國 立 交 通 大 學

電信工程研究所

博 士 論 文

一種韻律輔助中文語音辨認系統及其應用

A New Prosody-Assisted Mandarin ASR System and Its Application

研 究 生： 楊智合

指導教授： 陳信宏 博士
廖元甫 博士

中華民國 一百零一 年 六 月

一種韻律輔助中文語音辨認系統及其應用

A New Prosody-Assisted Mandarin ASR System and Its Application

研 究 生 ：楊智合

指導教授 ：陳信宏
廖元甫

Student : Jyh-Her Yang

Advisors : Sin-Horng Chen
Yuan-Fu Liao

國立交通大學
電信工程研究所
博士論文

A Dissertation
Submitted to Institute of Communication Engineering
College of Electrical and Computer Engineering
National Chiao Tung University
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
in

Communication Engineering

June 2012

Hsinchu, Taiwan, Republic of China

中華民國一百零一年六月

推　薦　函

中華民國一零一年五月十一日

一、 事由：本校電信研究所博士班研究生 楊智合 提出論文以參加國立交通大學博士班論文口試。

二、 說明：本校電信研究所博士班研究生 楊智合 已完成本校電信研究所規定之學科課程及論文研究之訓練。

有關學科部分，楊君已修滿十八學分之規定（請查閱學籍資料）並通過資格考試。

有關論文部分，楊君已完成其論文初稿，相關之論文亦分別發表或即將發表於國際期刊（請查閱附件）並滿足論文計點之要求。

總而言之，楊君已具備國立交通大學電信工程研究所應有之教育及訓練水準，因此特推薦

楊君參加國立交通大學電信工程研究所博士班論文口試。

交通大學電信工程研究所教授　　陳信宏

台北科技大學電腦與通訊研究所副教授　　廖元甫

一種韻律輔助中文語音辨認系統及其應用

研究生：楊智合　　指導教授：陳信宏　博士
廖元甫　博士

國立交通大學電信工程學系

中文摘要

本論文提出一種新的韻律輔助之中文語音辨識系統，它不同於以往較簡單的作法，是利用較精細的四層中文韻律結構模式來幫助中文語音辨認，本論文利用先前已開發的韻律標記與韻律模式演算法從大量未經人工標記的語料庫中自動產生訓練出 12 種韻律模型，並以兩階段方式將其加入到自動語音辨認系統中，對系統中第一個階段，也就是傳統隱藏式馬可夫模型(HMM)辨認器所產生的詞圖(word lattice)作重新評分的動作，如此可以得到更正確的詞辨認序列；此外，系統第二個階段同時解碼出多種資訊，包含詞性(POS)、標點符號(PM)以及用來建構測試語料之階層式韻律架構的兩種韻律標記。本論文實驗語料使用 TCC300 語料庫中的朗讀式長句，同時實驗中引入一個因子式語言模型，它是一個描繪詞、詞性及標點符號三者之間關係的模型，以此當作基準(baseline)辨認效能。本研究在加入所有韻律資訊後之實驗結果對於詞(word)、字(character)、音節(syllable)的錯誤率分別為 20.1%、13.6%及 9.4%，與 baseline 比較則分別改善了 4.1%、4.0%及 2.4%的絕對錯誤率(16.9%、22.6%及 20.6%的相對錯誤率)。由實驗結果分析，發現本系統能成功修正許多辨認錯誤是來自於搶詞與聲調錯誤。

在應用上，我們使用此辨認方法建立一種新的以模式為基礎的中文語音韻律編碼系統，在編碼端，以此韻律輔助語音辨認系統由輸入語音產生語言參數及韻律標記加以編碼；在解碼端，將這些語言參數及韻律標記資訊解碼，用以建構出音節基頻軌跡、音節長度、音節能量位準及音節間的停頓長度，接著以 HMM 語音合成器結合語音的頻譜參數合成出語音訊號，由 TCC300 語料之實驗證實，合成語音在低資料率 543 bits/sec 下仍有高的聲音品質。

A New Prosody-Assisted Mandarin ASR System and Its Application

Student: Jyh-Her Yang

Advisors: Dr. Sin-Horng Chen
Dr. Yuan-Fu Liao

Department of Communication Engineering, National Chiao Tung University
Hsinchu, Taiwan, Republic of China

Abstract

This dissertation 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 4-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.

With the success of the prosody-assisted ASR system, we conduct an application to speech coding. A new model-based Mandarin-speech coding system is proposed. It employs the prosody-assisted ASR with the hierarchical prosodic model (HPM) to generate from the input speech enriched transcriptions, including linguistic features, prosodic tags and spectral parameters in the encoder. By sending these features to the decoder, we can first reconstruct the prosodic-acoustic features of syllable pitch contour, syllable duration, syllable energy level, and inter-syllable pause duration by HPM using the linguistic features and prosodic tags; and then combined with spectral parameters to reconstruct the input speech signal by an HMM-based speech synthesizer. Experimental results show that the reconstructed speech has good quality at a low data rate of 543 bits/s.

致謝

最需要感謝的是我的指導教授：陳信宏老師及廖元甫老師，在我求學的過程中，他們給予專業上豐富的知識並讓我有許多參與國際學術會議的機會，提升我的研究視野，也感謝兩位老師平日生活上的諄諄教誨與不吝分享，當感到挫折時能被鼓舞而打起精神繼續在研究上打拼，從學術及心態等各方面都獲益良多。此外，也感謝王逸如老師的指教，平時在請教王老師一些問題時都可以得到很好的回應與幫助。也感謝冀泰石老師，在與冀老師合作的期間，學習到許多。最後，非常感謝王小川老師、王駿發老師、李琳山老師、張文輝老師四位口試委員對本研究的肯定和建議，對我而言是種莫大的鼓勵。

在實驗室裡，感謝羅文輝學長，常給一些建議提供我另一個思考方向；感謝郭威志學長在我剛進實驗室時的提攜與指引；也感謝振宇、阿德、希群和巴金，在研究上相互鼓勵與切磋；也感謝工研院的林政賢學長讓我有機會到工研院去打工學習；還要謝謝歷年來一起在實驗室同甘共苦的可愛學弟妹們，要感謝的人多到無法列舉，要感謝的話多到無法言盡，總而言之，在我生命中出現的人都是我要感謝的人，都是我的貴人，感謝有了你們讓我的生活更加豐富，才能拼湊出我不一樣的人生。

最後，特別感謝一直支持我的家人，謝謝爸媽包容我那麼多年沒有工作，讓我不用擔心負擔家中生計而令我能心無旁騖地把心思放在研究，沒有你們的支持就無法有今日的我；還有我的愛人淳郁，認識妳是我三生有幸，雖然我們之間經過好幾番波折，但最終妳總能不離不棄，多年下來，我們之間的感情變得更緊密堅定，也是妳讓我變得成熟許多，感謝有妳一路陪伴，分享我的喜怒哀樂。你們的支持及鼓勵是我生命中最大的力量，在此僅將此論文獻給你們！

Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Background

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 firstly 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.

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 pre-determined by an 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 level 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. Figure 1.1 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 lies in the need of a large well-annotated database with full prosodic cues being properly labeled. In the past, prosody labeling is usually done by human became of the lack of a good automatic labeling algorithm. But, preparing such a database by human is still difficult because the labeling work is highly time-consuming and it is not easy to maintain the consistency of fully labeling of all prosodic cues for the same annotators or between different annotators. So, most previous works of this class used databases annotated with only few obvious prosodic cues, such as pitch accent and intonational phrase boundary. This will highly limit the effectiveness 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, the prosodic cues they used were still very limited. Besides, 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 1.1 summarizes the primary features of prosody modeling and experiment setting for those previous studies on prosody-assisted ASR for comparison.

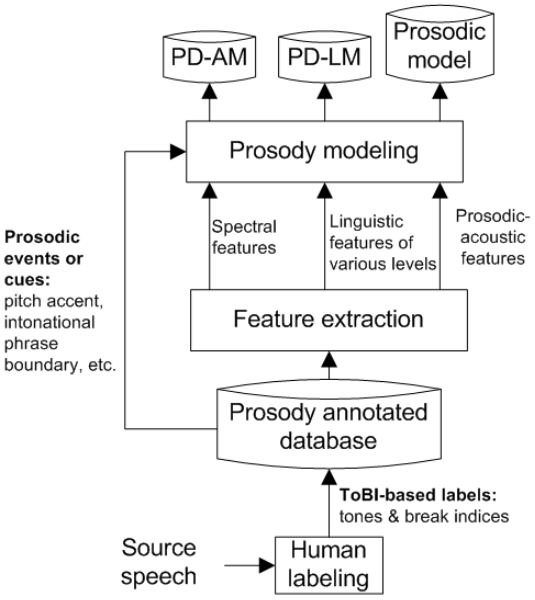


Figure 1.1: A conceptual block diagram of the prosody modeling class using intermediate abstract phonological categories. PD-AM and PD-LM denote prosody-dependent acoustic model and prosody-dependent language model.

Table 1.1: 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*/p	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*/p	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*/p	t/WB	M	B	100K	SD	1.45/7.5(character)
Ni [12]	Dir	-	-	F0/d/e/p	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*/p	SJ	M	R	110K	SD	1.1/4.2(character)
proposed	7B+PS	4-L	U	F0*/d*/e/p/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 | Dir: direct prosody modeling}; **PH: prosody hierarchy** = {L: layer}; **PL: prosody labeling** = {U: unsupervised | SS: semi-supervised | S: supervised | BS: bootstrapping | R: taking lexical word as potential PW}; **PAF: prosodic-acoustic feature** = {F0: fundamental frequency | d: duration | e: energy | pd: pause duration | *: with differential}; **LF: linguistic feature** = {t: tone | ph: phone | s: base-syllable type | W: word | 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 | MS: multi speaker}; **IMP: improvement in absolute/relative accuracy.**

1.2 Motivation

In this dissertation, 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 Figure 1.1 mainly on adopting a systematic way to perform prosody modeling on a large unlabeled database for automatically exploiting full prosodic cues of speech based on a 4-layer prosody-hierarchy model to assist in ASR. The general goal of our prosody modeling is to explore a wide-range, mixed context information of speech and the associated text via building prosodic models to properly describe the relations of the parameters of the 4-layer prosody-hierarchy model with the prosodic-acoustic features, provided by the input speech, and the linguistic features of the target text to be recognized. Figure 1.2 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]. The 4-layer model of prosody hierarchy of Mandarin speech defines two types of prosodic tags, break type and prosodic state, to specify its 4-layer structure and 4 types of constituents. 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 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 enriching information has also contribution on an application of the proposed prosody-assisted ASR system, speech coding system, in the post-processing. It differs from the conventional speech coding system on using the prosody-assisted ASR in the encoder to extract high-level linguistic and prosodic features to assist in improving the coding efficiency.

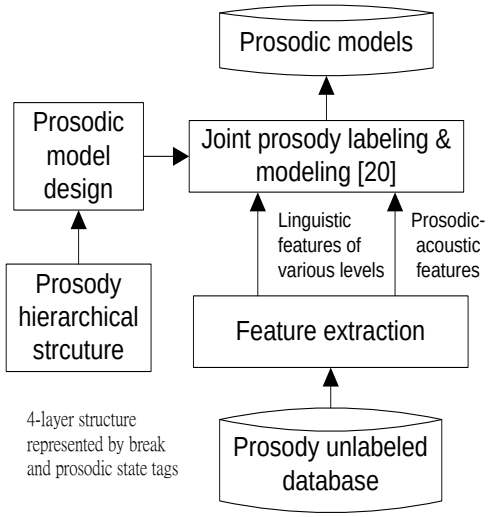


Figure 1.2: The prosody modeling approach in the proposed prosody-assisted ASR system.

