It is noted that $B0$ and $B1$ are not shown in the figure. From above discussions, we can conclude that the prosody hierarchical structure of the testing utterance constructed by the decoded break tags matched well with the cues provided by the prosodic-acoustic features.

Lastly, we analyzed the complexity of the second-stage rescoring process. Table IX shows the average number of nodes in the expanded lattice (NEL), the average number of arcs in the expanded lattice (AEL), the density of the expanded lattice (DEL), and the real time factor (RTF) of the baseline scheme and the proposed Scheme 2. NEL and AEL are defined as the average numbers of nodes and arcs for a testing utterance. DEL is defined as the number of arcs in the expanded lattice divided by the number of words in the true transcription. RTF is defined as the ratio of the time spent on rescoring to the length of the testing utterance. As shown in Table IX, the proposed system is about 2 times larger in NEL, AEL, and DEL than the baseline scheme; while the RTF is about 2.5 times larger.

IV. CONCLUSION

In this paper, we have discussed a new prosody-assisted ASR system in detail. The system employed a sophisticated prosody modeling method to generate 12 prosodic models to assist in improving the recognition performance as well as decoding more information from the testing utterance. Experimental results confirmed the effectiveness of the proposed system. Several advantages of the proposed system can be found. First, these 12 prosodic models were trained using an unlabeled speech database. This not only saved the costly hand-labeling effort, but also avoided the defects of human labeling, including inaccuracy and inconsistency. The resulting prosodic tag labels matched well with the cues provided by linguistic features and/or prosodic-acoustic features. Second, these 12 prosodic models described well the relationships of the two prosodic tags of the four-layer prosody-hierarchy model, various linguistic features of texts, and the eight prosodic-acoustic features of speech signals. Experimental results showed that parameters of these 12 well-trained prosodic models were all meaningful. Third, the recognition performance of the conventional HMM recognizer can be improved by the proposed system via correcting many word segmentation errors and tone recognition errors. Fourth, more information could be decoded from the testing utterance. Aside from the two linguistic features of POS and PM, the two decoded sequences of break type and prosodic state could be used to construct the prosody hierarchical structure of the testing utterance.

Some further works are worth doing in the future. First, we are interested in generalizing the proposed approach to spontaneous-speech ASR. To this end, we need to extend the three models of AM, LM and HPM to additionally consider the special characteristics, such as disfluency, of spontaneous speech. A preliminary study has been conducted to construct a hierarchical prosodic model for spontaneous Mandarin speech [35]. Second, it is also an interesting task to scale up the proposed approach to ASR for larger vocabulary comprising many compound words. The task can be attacked by modifying the first-stage recognition via first constructing an LM for a lexicon comprising both words and subwords, then generating a mixed-word/subword lattice using the new LM, and lastly forming compound words from subwords by applying some word-compounding rules. The second-stage recognition can be directly applied. Third, modifying the proposed approach to reduce its computational complexity is needed for online system implementation. The task can be attacked by applying some prosodic models to reduce the size of the word lattice generated by the first-stage recognition. Specifically, we can incorporate the syllable-juncture prosodic-acoustic model into the first-stage recognition to detect B3 and B4 from long silences and generate a word lattice for each PPh-like segment instead of a large word lattice for the whole utterance. The stage-stage recognition can then be operated in a way of PPh-by-PPh decoding process. This can greatly speed up the second-stage Viterbi decoding process as well as reduce the decoding delay. Besides, the size of a PPh word lattice can be further reduced by verifying its constituent words using the syllable-juncture prosodic-acoustic model to exclude unqualified words with prosodic features mismatching the intraword prosodic cues. Fourth, it is found from error analysis that the WER improvement of the proposed system is seriously hampered by OOVs. Since most OOVs are name entities, incorporating an LM for name entity should be helpful. Fifth, some high-level linguistic features, such as word chunk, phrase, and syntax, are still not used in this study. Design new prosodic models to include them should be useful for further improving the recognition performance as well as for decoding the syntactic structure of the testing utterance. Lastly, applying the same technique to other languages, such as English, must be interested to the speech processing society.

ACKNOWLEDGMENT

The authors would like to thank the ACLCLP for providing the TCC300 Corpus, Sinorama, NTCIR, and Sinica Corpus.

REFERENCES

[1] S. Ananthakrishnan and S. Narayanan, "Unsupervised adaptation of categorical prosody models for prosody labeling and speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 1, pp. 138–149, Jan. 2009.

[2] S. Ananthakrishnan and S. Narayanan, "Improved speech recognition using acoustic and lexical correlates of pitch accent in a n-best rescoring framework," in *Proc. ICASSP*, 2007, pp. IV-873–IV-873.