1.3 Organization of the Dissertation

The rest of this dissertation is organized as follows. Chapter 2 presents the proposed prosody-assisted ASR system. It introduces the design of the hierarchical prosody model (HPM), the training of HPM, the two-stage prosody-assisted ASR system, and experimental results. In Chapter 3, we introduce an application of the proposed prosody-assisted ASR system to the coding of prosodic information for Mandarin speech. Some conclusions and future works are given in the last chapter.

Chapter 2 The Proposed Prosody-Assisted ASR System

The proposed prosody-assisted Mandarin speech recognition system is discussed in detail in this chapter. The chapter is organized as follows. Section 2.1 presents the design of prosodic models for ASR. The training of the proposed prosodic models is discussed in Section 2.2. Section 2.3 describes the two-stage prosody-assisted ASR system. Section 2.4 discusses the experimental results. Some conclusions of this chapter are given in Section 2.5.

2.1 The 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 *B*0, *B*1, *B*2, *B*3, *B*4 and *B*5 [21]. First, *B*0 and *B*1 represent respectively non-breaks of reduced syllable boundary (or tightly-coupling syllable juncture) and normal syllable boundary, within a PW, which have no identifiable pauses between SYLs. Second, PW boundary *B*2 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. Thirdly, *B*4 and *B*5 are defined for BG and PG boundaries, respectively. *B*4 is a breathing pause and *B*5 is a

complete speech paragraph end characterized by final lengthening coupled with weakening of speech sounds.

In this dissertation, we adopt a 4-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 4-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 represent 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 *B*2 into three types, *B*2-1, *B*2-2 and *B*2-3, and combining *B*4 and *B*5 into one denoted simply by *B*4. Here, *B*2-1, *B*2-2 and *B*2-3 represent PW boundaries with F0 reset, short pause and pre-boundary syllable duration lengthening, respectively. The reason of refining *B*2 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 *B*4 and *B*5 is owing to the similarity of their prosodic-acoustic characteristics. Therefore, the break-type tag set used is {*B*0, *B*1, *B*2-1, *B*2-2, *B*2-3, *B*3, *B*4}. As shown in Figure 2.1, 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.

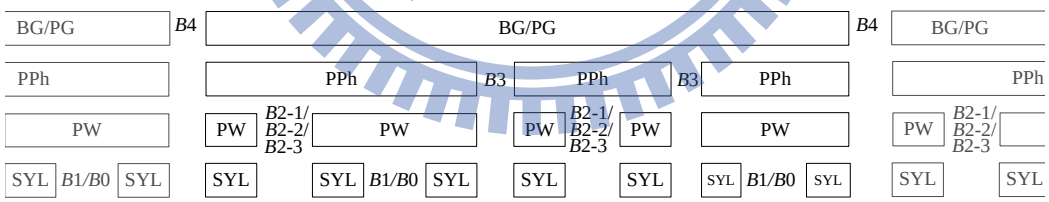


Figure 2.1: The prosody-hierarchy model of Mandarin speech used in this study [20], [21].

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 $\Lambda_a = \{\mathbf{X}_a, \mathbf{X}_p\}$:

$$\begin{aligned} \Lambda_l^*, \Lambda_p^*, \Upsilon_s^* &= \arg \max_{\Lambda_l, \Lambda_p, \Upsilon_s} P(\mathbf{W}, \mathbf{POS}, \mathbf{PM}, \mathbf{B}, \mathbf{P}, \Upsilon_s \mid \mathbf{X}_a, \mathbf{X}_p) \\ &= \arg \max_{\Lambda_l, \Lambda_p, \Upsilon_s} P(\mathbf{W}, \mathbf{POS}, \mathbf{PM}, \mathbf{B}, \mathbf{P}, \Upsilon_s, \mathbf{X}_a, \mathbf{X}_p) \end{aligned} \tag{2.1}$$

where $\mathbf{W} = \{w_1^M\}$ is a word sequence; $\mathbf{POS} = \{pos_1^M\}$ is a POS sequence associated with 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 $B_n \in \{B0, B1, B2\text{-}1, B2\text{-}2, B2\text{-}3, B3, B4\}$; N is the total number of syllables; $\mathbf{P} = \{\mathbf{p}, \mathbf{q}, \mathbf{r}\}$ 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; $\mathbf{X}_a$ is a frame-based spectral feature vector 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 **X**, **Y**, and **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 X; syllable-juncture pause duration (pd) and energy-dip level (ed) for Y; and normalized pitch-level jump (pj) and two normalized duration lengthening factors (dl and df) for Z. Notations of tags and features are summarized in Table 2.1.

Table 2.1: 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 durations **se**: syllable energy levels
	Y: syllable-juncture prosodic-acoustic features	**pd**: pause durations **ed**: energy-dip levels
	Z: inter-syllable differential prosodic-acoustic features	**pj**: normalized pitch-level jumps **dl**: normalized duration lengthening factor 1 **df**: normalized duration lengthening factor 2
Λ_l : linguistic features	**W**: words **POS**: part-of-speeches **PM**: punctuation marks	
	t: tones **s**: base-syllable types **f**: final types	

To make Equation (2.1) 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 **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 **X** is independent of syllable-juncture and inter-syllable differential prosodic-acoustic feature sequences, **Y** and **Z**; 4) Break tag sequence **B** depends mainly on contextual linguistic feature sequence Λ_l; and 5) Prosodic state sequence **P** depends on **B** only. The reason is that **P** is used to characterize the prosodic constituents' patterns which are mainly determined by the prosody hierarchy specified by the break type sequence **B**. The relation between linguistic features and prosody hierarchy is built through the modeling of **B**. In other words, the linguistic feature Λ_l can influence the prosodic state through **B**. We therefore ignore the direct dependency of **P** on Λ_l for simplicity. Based on these assumptions, Equation (2.1) is rewritten as

$$\Lambda_l^*, \Lambda_p^*, \Upsilon_s^* \approx \arg \max_{\Lambda_l, \Lambda_p, \Upsilon_s} \{P(\mathbf{X}_a, \Upsilon_s|\mathbf{W})P(\mathbf{W}, \mathbf{POS}, \mathbf{PM}) \cdot P(\mathbf{B}|\Lambda_l)P(\mathbf{P}|\mathbf{B})P(\mathbf{X}|\Upsilon_s, \Lambda_p, \Lambda_l)P(\mathbf{Y}, \mathbf{Z}|\Upsilon_s, \Lambda_p, \Lambda_l)\} \quad (2.2)$$

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 **W**, **POS** and **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 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. Figure 2.2 shows the relationships of features involved in the four prosodic models, LM, and AM.

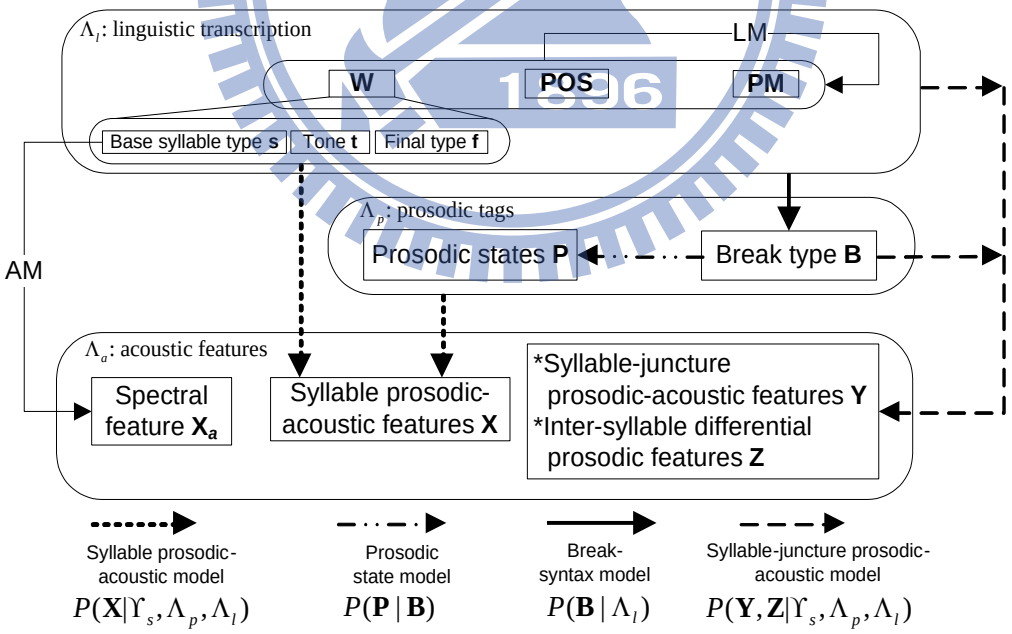


Figure 2.2: The relationships of AM, LM, and four prosodic models with prosodic tags, linguistic features and prosodic-acoustic features.

In implementation, we need to further elaborate these four prosodic models.

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

$$P(\mathbf{B}\,|\,\Lambda_l) \approx \prod_{n=1}^{N-1} P(B_n\,|\,\Lambda_{l,n}) \tag{2.3}$$

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.

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

$$\begin{aligned} P(\mathbf{P}\,|\,\mathbf{B}) &\approx P(\mathbf{p}|\mathbf{B})P(\mathbf{q}|\mathbf{B})P(\mathbf{r}|\mathbf{B}) \\ &\approx P(p_1)P(q_1)P(r_1)\left[\prod_{n=2}^{N} P(p_n\,|\,p_{n-1},B_{n-1})P(q_n\,|\,q_{n-1},B_{n-1})P(r_n\,|\,r_{n-1},B_{n-1})\right] \end{aligned} \tag{2.4}$$

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 these 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.

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

$$\begin{aligned} P(\mathbf{X}|\Upsilon_s,\Lambda_p,\Lambda_l) &\approx P(\mathbf{sp}|\Upsilon_s,\mathbf{B},\mathbf{p},\mathbf{t})P(\mathbf{sd}|\Upsilon_s,\mathbf{B},\mathbf{q},\mathbf{t},\mathbf{s})P(\mathbf{se}|\Upsilon_s,\mathbf{B},\mathbf{r},\mathbf{t},\mathbf{f}) \\ &\approx \prod_{n=1}^{N} P(sp_n|p_n,B_{n-1}^{n},t_{n-1}^{n+1})P(sd_n\,|\,q_n,s_n,t_n)P(se_n\,|\,r_n,\,f_n,t_n) \end{aligned} \tag{2.5}$$

where $P(sp_n \mid B_{n-1}^n, p_n, t_{n-1}^{n+1})$, $P(sd_n \mid q_n, s_n, t_n)$, and $P(se_n \mid r_n, f_n, t_n)$ 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

$$sp_n = sp_n^r + \beta_{t_n} + \beta_{p_n} + \beta^f_{B_{n-1},t^n_{n-1}} + \beta^b_{B_n,t_n^{n+1}} + \mu_{sp} \tag{2.6}$$

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; $\beta^f_{B_{n-1},t^n_{n-1}}$ 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

$$P(sp_n \mid p_n, B_{n-1}^n, t_{n-1}^{n+1}) = N(sp_n; \beta_{t_n} + \beta_{p_n} + \beta^f_{B_{n-1},t^n_{n-1}} + \beta^b_{B_n,t_n^{n+1}} + \mu_{sp}, R_{sp}) \tag{2.7}$$

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 $P(sd_n \mid q_n, s_n, t_n)$ and the syllable energy level model $P(se_n \mid r_n, f_n, t_n)$ are formulated by

$$P(sd_n \mid q_n, s_n, t_n) = N(sd_n; \gamma_{t_n} + \gamma_{s_n} + \gamma_{q_n} + \mu_{sd}, R_{sd}). \tag{2.8}$$

$$P(se_n \mid r_n, f_n, t_n) = N(se_n; \omega_{t_n} + \omega_{f_n} + \omega_{r_n} + \mu_{se}, R_{se}). \tag{2.9}$$

where sd_n and se_n are the observed duration and energy level of syllable *n*, respectively; γ's and ω's represent APs for syllable duration and syllable energy

level; μ_{sd} and μ_{se} are their global means; and R_{sd} and R_{se} are variances of modeling residues.

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

$$\begin{aligned} P(\mathbf{Y},\mathbf{Z}|\Upsilon_s,\Lambda_p,\Lambda_l) &\approx P(\mathbf{pd},\mathbf{ed},\mathbf{pj},\mathbf{dl},\mathbf{df}|\Upsilon_s,\Lambda_p,\Lambda_l) \\ &\approx \prod_{n=1}^{N-1} P(pd_n,ed_n,pj_n,dl_n,df_n|\Upsilon_s,B_n,\Lambda_{l,n}) \\ &\approx \prod_{n=1}^{N-1}\Big\{ g(pd_n;\alpha_{B_n,\Lambda_{l,n}},\eta_{B_n,\Lambda_{l,n}}) N(ed_n;\mu_{ed,B_n,\Lambda_{l,n}},\sigma^2_{ed,B_n,\Lambda_{l,n}}) \\ &\quad \cdot N(pj_n;\mu_{pj,B_n,\Lambda_{l,n}},\sigma^2_{pj,B_n,\Lambda_{l,n}}) N(dl_n;\mu_{dl,B_n,\Lambda_{l,n}},\sigma^2_{dl,B_n,\Lambda_{l,n}}) \\ &\quad \cdot N(df_n;\mu_{df,B_n,\Lambda_{l,n}},\sigma^2_{df,B_n,\Lambda_{l,n}})\Big\} \end{aligned} \tag{2.10}$$

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;

$$pj_n = (sp_{n+1}(1) - \beta_{t_{n+1}}(1)) - (sp_n(1) - \beta_{t_n}(1)) \tag{2.11}$$

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;

$$dl_n = (sd_n - \gamma_{t_n} - \gamma_{s_n}) - (sd_{n-1} - \gamma_{t_{n-1}} - \gamma_{s_{n-1}}) \tag{2.12}$$

$$df_n = (sd_n - \gamma_{t_n} - \gamma_{s_n}) - (sd_{n+1} - \gamma_{t_{n+1}} - \gamma_{s_{n+1}}) \tag{2.13}$$

are two normalized duration lengthening factors before and across juncture n. Both dl_n and df_n are modeled as normal distributions. 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 *B*0 and *B*1. Like the state duration of phone HMM model, Gamma distribution is suitable for this kind of data.

2.2 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

$$
\begin{aligned}
Q = & \left(\prod_{n=1}^{N-1} P(B_n \mid \Lambda_{l,n}) \right) \left(P(p_1)P(q_1)P(r_1) \left[\prod_{n=2}^{N} P(p_n \mid p_{n-1}, B_{n-1})P(q_n \mid q_{n-1}, B_{n-1})P(r_n \mid r_{n-1}, B_{n-1}) \right] \right) \\
& \left(\prod_{n=1}^{N} P(sp_n \mid p_n, B_{n-1}^{n}, t_{n-1}^{n+1})P(sd_n \mid q_n, s_n, t_n)P(se_n \mid r_n, f_n, t_n) \right) \\
& \left(\prod_{n=1}^{N-1} g(pd_n; \alpha_{B_n,\Lambda_{l,n}}, \eta_{B_n,\Lambda_{l,n}}) N(ed_n; \mu_{ed,B_n,\Lambda_{l,n}}, \sigma^2_{ed,B_n,\Lambda_{l,n}}) N(pj_n; \mu_{pj,B_n,\Lambda_{l,n}}, \sigma^2_{pj,B_n,\Lambda_{l,n}}) \right. \\
& \left. N(dl_n; \mu_{dl,B_n,\Lambda_{l,n}}, \sigma^2_{dl,B_n,\Lambda_{l,n}}) N(df_n; \mu_{df,B_n,\Lambda_{l,n}}, \sigma^2_{df,B_n,\Lambda_{l,n}}) \right)
\end{aligned}
\tag{2.14}
$$

It then performs a multi-step iterative procedure to re-label 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.

2.3.1 Initialization

(a) Initial labeling of break indices

The initial break index of each syllable juncture is determined by a decision tree

shown in Figure 2.3. 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 *B*4 if its pause duration is longer than a large threshold *Th*1. Then, it is assigned as *B*3 if its pause duration is longer than *Th*2. Then, all intrawords are labeled as *B*0/*B*1. We then mark interwords with medium pause duration ($\geq Th3$) as *B*2-2, with medium pitch jump ($\geq Th4$) as *B*2-1, and with medium pre- or post-syllable lengthening ($\geq Th5$ and $\geq Th6$) as *B*2-3. All remaining interwords are labeled as *B*0/*B*1. Lastly, *B*0/*B*1 are refined as *B*0 if the syllable juncture has continuous F0 trajectory, otherwise it is labeled as *B*1. 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.

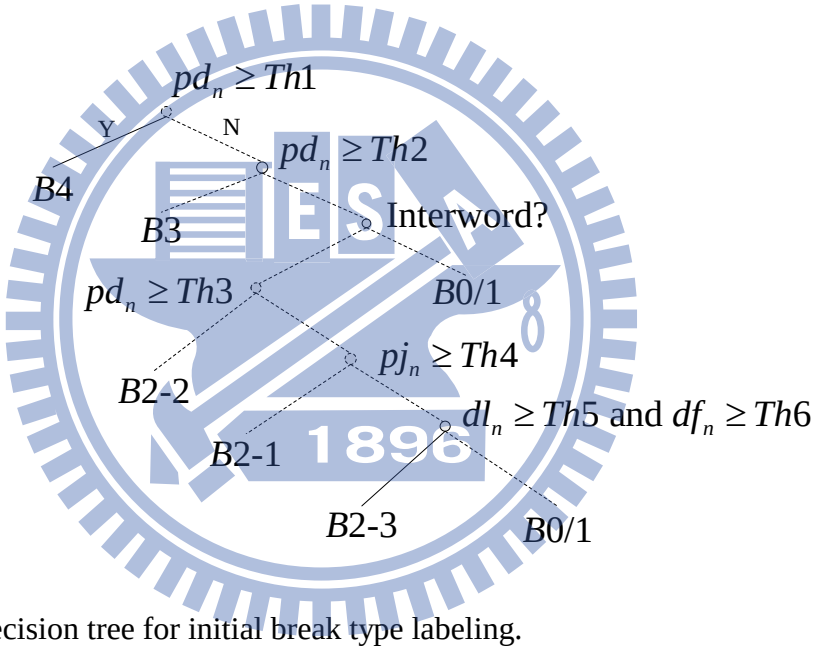


Figure 2.3: The decision tree for initial break type labeling.

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 *B*3 or *B*4. 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 *B*4 and *B*3, i.e. $f_{B3}(pd)$ and $f_{B4}(pd)$, respectively. The threshold *Th*1 is then set to be the equal probability intersection between the two distributions. Then, we construct a Gamma distribution $f_{B0/1}(pd)$ for *B*0/*B*1 by using the pause durations of all intrawords. Another Gamma distribution $f_{B2\text{-}2}(pd)$ for *B*2-2 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 *B*0/*B*1. The thresholds *Th*2 and *Th*3 are then set to be the equal probability intersections of $f_{B2\text{-}2}(pd)/f_{B3}(pd)$ and $f_{B2\text{-}2}(pd)/f_{B0/1}(pd)$.

We then determine the three thresholds, *Th*4, *Th*5, and *Th*6, which are used to label initial *B*2-1 and *B*2-3. 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 *B*2-1, i.e., $f_{B2\text{-}1}(pj)$, 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 *B*2-3, i.e., $f_{B2\text{-}3}(dl)$ and $f_{B2\text{-}3}(df)$, 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, *Th*4, *Th*5 and *Th*6 are set to be the equal probability intersections of $f_{\text{intra}}(pj)/f_{B2\text{-}1}(pj)$, $f_{\text{intra}}(dl)/f_{B2\text{-}3}(dl)$ and $f_{\text{intra}}(df)/f_{B2\text{-}3}(df)$.

(b) 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 of initial 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.

2.3.2 Iteration

The iteration is a multi-step procedure listed below:

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: Re-label the prosodic state sequence of each utterance by the Viterbi algorithm so as to maximize Q defined in Equation (2.14).

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: Re-label the break type sequence of each utterance by the Viterbi algorithm so as to maximize Q defined in Equation (2.14).

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.

2.3 The Two-Stage Prosody-Assisted ASR System

Figure 2.4 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 W, the part-of-speech sequence POS, and the punctuation mark sequence PM. The FLM is composed of a word-trigram model, a factored POS model and a factored PM model, and is formulated as

$$P(\mathbf{W},\mathbf{PM},\mathbf{POS}) \approx \prod_{i=1}^{M} \left\{ \underbrace{P(w_i \mid w_{i-2}^{i-1})}_{\text{word-trigram LM}} \cdot \underbrace{P(pos_i \mid pos_{i-1}, w_i)}_{\text{factored POS model}} \cdot \underbrace{P(pm_{i-1} \mid pos_{i-1}^{i}, w_{i-1})}_{\text{factored PM model}} \right\} \quad (2.16)$$

Here, the FLM approach used in [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.

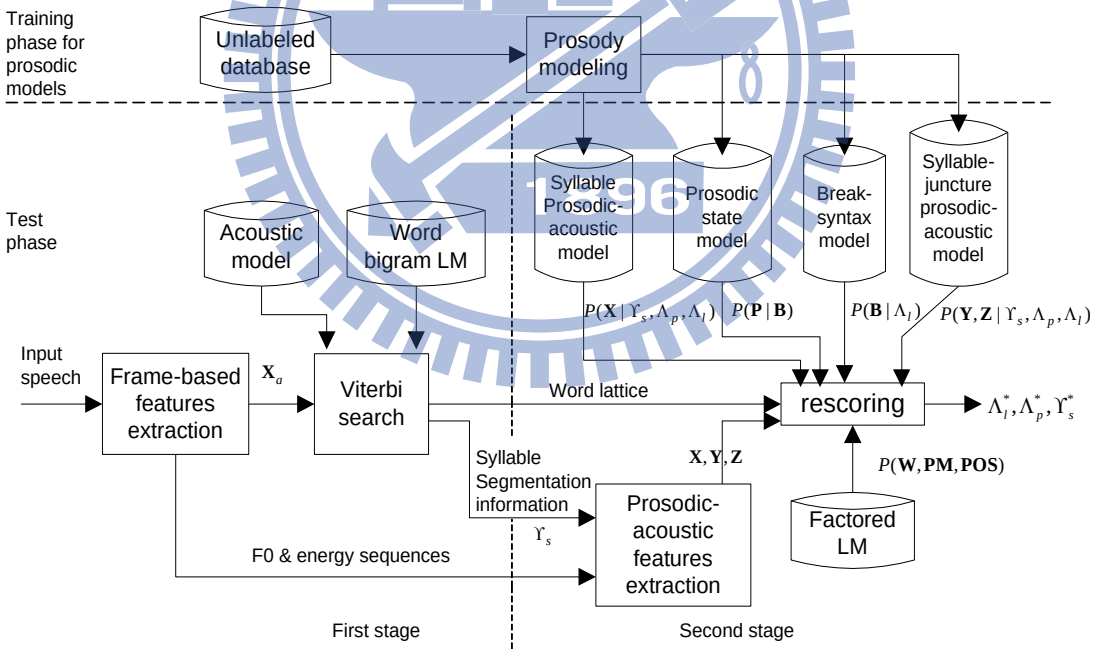


Figure 2.4: A block diagram of the two-stage prosody-assisted ASR system.