[3] S. Ananthakrishnan and S. Narayanan, "Prosody-enriched lattices for improved syllable recognition," in *Proc. Interspeech*, 2007, pp. 1813–1816.

[4] 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.*, vol. 14, no. 1, pp. 232–245, Jan. 2006.

[5] D. H. Milone and A. J. Rubio, "Prosodic and accentual information for automatic speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 11, no. 4, pp. 321–333, Jul. 2003.

[6] D. Vergyri, A. Stolcke, V. R. R. Gadde, L. Ferrer, and E. Shriberg, "Prosodic knowledge sources for automatic speech recognition," in *Proc. ICASSP*, 2003, pp. I-208–I-211.

[7] M. Ostendorf, I. Shafran, and R. Bates, "Prosody models for conversational speech recognition," in *Proc. 2nd Plenary Meeting Symp. Prosody and Speech Process.*, 2003, pp. 147–154.

[8] W.-J. Wang, Y.-F. Liao, and S.-H. Chen, "RNN-based prosodic modeling for mandarin speech and its application to speech-to-text conversion," *Speech Commun.*, vol. 36, pp. 247–265, 2002.

[9] J.-T. Huang and L.-S. Lee, "Improved large vocabulary Mandarin speech recognition using prosodic features," in *Proc. Speech Prosody*, 2006.

[10] J.-T. Huang and L.-S. Lee, "Prosodic modeling in large vocabulary Mandarin speech recognition," in *Proc. ICSLP*, 2006.

[11] X. Lei and M. Ostendorf, "Word-level tone modeling for Mandarin speech recognition," in *Proc. ICASSP*, 2007, pp. IV-665–IV-668.

[12] C. Ni, W. Liu, and B. Xu, "Improved large vocabulary Mandarin speech recognition using prosodic and lexical information in maximum entropy framework," in *Proc. CCPR*, 2009.

[13] C. Ni, W. Liu, and B. Xu, "Using prosody to improve Mandarin automatic speech recognition," in *Proc. Interspeech*, 2010.

[14] Y. Liu, E. Shriberg, A. 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.*, vol. 14, no. 5, pp. 1526–1540, Sep. 2006.

[15] E. Shriberg and A. Stolcke, "Prosody modeling for automatic speech recognition and understanding," in *Proc. Workshop Math. Found. Natural Lang. Modeling*, 2002.

[16] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling English prosody," in *Proc. ICSLP*, 1992, vol. 2, pp. 867–870.

[17] D. Hirst and A. D. Cristo, *Intonation Systems. A Survey of Twenty Languages*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[18] V. K. R. Sridhar, S. Bangalore, and S. S. Narayanan, "Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 797–811, May 2008.

[19] J.-H. Jeon and Y. Liu, "Automatic prosodic events detection suing syllable-based acoustic and syntactic features," in *Proc. ICASSP*, 2009, pp. 4565–4568.

[20] C.-Y. Chiang, S.-H. Chen, H.-M. Yu, and Y.-R. Wang, "Unsupervised joint prosody labeling and modeling for Mandarin speech," *J. Acoust. Soc. Amer.*, vol. 125, no. 2, pp. 1164–1183, Feb. 2009.

[21] C.-Y. Tseng, S.-H. Pin, Y.-L. Lee, H.-M. Wang, and Y.-C. Chen, "Fluent speech prosody: Framework and modeling," *Speech Commun.*, vol. 46, pp. 284–309, 2005.

[22] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Tree*. Belmont, CA: Wadsworth, 1984.

[23] S.-H. Chen and Y.-R. Wang, "Vector quantization of pitch information in Mandarin speech," *IEEE Trans. Commun.*, vol. 38, no. 9, pp. 1317–1320, Sep. 1990.

[24] J. A. Bilmes and K. Kirchhoff, "Factor language models and generalized parallel backoff," in *Proc. HLT/NACCL*, 2003, pp. 4–6.

[25] A. Stolcke, "SRILM – An extensible language modeling toolkit," in *Proc. ICSLP*, 2002.

[26] P. Beyerlein, "Discriminative model combination," in *Proc. ICASSP*, 1998, pp. 481–484.

[27] B.-H. Juang, W. Chou, and C.-H. Lee, "Statistical and discriminative methods for speech recognition," in *Speech Recognition and Coding – New Advances and Trend*, A.J. Rubio Ayuso and J. M. Lopez Soler, Eds. Berlin-Hheidelberg, Germany: Springer-Verlag, 1995.