In the second-stage rescoring process, a product of sixteen 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 Equation (2.2). For considering the relative importance of each individual model to ASR, a log-linear

combination scheme to integrate these sixteen probabilities is adopted in this study:

$$L(S,\Lambda_\alpha)=\log C(\Lambda_\alpha)+\sum_{j=1}^{16}\alpha_j \log p_j \qquad (2.17)$$

where $S=[p_1\cdots p_{16}]$ is a 16-dimensional vector formed by these sixteen probabilities; $\Lambda_\alpha=[\alpha_1\cdots\alpha_{16}]$ is a weighting vector; and $C(\Lambda_\alpha)$ 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.

2.4 Experimental Results

2.4.1 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. 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 8-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 14993. 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.

2.4.2 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 Equation (2.11)-(2.13)) 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 a 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.

A quantitative analysis of the prosody modeling result is given as follows. Table 2.2 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].

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. The excluding of ill-pronounced training utterances is to 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 Equation (2.11)-(2.13)) 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 a node splitting. Finally, the total numbers of nodes (leaf nodes) obtained were 63(31) and 46(27) for these two models, respectively.

A quantitative analysis of the prosody modeling result is given as follows. Table 2.2 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. These agreed with the prior linguistic knowledge and generally matched with those of other previous studies [32], [33].

Table 2.2: APs of Five Tones

Tone	1	2	3	4	5
Pitch mean (log-Hz)	0.097	-0.05	-0.11	0.065	-0.069
Duration (ms)	9	5	-5	5	-54
Energy level (dB)	0.874	-0.623	-0.785	0.840	-1.567

Figure 2.5 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 a previous study [33].

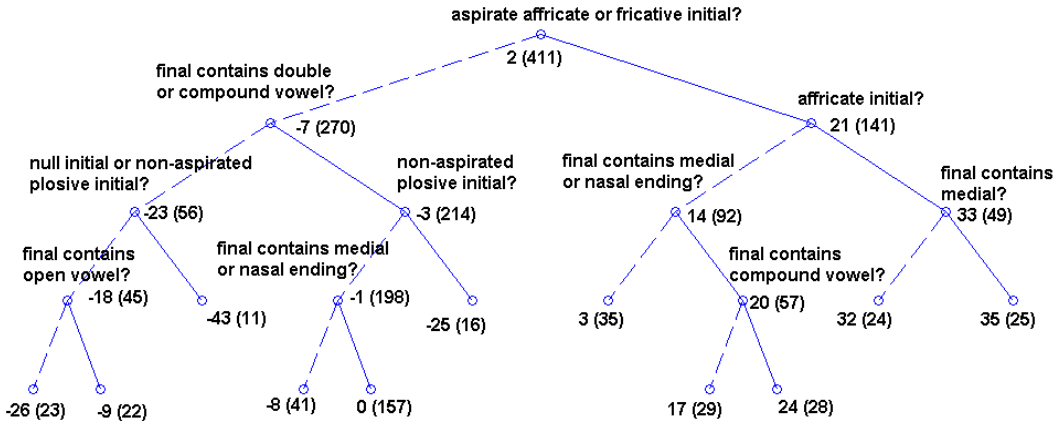


Figure 2.5: Decision tree analysis of duration APs of all 411 base-syllables. Numbers associated with each leaf node represents the average length (ms) of the APs and the sample count (in the bracket). Solid line indicates positive answer to the question and dashed line indicates negative answer.

Figure 2.6 depicts the forward and backward coarticulation patterns for the three extreme cases of break types, i.e., *B*0 (tightly coupling), *B*1 (normal) and *B*4 (major break). Several characteristics of these APs can be found. Firstly, 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 *B*0 junctures and the least for *B*4 junctures. Thirdly, for tightly coupling *B*0 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 $\{\beta^{f}_{B,t} \mid t^{n}_{n-1} = (1,2), (1,3), (2,2), (2,3)\}$ were due to H-L mismatches, while the downward bending at the beginning parts of $\{\beta^{f}_{B,t} \mid t^{n}_{n-1} = (3,1), (3,4)\}$ corresponded to L-H mismatches. Figure 2.7(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. Fourthly, the well-known *sandhi* rule that Tone 3-Tone 3 will change to Tone 2-Tone 3 had been learned in the backward coarticulation AP of 3-3 tone pair. Figure 2.7(b) illustrates this effect. Lastly, the forward coarticulations were generally larger than the backward coarticulations. The above mentioned characteristics generally conformed well to the observation found by Xu [34].

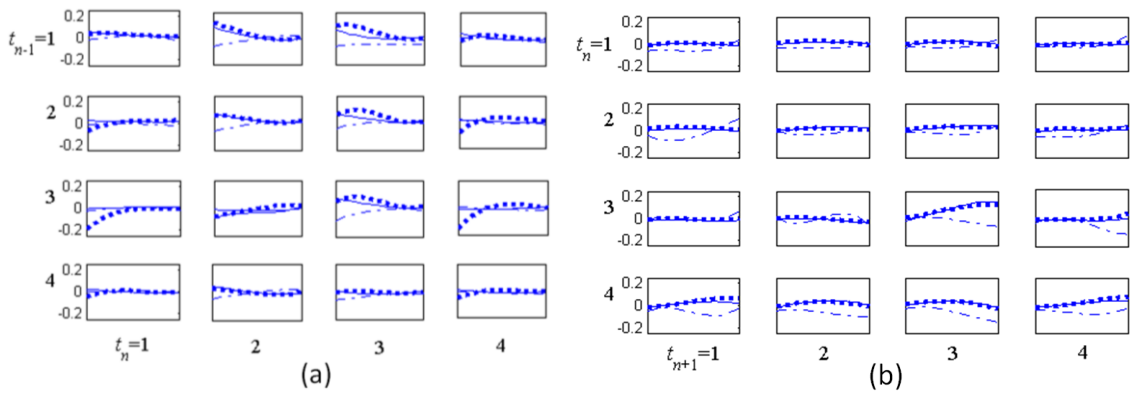


Figure 2.6: (a) Forward and (b) backward coarticulation patterns, $\boldsymbol{\beta}^{f}_{B_{n-1},t^{n}_{n-1}}$ and $\boldsymbol{\beta}^{b}_{B_{n},t^{n+1}_{n}}$, for *B*0 (point line), *B*1(solid line), and *B*4(dashed line).

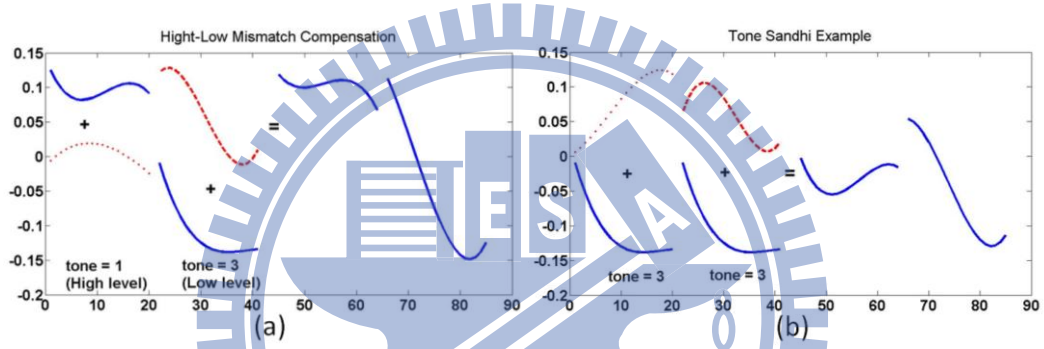


Figure 2.7: Two examples demonstrate the effects of coarticulation APs: (a) Tone 1-Tone 3 and (b) the sandhi rule of Tone 3-Tone 3. Solid lines (left): basic tone pitch patterns; point lines: backward APs; dashed lines: forward APs; and solid lines (right): the resulting pitch patterns.

Figure 2.8 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 *B*3 and *B*4, while they were on *B*0 and *B*1 for nodes in the intraword sub-tree. Moreover, phonetic information was important for the intraword sub-tree to further discriminate between *B*0 and *B*1. For the PM sub-tree, the type of PM was

important. Fig. 2.9 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.

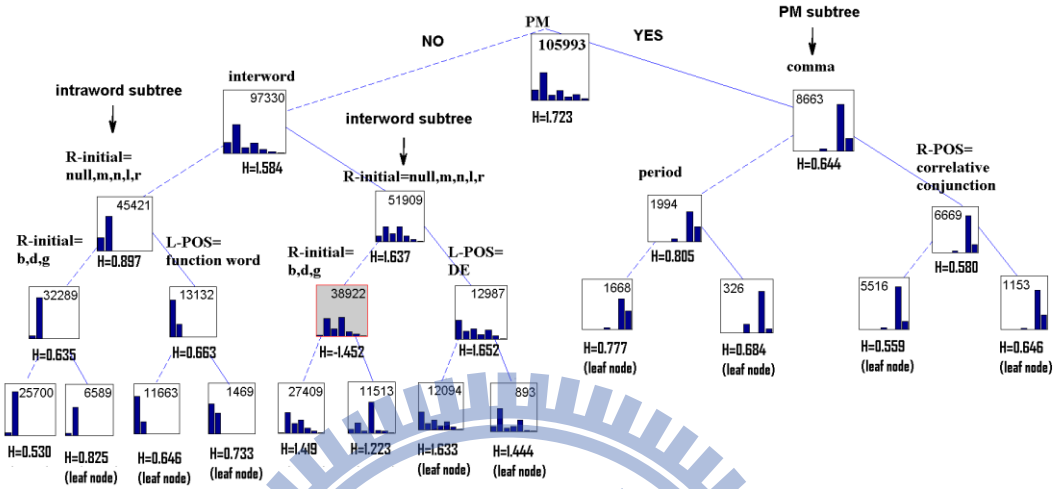


Figure 2.8: Decision tree for the break-syntax model. The bar plot associated with a node denotes the distribution of these seven break types (*B*0, *B*1, *B*2-1, *B*2-2, *B*2-3, *B*3, *B*4, from left to right) and the number is the total sample count of the node. *H* is the Shannon entropy to measure the uncertainty of break type distribution.