[28] "Mandarin Microphone Speech Corpus—TCC300," [Online]. Available: http://www.aclclp.org.tw/use_mat.php#tcc300edu

[29] HTK Web-Site. 2009 [Online]. Available: http://htk.eng.cam.ac.uk

[30] L. R. Bahl, R. F. Brown, P. V. de Souza, and R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proc. ICASSP*, 1986, pp. 49–52.

[31] C.-R. Huang, K.-J. Chen, F.-Y. Chen, Z.-M. Gao, and K.-Y. Chen, "Sinica treebank: Design criteria, annotation guidelines, and on-line interface," in *Proc. 2nd Chinese Lang. Process. Workshop '00*, Hong Kong, 2000, pp. 29–37.

[32] S.-H. Chen, W.-H. Lai, and Y.-R. Wang, "A statistics-based pitch contour model for Mandarin speech," *J. Acoust. Soc. Amer.*, vol. 117, no. 2, pp. 908–925, Feb. 2005.

[33] S.-H. Chen, W.-H. Lai, and Y.-R. Wang, "A new duration modeling approach for Mandarin speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 11, no. 4, pp. 308–320, Jul. 2003.

[34] Y. Xu, "Contextual tonal variations in Mandarin," *J. Phonetics*, vol. 25, pp. 61–83, 2007.

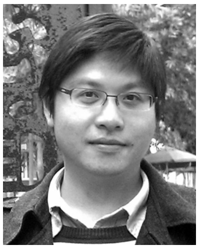
[35] Y.-L. Chou, C.-Y. Chiang, Y.-R. Wang, H.-M. Yu, and S.-H. Chen, "Prosody labeling and modeling for Mandarin spontaneous speech," in *Proc. Speech Prosody '10*, Chicago, IL, May 2010.

Sin-Horng Chen (SM'94) received the B.S. degree in communication engineering and the M.S. degree in electronics engineering from National Chiao Tung University (NCTU), Hsinchu, Taiwan, in 1976 and 1978, respectively, and the Ph.D. degree in electrical engineering from Texas Tech University, Lubbock, in 1983.

He became an Associate Professor and a Professor in the Department of Communications Engineering, NCTU, in 1983 and 1990, respectively. Currently, he is the Dean of the Electrical and Computer Engineering College, NCTU. His major research interest is in speech signal processing, especially in Mandarin speech recognition and text-to-speech.

Jyh-Her Yang received the B.S. degree in electrical engineering from National Yunlin University of Science and Technology, Yunlin, Taiwan, in 2002, and the M.S. degree in computer and communication engineering from National Taipei University of Technology, Taipei, Taiwan, in 2004. He is currently pursuing the Ph.D. degree in communications engineering from National Chiao Tung University, Hsinchu, Taiwan.

He is an Associate Researcher at Telecommunication Labs, Chunghwa Telecom Co., Taiwan. His major research area is large-vocabulary Mandarin speech recognition.



Chen-Yu Chiang (M'09) was born in Taipei, Taiwan, in 1980. He received the B.S., M.S., Ph.D. degrees in communication engineering from National Chiao Tung University (NCTU), Hsinchu, Taiwan, in 2002, 2004, and 2009, respectively.

In 2009, he was a Postdoctoral Fellow at the Department of Electrical Engineering, NCTU, where he primarily worked on prosody modeling for automatic speech recognition and text-to-speech system, under the guidance of Prof. Sin-Horng Chen. He is now a Visiting Scholar at the Center for Signal and Image Processing (CSIP), Georgia Institute of Technology, Atlanta. His main research interests are in speech processing, in particular prosody modeling, automatic speech recognition and text-to-speech systems.

Dr. Chen-Yu Chiang is a member of ISCA and ASA.I

Ming-Chieh Liu received the B.S. degree in electrical engineering from Chung Yuan Christian University, Chung Li City, Taiwan, in 2009, and the M.S. degree in communications engineering from National Chiao Tung University, Hsinchu, Taiwan, in 2011.

He is an Engineer with the Taiwan Semiconductor Manufacturing Company, Hsinchu, Taiwan. His major research area is 3-D IC techniques His research interest is prosody-assisted speech recognition.

Yih-Ru Wang (M'06) received the B.S. and M.S. degrees from the Department of Communication Engineering, National Chaio Tung University (NCTU), Hsinchu, Taiwan, in 1982 and 1987, respectively, and the Ph.D. degree from the Institute of Electronic Engineering, NCTU, in 1995.

He was an Instructor in the Department of Communication Engineering, NCTU, from 1987 to 1995. In 1995, he became an Associate Professor. His general research interests are Mandarin spontaneous speech recognition and the application of neural networks in speech processing.