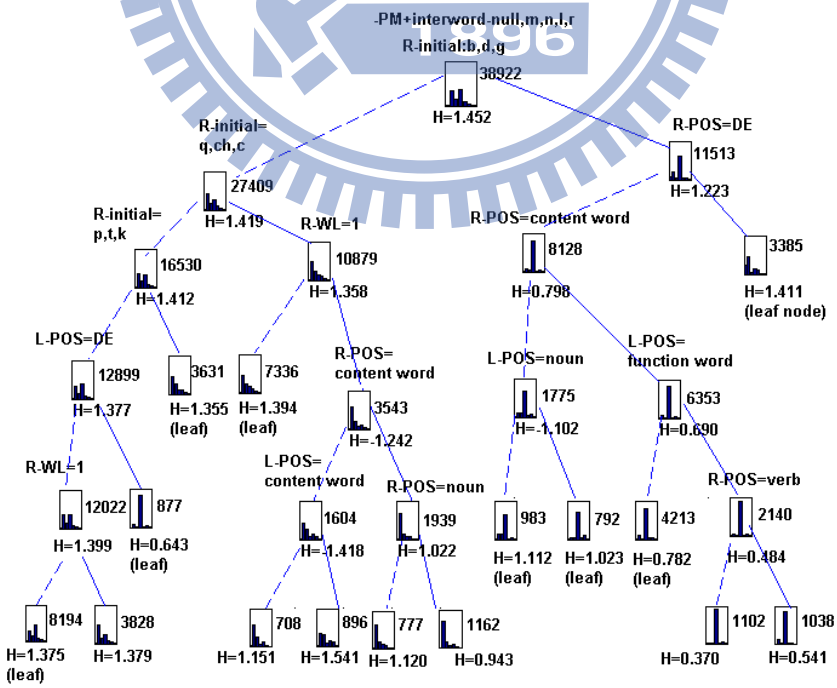


Figure 2.9: The deeper part of the decision tree for the break-syntax model. It is the sub-tree starting from the shaded node shown in Figure 2.8.

Figure 2.10 shows the major parts of decision trees of the break-acoustic model for the 7 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, *B*2-3 was similar to *B*1 and *B*2-1 in the distributions of pause duration, and energy-dip level. *B*2-1, *B*3, and *B*4 had positive normalized pitch jumps in average, while *B*0, *B*1, and *B*2-3 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 *B*2-2, *B*2-3, *B*3, and *B*4 were relatively larger than those of *B*0, *B*1, and *B*2-1. 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 increased 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 *B*4 due to the relative uniformity on the prosodic-acoustic features of its data. The questions used to split trees of pause-related break types (i.e., *B*3 and *B*2-2) tended to be related to higher-level linguistic 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 other non-pause break types (i.e., *B*0, *B*1, *B*2-1 and *B*2-3).

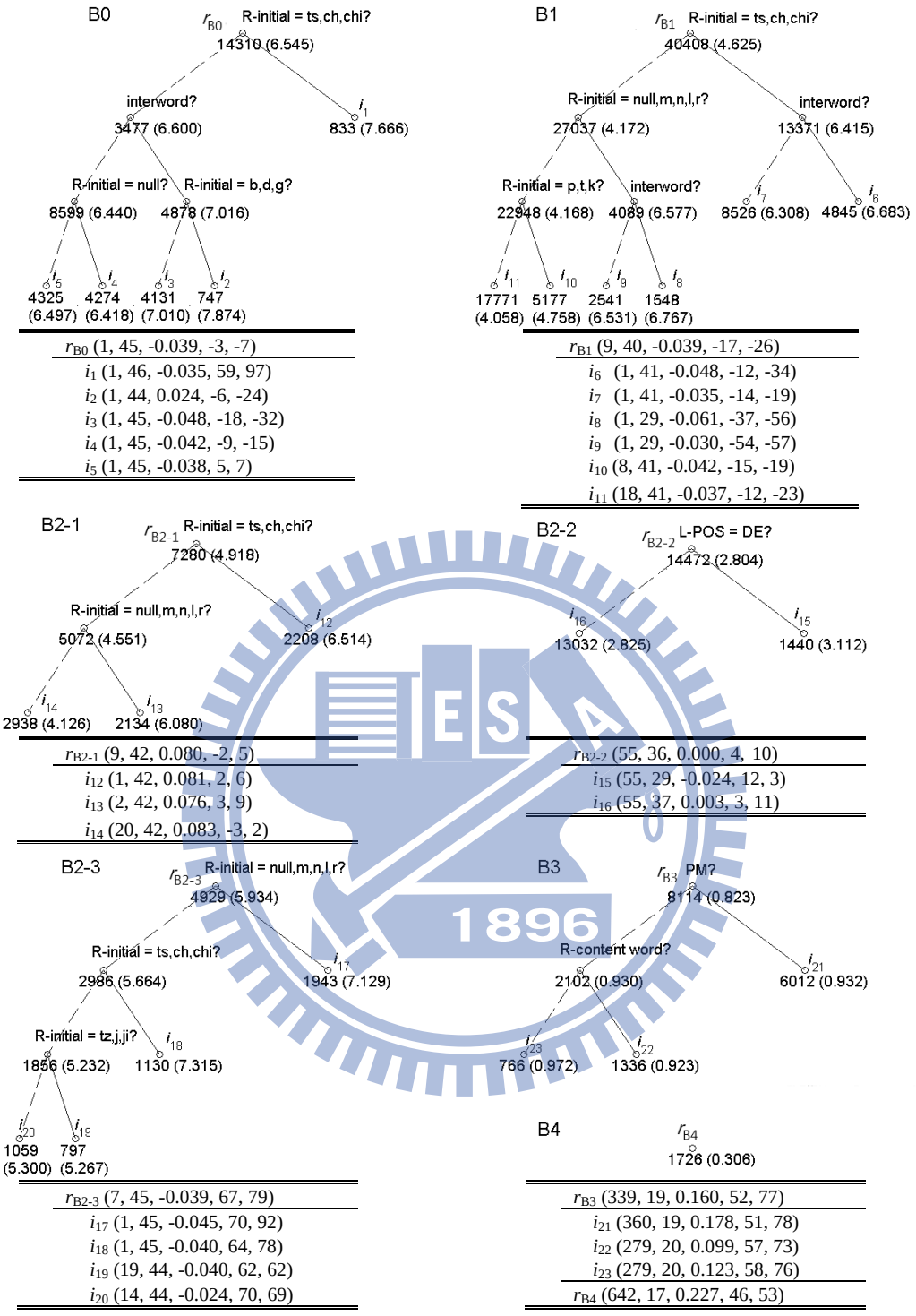


Figure 2.10: Decision trees of the break-acoustics model for 7 break types. Solid (dash) line indicates positive (negative) answer to the question. Numbers in a node are sample count and average likelihood per sample (in a bracket). The statistics for each node are shown in the bracket of the tables 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.

Figure 2.11 illustrates the transitions of pitch prosodic state $P(p_n \mid p_{n-1}, B_{n-1})$ for seven break types. For *B*0 and *B*1, the general high-to-low, nearby-state transitions showed that the syllable log-F0 level declined slowly within PWs. For *B*2-2, it had both high-to-low and low-to-high state transitions. For *B*2-1, *B*3, and *B*4, their low-to-high state transitions showed clearly the phenomena of syllable log-F0 level resets across PWs, PPhs, and BG/PGs. Comparing with these clear log-F0 level resets, the resets of *B*2-2 were insignificant. The transition of *B*2-3 is similar to those of *B*0 and *B*1. This implies no apparent pitch reset exists at the duration-lengthening juncture of *B*2-3. These phenomena were similar to those found in our previous study on the database of a single female speaker [20]. Table 2.3 lists a summary of the parameter numbers (#para) used in these 12 prosodic models

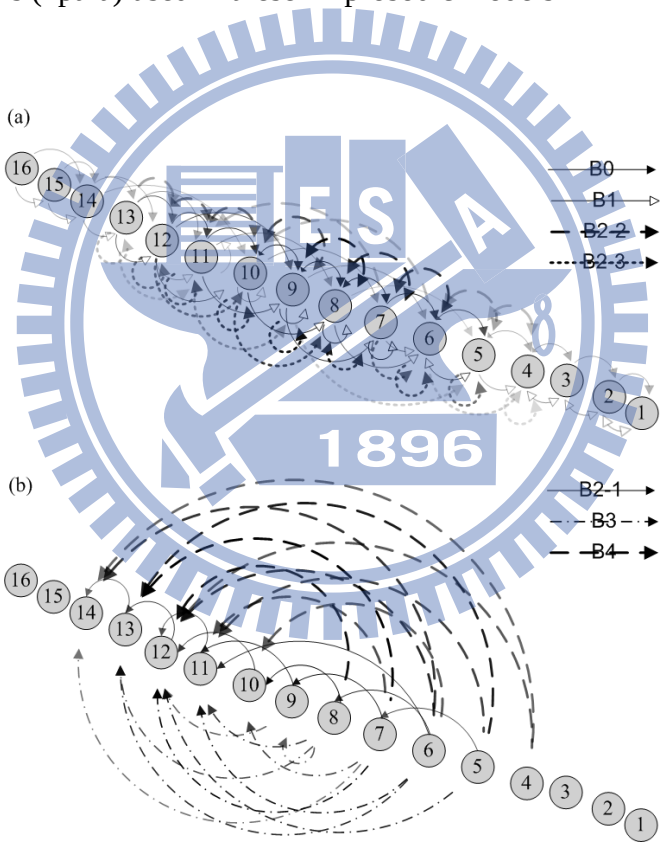


Figure 2.11: The most significant prosodic state transitions for (a) *B*0, *B*1, *B*2-2 and *B*2-3, and (b) *B*2-1, *B*3 and *B*4. Here, the number in each node represents the index of the prosodic state. Note that larger state index represents higher log-F0 value and darker lines represent more important state transitions.

Table 2.3: 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	1597	APs: (5 tones+ 16 states) ×3, 1400 coarticulations, 82 base-syllables, 40 final types, 12 means & variances

2.4.3 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 firstly 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 4 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 2.4 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 those of 29.6%, 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 it was expected to be useful for tone/word discrimination. We hence designed and tested two schemes of incorporating prosodic models. Scheme 1 incorporated the 6 juncture-based prosodic models, i.e., the break-syntax model and the 5 syllable-juncture 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 syllable 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 2.4. 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 2.4: 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

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 2.5 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, $N_{\text{corretW,corretPOS}}$, to the total number of correctly recognized words; while recall is defined as the ratio of $N_{\text{corretW,corretPOS}}$ 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 2.5: 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 2.6 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 re-evaluated 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 2.6: 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

Table 2.7 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 proper use of tone information in the prosody modeling for syllable pitch contour and syllable duration.

Table 2.7: 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

An error analysis was conducted to examine the recognition results in more detail. Firstly, 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. Figure 2.12 illustrates an example. As shown in the figure, there were four prosodic phrases (PPh's) separated by *B*3. 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 *B*2-1 break after the corrected word.

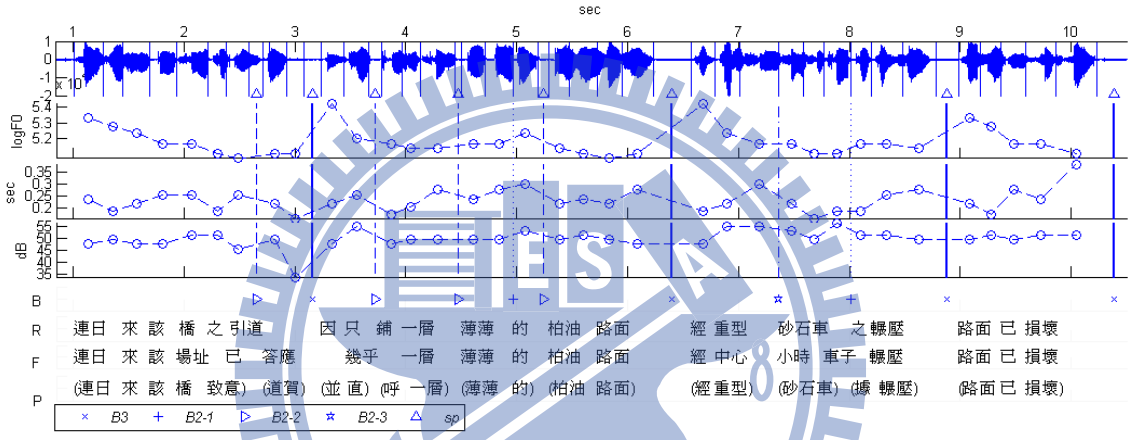


Figure 2.12: An example of recognition results for a partial paragraph. Eight panels represent, respectively, waveform, prosodic state AP+global mean of syllable log-F0 level, syllable duration, and syllable energy level, break type (B), reference transcription (R), result of baseline scheme (F) and 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).

Secondly, 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. Figure 2.13 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 *B*3 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.

(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) *B*2-2 公會(association) *B*0 理事長(council chairman) *B*3 或(or) *B*2-2 真心(true heart) *B*3…

Figure 2.13: An 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 Scheme 2 system.

Thirdly, 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 $P(sd_n \mid q_n, s_n, t_n)$ shown in Equation (2.8). 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 4-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 Figure 2.12 as an example, we can describe the prosody structure of the utterance as follows. On the top level, there are four prosodic phrases (PPh's) separated by three *B*3 breaks. From the first two panels of Figure 2.12, we find that all three *B*3 breaks were associated with long pauses and large pitch resets. So, these three *B*3 breaks were all labeled well. On the next level, there are 2, 5, 3 and 1 prosodic words (PWs) in these four PPh's, respectively. Within these four PPh's, PWs were separated by (*B*2-2), (*B*2-2, *B*2-2, *B*2-1, *B*2-2), (*B*2-3, *B*2-1) and (-). As shown in the first three panels of Figure 2.12, all four *B*2-2 breaks were associated with short

pauses, the *B*2-3 break was associated with a pre-boundary lengthening, and the two *B*2-1 breaks were associated with medium pitch resets. So, they were all properly labeled. Lastly, the bottom level is composed of syllables separated by *B*0 or *B*1 breaks. It is noted that *B*0 and *B*1 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 2.7 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 2.7, 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.

Table 2.7: 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

2.5 Conclusions for Chapter 2

In this chapter, 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 4-layer prosody-hierarchy model, various linguistic features of texts, and the 8 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.

Chapter 3 An Application of Prosody-Assisted Mandarin ASR to Speech Coding

Motivated by the success of the new prosody-assisted ASR system discussed in Chapter 2, we apply it to the coding of prosodic information. Section 3.1 presents the proposed speech coding system. Performance evaluation of the new speech coding system is discussed in Section 3.2. Lastly, some conclusions are given in Section 3.3.

3.1 The Proposed Speech Coding System

Figure 3.1 shows a schematic diagram of the proposed Mandarin-speech coding system. In the encoder, input speech signal is firstly recognized by the prosody-assisted Mandarin ASR system (PA-ASR) [35],[36] with an HMM-based acoustic model (AM), a factored language model (FLM) [24] and a hierarchical prosodic model (HPM) [20]. Three types of information are transcribed by the speech recognition. One is linguistic features including strings of base-syllable, tone, word, POS and PM. Another is prosodic features including tag sequences of syllable prosodic state and inter-syllable break type. It is worth to note that these two prosodic tag sequences can be used to form a hierarchical prosody structure of the input speech. The other is the segmentation information of various levels from HMM state to word.

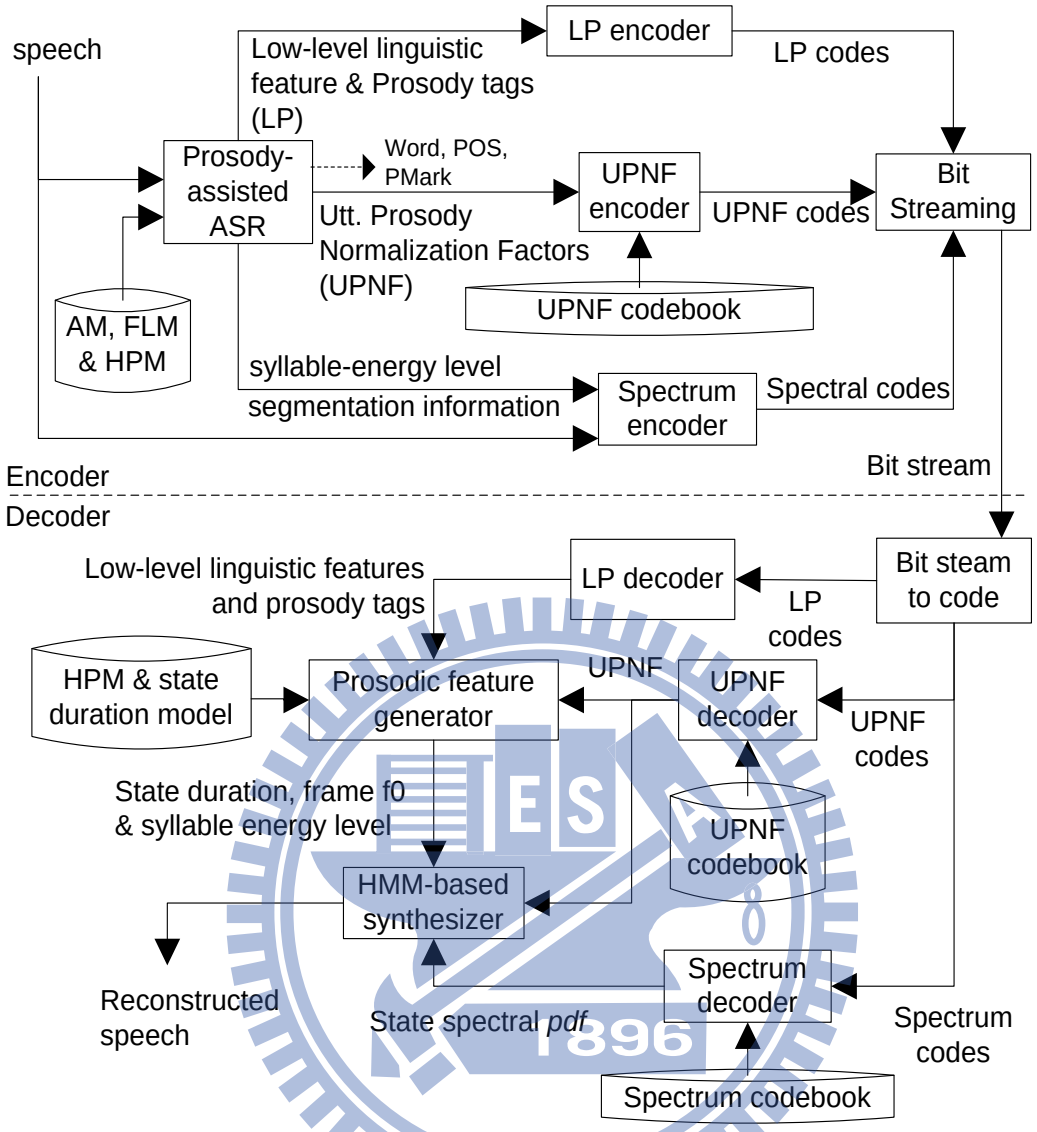


Figure 3.1: A schematic diagram of the proposed speech coding system.

By using some low-level linguistic features and prosodic tags (LP), we can reconstruct prosodic-acoustic features, including syllable pitch contour, syllable duration, syllable energy level, and inter-syllable pause duration with the help of HPM. So, we only need to encode those LP features for prosody reconstruction in the decoder. It is noted that prosodic features used in PA-ASR are pre-normalized by speaker-level (training phase) or utterance-level (test phase) mean and variance. Therefore, an additional utterance prosody normalization factor (UPNF) encoder is required for encoding these prosody normalization factors. By using the HMM-state segmentation information, we can extract state-based spectral features and encode them by vector quantization (VQ).

In the decoder, we first use the decoded LP features to reconstruct the four prosodic-acoustic features by HPM whose parameters are sent to the decoder in advance as side information. We then use base-syllable type and syllable duration to predict state durations by a state duration model. Lastly, by using the decoded state spectral features, the reconstructed prosodic-acoustic features, and the predicted state durations, an HMM-based speech synthesizer generates the output speech.

In the following subsections, we discuss the encoder and the decoder in more detail.

3.1.1 The Speech Encoder

As shown in Figure 3.1, the speech encoder is composed of four parts including a PA-ASR [35],[36], an LP encoder, a UPNF encoder, and a spectrum encoder. The PA-ASR system is a sophisticated speech recognizer discussed in Chapter 2 [35],[36]. Figure 2.4 displays its functional block diagram. It is a two-stage system to firstly use an AM and a bigram LM to generate a word lattice in the first stage decoding, and to then use an FLM [24] and an HPM [20] to finely decode from the word lattice the best linguistic sequences (i.e. base-syllable, tone, word, POS and PM) and their corresponding segmentation information, as well as prosody tag sequences (i.e. prosodic states and break types) that represent a hierarchical prosody structure of the input utterance. The AM is a syllable-based HMM model. It models each of 411 base-syllables as an 8-state left-to-right HMM. The FLM is an extension of the conventional trigram model to additionally consider POS and PM aside from word. The HPM consists of various prosodic sub-models to describe the relationship of prosodic tags, prosodic-acoustic features, and linguistic features.

Four sub-models of the HPM are involved in the coding process. They include three syllable prosodic-acoustic models, which are used to describe the variations of syllable pitch contour, duration and energy level, and one prosodic-acoustic model which describes the variation of syllable-juncture pause duration influenced by some linguistic features. For syllable pitch contour, it is formulated as an additive model:

$$sp_n = sp_n^r + \beta_{t_n} + \beta_{p_n} + \beta^f_{B_{n-1},t^n_{n-1}} + \beta^b_{B_n,t^{n+1}_n} + \mu_{sp} \tag{3.1}$$

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 residual of modeling sp_n; β_{t_n} and β_{p_n} are the affecting patterns (APs) for tone t_n and prosodic state tag p_n, respectively; $\beta^f_{B_{n-1},t^n_{n-1}}$ and $\beta^b_{B_n,t^{n+1}_n}$ are the forward and backward coarticulation APs contributed from syllable $n-1$ and syllable $n+1$, respectively; and μ_{sp} is the global mean of pitch vector. Here, B_n is the break tag after syllable n. Similarly, syllable duration and energy level are modeled as

$$sd_n = sd_n^r + \gamma_{t_n} + \gamma_{s_n} + \gamma_{q_n} + \mu_{sd} \tag{3.2}$$

$$se_n = se_n^r + \omega_{t_n} + \omega_{f_n} + \omega_{r_n} + \mu_{se} \tag{3.3}$$

where $\gamma_{t_n}/\omega_{t_n}$, γ_{s_n}, ω_{f_n} and $\gamma_{q_n}/\omega_{r_n}$ are the APs of tone t_n, base-syllable s_n, final type f_n, and prosodic state tags q_n/r_n; and μ_{sd} and μ_{se} are global means. To reconstruct these three prosodic-acoustic features using the three sub-models in the decoder, we need to encode and transmit low-level linguistic features of tone, base-syllable and final types as well as prosodic features of break type and prosodic state tags. Besides, all affecting patterns are sent as side information. It is noted that we neglect the coding of the residuals because they all have small variances.

The fourth sub-model describes the variation of inter-syllable pause duration by break-dependent decision trees (BDTs). For each break type, a decision tree is used to determine the pdf of pause duration according to linguistic features. For reconstructing the pause duration, we needs to send the information of the break tag and the residing leaf node of the associated decision tree for each inter-syllable juncture to the decoder. All pdfs of leaf nodes in these seven decision trees are also sent to the decoder as side information.

Table 3.1 shows the bit assignment of the encodings of these low-level linguistic features of tone, base-syllable and final types, prosodic tags of prosodic state and break type, and leaf nodes of BDTs. Notice that the BDT is constructed for each break type, and each BDT has different number of leaf nodes. Therefore, the bit length is variable for each given known break type.

Table 3.1: Bit assignment for encoding linguistic features and prosody tags

Symbol	# of symbol	bit
Lexical tone t_n	5	3
Base-syllable type s_n	411	9
Pitch prosodic state p_n	16	4
Duration prosodic state q_n	16	4
Energy prosodic state r_n	16	4
Break type B_n	7	3
BDT leaf node index T_n for $B0$, $B1$, $B2$-1, $B2$-2, $B2$-3, $B3$, $B4$	5/7/3/2/4/3/1	3/3/2/1/2/2/0
Total bits per syllable (maximum)		30

For avoiding taking care of the speaker/utterance variability of prosodic-acoustic features in HPM, they are pre-normalized. For syllable pitch contour, a scheme of frame-based F0 value normalized by speaker-level (training phase) or utterance-level (test phase) mean and variance is adopted; while for both syllable duration and syllable energy level, they are simply normalized by their corresponding speaker-/utterance-level means and variances. These normalization factors are needed to be encoded and sent to the decoder. In this study, they are scalar-quantized independently by the UPNF encoder. Their codebooks are also sent to the decoder as side information.

Since we want to use the HMM-based speech synthesizer in the decoder to generate the output speech, we extract 25-dimensional mel-generalized cepstral (MGC) [37] vector including the zero-th coefficient for each 25ms frame with 5ms shift. Blackman window is used in the feature extraction. Besides, delta and delta-delta MGCs are also extracted. In the training phase, we calculate the pdf parameters (i.e., mean and variance) of each MGC coefficient for each HMM state using the training data with the time-aligned segmentation information provided by the PA-ASR system. In the test phase, we first calculate the mean vector of 25-dimentional MGC vectors for each state segment and then subtract the mean MGC vector of the corresponding state of the recognized base-syllable to obtain a residual vector. Lastly, we encode all state-based residual vectors by vector quantization (15 bit for each state). Both the pdf

parameters of all HMM states and the VQ codebooks are sent to the decoder as side information. It is noted that the energy coefficient in each state MGC vector is pre-normalized by the energy level of the associated syllable. Table 3.2 summarizes the side information of the coding system.

Table 3.2: Side information of the proposed coding system

Type	parameter #
Lexical tone APs: $\beta_t / \gamma_t / \omega_t$	5/5/5
Coarticulation APs: $\beta^f_{B,t} / \beta^b_{B,t}$	180/180
Prosodic state APs: $\beta_p / \gamma_q / \omega_r$	16/16/16
Global mean APs: $\mu_{sp} / \mu_{sd} / \mu_{se}$	1/1/1
Base-syllable type and final type APs: γ_s / ω_{f_n}	411/40
BDT leaf node mean: $\mu^{pd}_{T_n}$	25
Spectrum codebook	1056
MGC *pdf*s of all HMM states	26304
Normalization factor codebooks	384
Total	28646

3.1.2 The Speech Decoder

The task of the speech decoder is to reconstruct speech signal by using the decoded linguistic, prosodic and spectral parameters. As shown in Figure 3.1, the speech decoder consists of five parts including the LP decoder, the UPNF decoder, the spectrum decoder, the prosodic-acoustic feature generator, and an HMM-based speech synthesizer [38]. The LP decoder generates low-level linguistic features and prosody tags by looking up tables. The spectrum decoder uses the spectrum codebook to generate the output spectral features of each state from the input codeword index. The prosodic feature generator reconstructs the three prosodic-acoustic features and pause duration by HPM using the decoded low-level linguistic features and prosody tags. These three prosodic-acoustic features are de-normalized by using the decoded utterance-level factors. After obtaining syllable duration, we then predict state durations. Lastly, the HMM-based speech synthesizer reconstructs the input speech signal by using the state spectral features, state duration and the associated prosodic-acoustic features.

In state duration prediction, we assume that the state duration is normally distributed and affected by base-syllable type s_n, i.e

$$P(d_{n,c} \mid s_n, c) = N(d_{n,c}; \mu_c^{s_n}, \sigma_c^{s_n}) \tag{3.4}$$

where $d_{n,c}$ denotes the duration of the c-th state in the n-th syllable. Given the reconstructed syllable duration sd_n, state durations of the syllable can be obtained by maximizing the summed log likelihood, i.e.

$$d_{n,1}^* \ldots d_{n,C}^* = \arg \max_{d_{n,1} \ldots d_{n,C}} \sum_{c=1}^{C} \log P(d_{n,c} \mid s_n, c) \tag{3.5}$$

under the constraint

$$sd_n = \sum_{c=1}^{C} d_{n,c} \tag{3.6}$$

The resulting state duration can be obtained by

$$d_{n,c} = \mu_c^{s_n} + \rho \cdot (\sigma_c^{s_n})^2 \tag{3.7}$$

where

$$\rho = \left(sd_n - \sum_{c=1}^{C} \mu_c^{s_n} \right) / \left(\sum_{c=1}^{C} (\sigma_c^{s_n})^2 \right) \tag{3.8}$$

3.2 Performance Evaluation

The proposed model-based Mandarin-speech coding system was evaluated on a large Mandarin read speech database TCC300 [28] that mentioned in Section 2.4.1. Table 3.3 shows the performance of the PA-ASR system. Word, character, and base-syllable error rates of 20.7%, 14.4%, and 9.6% were achieved, respectively. This performance is very good as compared with most conventional HMM-based ASR methods. Since syllable insertion and deletion errors were expected to cause more serious degradation on the coding performance, we also list them in Table 3.3. As seen from the table, both of them are small.

Table 3.3: The performance of the PA-ASR (%)

WER	CER	SER	Syll-INS	Syll-DEL	Syll-SUB
20.7	14.4	9.6	0.55	0.83	8.5

We then examined the performance of the coding system. Two cases were examined. One was the inside test in which both the speech utterance and the associated text were given. In this case, we first segmented the speech by time-alignment using the AM, and then labeled the prosodic tags automatically by the HPM. We then performed the encoding and decoding operations to reconstruct the speech. The other case was the outside test in which only the speech utterance was given. This is the case of the proposed coding system discussed in Section 2.

Table 3.4 shows the root-mean-square errors (RMSE) of the reconstructed four prosodic features. Here, all six utterance-level normalization factors were encoded using 6-bit scalar quantizers. Table 3.5 shows the RMSE of the reconstructed pause duration for different break types. Since major breaks like *B*3 and *B*4 are tolerant of larger errors, the performance was good. The average bit rates were 528 and 543 bits/s for the inside and outside tests, respectively. These data rates are low. Figure 3.2 shows an example of the reconstructed prosodic features of an utterance of the outside test. As shown in the figure, most reconstructed prosodic features were close to their reference values.

Table 3.4: The RMSE of the reconstructed prosodic features

	F0 (Hz)	Syllable duration (ms)	Syllable energy level (dB)	Pause duration (ms)
Inside test	11.4	18.4	0.52	73.8
Outside test	14.7	16.8	0.20	75.6

Table 3.5: The RMSE (ms) performance of the reconstructed pause duration with respect to different break types

	*B*0	*B*1	*B*2-1	*B*2-2	*B*2-3	*B*3	*B*4
Inside	19.3	26.5	75.6	149.2	35.0	177.9	312.9
Outside	12.4	17.1	88.3	178.4	39.6	176.9	292.7

Table 3.6: Bit rates for inside and outside tests

		Average	Max	Min
Inside	prosody	104.56	163.79	42.23
	spectral	423.73	661.90	178.07
outside	prosody	107.55	147.20	78.00
	spectral	435.06	594.44	318.05

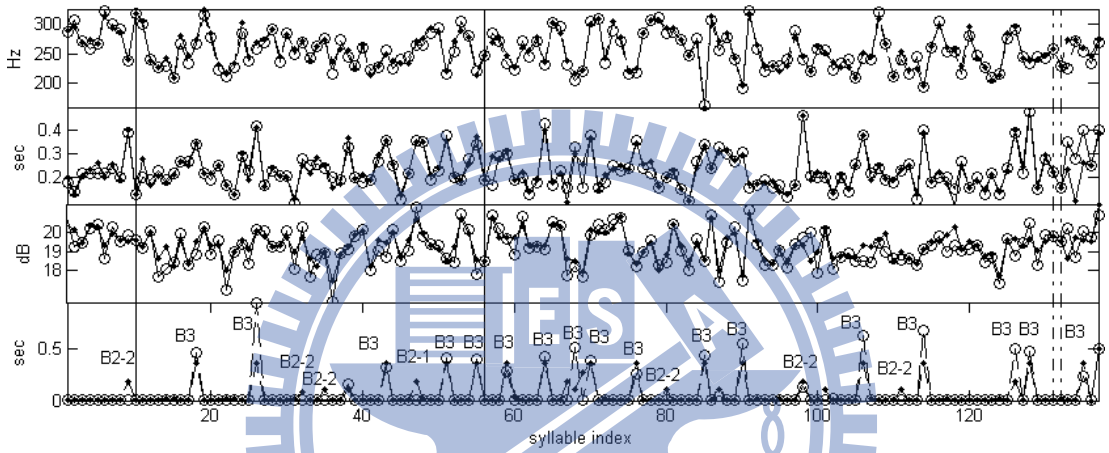


Figure 3.2: An example of the reconstructed prosodic features of an utterance. From top to bottom: syllable pitch mean, syllable duration, syllable energy level, and pause duration. (open circle: reference, dot: recognition result, solid line: deletion, dash dot line: insertion).

Lastly, an informal listening test was performed. Generally, all reconstructed speeches sounded good. The effects of recognition errors were not serious. Most substitution, deletion, and insertion errors were slightly perceptible. This mainly resulted from encoding and sending the spectral features to the decoder.

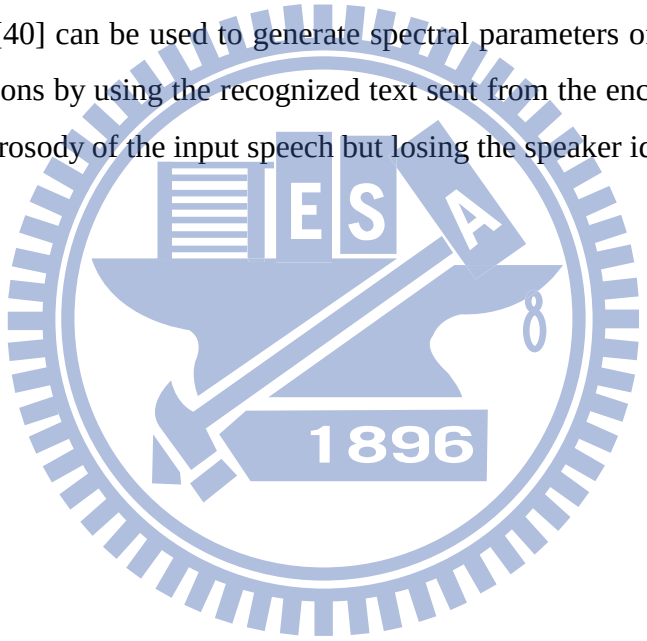
3.3 Conclusions for Chapter 3

In this chapter, a model-based Mandarin-speech coding system has been discussed. It differs from the conventional speech coding system on using a prosody-assisted ASR in the encoder to extract high-level linguistic and prosodic features to assist in improving the coding efficiency. Experimental results showed that

high-quality reconstructed speech can be obtained at a low data rate of 543 bits/s.

Another advantage of the proposed coding system can be found. By properly adjusting the prosodic features, we may modify the prosody of the reconstructed speech, e.g. changing the speech rate.

The proposed coding system can also operate on another two modes. One is the case of knowing both the speech signal and the associated text. This case has been examined as the inside test discussed in Section 3.2. An application of the mode is the speech coding of story readings in an electronic book. Prosody modification will be the most attractive feature of the application. The other mode is the case of low-rate speech coding without transmitting the spectral parameters. A text-to-speech system, such as the HTS [40] can be used to generate spectral parameters of a standard voice for their substitutions by using the recognized text sent from the encoder. In this case, we can keep the prosody of the input speech but losing the speaker identity.



Chapter 4 Conclusions and Future Works

4.1 Conclusions

In this dissertation, we present a study on involving abundant prosodic cues in the prosody modeling for assisting in ASR. Experimental results confirmed that the new prosody-assisted ASR system performs effectively on improving the syllable/character/word error rates. Several advantages of the proposed system can be found. First, these 12 prosodic models of the HPM 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 4-layer prosody-hierarchy model, various linguistic features of texts, and the 8 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.

We also present a study on applying the new prosody-assisted ASR to the coding of prosodic information. It demonstrates the feasibility of using a prosody-assisted ASR in the encoder to extract high-level linguistic and prosodic features to assist in improving the coding efficiency. Experimental results showed that high-quality reconstructed speech can be obtained at a low data rate of 543 bits/s. Aside from coding efficiency, another advantage of the proposed coding system can be found. By properly adjusting the parameters of the HPM, we may modify the prosody of the reconstructed speech, e.g. changing the speech rate. The proposed coding system can also operate on another two modes. One is the case of knowing both the speech signal

and the associated text. This case has been examined as the inside test discussed in Section 3.2. An application of the mode is the speech coding of story readings in an electronic book. Prosody modification will be the most attractive feature of the application. The other mode is the case of low-rate speech coding without transmitting the spectral parameters. A text-to-speech system, such as the HTS [40] can be used to generate spectral parameters of a standard voice for their substitutions by using the recognized text sent from the encoder. In this case, we can keep the prosody of the input speech but losing the speaker identity.

4.2 Future Works

Some further works are worth doing in the future. Firstly, 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]. Secondly, 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 firstly 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. Thirdly, modifying the proposed approach to reduce its computational complexity is needed for on-line 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 *B*3 and *B*4 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. Fourthly, 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. Fifthly, 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.

Bibliography

[1] S. Ananthakrishnan and S. Narayanan, "Unsupervised adaptation of categorical prosody models for prosody labeling and speech recognition," IEEE Transactions on Audio, Speech and Language Processing, 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," Proc. of ICASSP, pp. IV-873-IV876, 2007

[3] S. Ananthakrishnan and S. Narayanan, "Prosody-enriched lattices for improved syllable recognition," Proc. of INTERSPEECH, pp. 1813-1816, 2007

[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 Transactions on Audio, Speech and Language Processing, vol. 14 no.1, pp.232-245, January 2006.

[5] D. H. Milone and A. J. Rubio, "Prosodic and accentual information for automatic speech recognition," IEEE Transactions on Audio, Speech and Language Processing, vol. 11, no. 4, pp. 321-333, July 2003.

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

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

[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 Communication, vol. 36, pp. 247-265, 2002

[9] J.-T. Huang and L.-S. Lee, "Improved large vocabulary Mandarin speech recognition using prosodic features," Proc. of SPEECH PROSODY, 2006.

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

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

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

[13] C. Ni, W. Liu, and B. Xu, "Using prosody to improve Mandarin automatic speech recognition," Proc. of 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 Transactions on Audio, Speech and Language Processing, vol. 14, no. 5, pp. 1526-1540, September 2006.

[15] E. Shriberg and A. Stolcke, "Prosody modeling for automatic speech recognition and understanding," in Proc. workshop on mathematical foundations of natural language 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," Proc. of ICSLP, vol. 2, pp. 867-870, 1992.

[17] D. Hirst, and A. D. Cristo, "Intonation systems. a survey of twenty Languages," Cambridge University 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. on Audio, Speech and Language Processing, 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," Proc. of ICASSP, pp. 4565-4568, 2009.

[20] C.-Y. Chiang, S.-H. Chen, H.-M. Yu, and Y.-R. Wang, "Unsupervised joint prosody labeling and modeling for Mandarin speech," Journal of the Acoustic Society of America, 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 Communication, **46**, pp. 284-309, 2005.

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

[23] S.-H. Chen and Y.-R. Wang, "Vector quantization of pitch information in Mandarin speech," IEEE Transactions on Communications, vol. 38, no. 9, pp. 1317-1320, September 1990.

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

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

[26] P. Beyerlein, "Discriminative model combination," Proc. of ICASSP, pp. 481-484, 1998.

[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 Trends, ed. A.J. Rubio Ayuso, J.M. Lopez Soler, Springer-Verlag, Berlin-Hheidelberg, 1995.

[28] Mandarin microphone speech corpus – TCC300, http://www.aclclp.org.tw/use_mat.php#tcc300edu.

[29] "HTK Web-Site", http://htk.eng.cam.ac.uk. Accessed 2009

[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, pp. 49-52, 1986.

[31] C.-R. Huang, K.-J. Chen, F.-Y. Chen, Z.-M. Gao and K.-Y. Chen. 2000, “Sinica treebank: design criteria, annotation guidelines, and on-line interface,” Proceedings of 2nd Chinese Language Processing Workshop, Hong Kong, pp. 29-37, 2000.

[32] S.-H. Chen, W.-H. Lai, and Y.-R. Wang, “A statistics-based pitch contour model for Mandarin speech,” Journal of the Acoustical Society of America, vol. 117, no. 2, pp. 908–925, February 2005.

[33] S.-H. Chen, W.-H. Lai, and Y.-R. Wang, “A new duration modeling approach for Mandarin speech,” IEEE Transactions on Audio, Speech and Language Processing, vol. 11, no. 4, pp. 308–320, July 2003.

[34] Y. Xu, “Contextual tonal variations in Mandarin,” J. Phonetics 25, 61-83, 2007.

[35] Y.-L. Chou, C.-Y. Chiang, Y.-R. Wang, H.-M. Yu, S.-H. Chen, “Prosody labeling and modeling for Mandarin spontaneous speech,” Proc. of SPEECH PROSODY, Chicago, USA, May 2010.

[36] J.-H. Yang, M.-J. Liu, H.-H. Chang, C.-Y. Chiang, Y.-R. Wang, and S.-H. Chen, , “Enriching Mandarin speech recognition by incorporating a hierarchical prosody model”, Proc. of ICASSP, May 2011.

[37] S.-H. Chen, J.-H. Yang, C.-Y. Chiang, M.-C. Liu, and Y.-R. Wang, "A New Prosody-Assisted Mandarin ASR System", to appear in IEEE Transactions on Audio, Speech, & Language Processing , vol. 20, no. 5, July 2012.

[38] Tokuda, K., Masuko, T., Kobayashi, T. and Imai, S., “Mel-generalized cepstral analysis-a unified approach to speech spectral estimation,” Proceedings of the International Conference on Spoken Language Processing, pp. 1043–1046, Yokohama, Japan, September 1994.

[39] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T., “Speech parameter generation algorithms for HMM-based speech synthesis,” Proc. of ICASSP, pp.1315-1318, June 2000.

[40] Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A.W., Tokuda, K., “The HMM-based speech synthesis system version 2.0,” Proc. of ISCA SSW6, Bonn, Germany, Aug. 2007.

Publication List

Journal Paper

[1] Sin-Horng Chen, **Jyh-Her Yang**, Chen-Yu Chiang, Ming-Chieh Liu, and Yih-Ru Wang, "A New Prosody-Assisted Mandarin ASR System", to appear in *Trans. on IEEE Audio, Speech, & Language Processing*, VOL. 20, NO. 5, JULY 2012.

[2] Yuan-Fu Liao, **Jyh-Her Yang** and Sin-Horng Chen, "Soft-decision A Priori Knowledge Interpolation for Robust Telephone Speaker Identification", *Journal of the Chinese Institute of Engineers*, pp. 627-637, July 2009.

Conference Papers

[1] Yu, Hsiu-min, Hsiu-hsueh Liu, **Jyh-her Yang**, Chen-yu Chiang, and Sin-horng Chen, "Tonal Contrast and Pitch Range in L2 Taiwan Min Produced by Native Si-Xien Hakka Speakers," Presented at The 12th Conference on Min Languages. (第 12 屆閩語國際學術研討會). In Proceedings of the 12th Conference on Min Languages, pp. 333-347, Taipei, Taiwan, 2011.

[2] Tzu-Hsuan Chiu, Chen-Yu Chiang, Yuan-Fu Liao, **Jyh-Her Yang**, Yih-Ru Wang and Sin-Horng Chen, "Prosody-dependent Acoustic Modeling for Mandarin Speech Recognition," accepted by Speech Prosody 2012

[3] **Jyh-Her Yang**, Ming-Chieh Liu, Hao-Hsiang Chang, Chen-Yu Chiang, Yih-Ru Wang, and Sin-Horng Chen, "Enriching Mandarin speech recognition by incorporating a hierarchical prosody model", in Proc. of ICASSP 2011, pp. 5052-5055, 2011.

[4] Chen-Yu Chiang, **Jyh-Her Yang**, Ming-Chieh Liu, Yih-Ru Wang, Yuan-Fu Liao, and Sin-Horng Chen, "A New Model-based Mandarin-speech Coding System," in Proc. Interspeech 2011, Florence, Italy, pp 2561-2564, Sept. 2011.

[5] Yuan-Fu Liao, Zhi-Xian Zhuang, and **Jyh-Her Yang**, "Maximum Likelihood A Priori Knowledge Interpolation-Based Handset Mismatch Compensation for Robust Speaker Identification", to appear in Tsinghua Science and Technology, 2008.

[6] Yuan-Fu Liao, **Jyh-Her Yang**, Chi-Hui Hsu, Cheng-Chang Lee, and Jing-Teng Zeng, "A Reference Model Weighting-based Method for Robust Speech Recognition", in Proc. of *InterSpeech*, 2007

[7] Yuan-Fu Liao, Zhi-Xian Zhuang, and **Jyh-Her Yang**, "Maximum Likelihood A Priori Knowledge Interpolation-Based Handset Mismatch Compensation for Robust Speaker Identification", NCMMSC, 2007.

[8] **Jyh-Her Yang**, Yuan-Fu Liao, Yih-Ru Wang, and Sin-Horng Chen, "A New Approach of Using Temporal Information in Mandarin Speech Recognition", Speech Prosody' 2006, Dresden, Germany, May 2006.

[9] **Jyh-Her Yang** and Yuan-Fu Liao, "Unseen Handset Mismatch Compensation Based on A Priori Knowledge Interpolation for Robust Speaker Recognition", in Proc. of *ICSLP*'2004.

[10] **Jyh-Her Yang** and Yuan-Fu Liao, "Unseen Handset Mismatch Compensation Based On Feature/Model-Space A Priori Knowledge Interpolation For Robust Speaker Recognition", in Proc. of *ISCLSP*'2004.

博士候選人資料

姓　名：楊智合

性　別：男

出生年月日：民國 68 年 12 月 10 日

籍　貫：桃園縣

學　歷：

國立雲林科技大學電機系學士班畢業(87 年 8 月～91 年 6 月)

國立台北科技大學電通所碩士班畢業(91 年 8 月～93 年 7 月)

國立交通大學電信工程研究所博士班(93 年 8 月～101 年 7 月)

論文題目：

新韻律輔助中文語音辨認系統及其應用

A New Prosody-Assisted Mandarin ASR System and Its Application