

國立交通大學

電信工程學系

博士論文

非監督式中文語音韻律標記及韻律模式

Unsupervised Joint Prosody Labeling and  
Modeling for Mandarin Speech

研究生： 江振宇

指導教授： 陳信宏博士  
王逸如博士

中華民國九十八年三月

非監督式中文語音韻律標記及韻律模式

Unsupervised Joint Prosody Labeling and  
Modeling for Mandarin Speech

研究生：江振宇

Student: Chen-Yu Chiang

指導教授：陳信宏 博士

Advisors: Dr. Sin-Horng Chen

王逸如 博士

Dr. Yih-Ru Wang



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  
Hsinchu, Taiwan

2009 年 3 月

# 推 薦 函

中華民國九十八年二月十一日

一、事由：本校電信研究所博士班研究生 江振宇 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 江振宇 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，江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，江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，江君已具備國立交通大學電信研究所應有之教育及  
訓練水準，因此特推薦

江君參加國立交通大學電信工程學系博士班論文口試。

交通大學電信工程學系教授 陳信宏

交通大學電信工程學系副教授 王逸如

# 國立交通大學

## 論文口試委員審定書

本校 電信工程 學系博士班 江振宇 君

所提論文 非監督式中文語音韻律標記及韻律模式

Unsupervised Joint Prosody Labeling and Modeling  
for Mandarin Speech

合於博士資格水準、業經本委員會評審認可。

口試委員：  
王州  
張文輝  
鄭秋汎  
李琳山  
陳作宏  
王逸如

指導教授：  
陳作宏  
陳仰寧  
王逸如

系主任：陳仰寧 教授

中華民國九十八年三月十三日

Department of Communication Engineering  
National Chiao Tung University  
Hsinchu, Taiwan, R.O.C.

Date: Mar 13, 2009

We have carefully read the dissertation entitled  
Unsupervised Joint Prosody Labeling and Modeling for  
Mandarin Speech

submitted by Chen-Yu Chiang in partial  
fulfillment of the requirements of the degree of DOCTOR OF  
PHILOSOPHY and recommend its acceptance.

Hsienshuan Wang

Wen-Hsi Chang

Jing-Hu Wang

Yih-Ru Wang

Chang-Hong

Je-Jee

Sim-Hyung Chen

Thesis Advisor: Sim-Hyung Chen Yih-Ru Wang

Chairman

Department of Communication Engineering: Bo-King Chen

# 非監督式中文語音韻律標記及韻律模式

研究生：江振宇

指導教授：陳信宏 博士  
王逸如 博士

國立交通大學電信工程學系

## 中文摘要

韻律模式可使用在許多語音處理應用上，如語音合成及語音辨認。一般傳統建構韻律模式的方法，是先對語音信號標示出韻律標記以表示重要的韻律訊息，進而建構韻律模式。傳統韻律標記的方法是以人工觀察並聆聽語音信號進行標記，此方法之缺點為：（1）因為不同標記人的主觀認定不同造成標記結果不一致，（2）即使是同一個標記人進行標記，長時間進行下來，亦難以保持一致性，（3）耗時。上述所論及的不一致性，進而可能使得韻律模式在語音處理應用上的表現不佳。為了改善以上缺點，在本研究中，我們設計出一個包含四個子模型的「非監督式中文韻律標記及韻律模式」(Unsupervised joint prosody labeling and modeling, UJPLM)演算法，自動化地對語料同時進行韻律模式以及韻律標記，試圖更客觀且一致地標記出韻律標記。本研究標記的韻律標記為停頓標記及韻律狀態，其中停頓標記表示韻律單位的邊界，而韻律狀態的序列代表上層韻律單位(韻律詞、韻律短語以及呼吸組/韻律句組)的音高變化。實驗語料由一位專業女播音員朗讀中文文稿，文稿內容則從「中央研究院詞庫小組—中文句結構樹資料庫」中選出的短篇文字。透過分析訓練出的模型參數，我們探討此語者之：（1）音節的音高輪廓變化、韻律標記及語言參數的關係，（2）停頓標記、韻律參數及語言參數的關係，（3）由韻律狀態所表示的上層韻律單位之音高變化。藉由停頓標記和其對應詞關係之深入分析，除了探討韻律參數與語言參數的連結，同時也驗證本研究所提出方法之標記能力。另外，經由和人工停頓標記之比較，發現以本研究方法標記出來的停頓標記，其對應的韻律參數擁有較一致的統計特性，相較傳統以人工標記所造成的不一致統計特性，本研究的方法更能真實地（或客

觀地) 描述語者之韻律特性。基於UJPLM演算法，本研究接著提出「進階非監督式中文韻律標記及韻律模式」(Advanced-UJPLM, A-UJPLM)演算法，增加一個次要停頓韻律標記及同時對於音高、音長和音強進行模式建立。實驗結果顯示此方法可以更豐富地描述語者之韻律特性，停頓標記的結果顯示在主要停頓及無停頓的標示上，與UJPLM標示的結果相當一致，而A-UJPLM能夠標記出較多的次要停頓，使得次要停頓標記結果與人工標記結果更一致。最後本研究提出一個以A-UJPLM演算法為基礎之語音合成韻律產生法，實驗結果顯示此方法產生之韻律參數大致符合實際語音的韻律參數，驗證A-UJPLM演算法在韻律標記及韻律模式上擁有不錯的表現。



# Unsupervised Joint Prosody Labeling and Modeling for Mandarin Speech

Student: Chen-Yu Chiang

Advisors: Dr. Sin-Horng Chen  
Dr. Yih-Ru Wang

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

## Abstract

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



UJPLM (A-UJPLM) method was designed based on UJPLM to jointly label seven prosodic tags and model syllable pitch contour, duration and energy level. Experimental results showed that A-UJPLM performed quite well. The break labeling result showed that A-UJPLM inserted more minor breaks than UJPLM to result in a more consistent labeling of minor breaks to the human labeling. Lastly, an application of A-UJPLM to the prosody generation for Mandarin TTS is proposed. Experimental results showed that the proposed method performed well. Most predicted values of syllable pitch mean, duration and energy level matched well to their original counterparts. This also reconfirmed the effectiveness of the A-UJPLM method.



## 致謝

首先，最需要感謝影響我至深的指導教授：陳信宏老師及王逸如老師，在七年多的指導過程中，他們給予我許多參與國外學術交流和學習的機會，更感謝兩位老師平日生活上的諄諄教誨與不吝分享，讓我從中獲益良多。此外，也謝謝余秀敏老師及潘荷仙老師的大力協助，使本研究的內容更臻於完善，並且開拓我的研究視野。同時，亦需要感謝廣瀨 啓吉老師 (Professor Keikichi Hirose) 和廖元甫老師，於為期三個月在東京大學的訪問研究時，給予我許多磨鍊的機會，使我獲得更多成長。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、鄭秋豫老師五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。

回首在實驗室的點滴，總是一片歡愉的氣氛；感謝羅文輝學長、郭威志學長以及賴玟杏學姊，在研究與生活上的提攜與指引，常不厭其煩地讓我請教；也感謝智合、阿德、希群和巴金，大家不只在研究上相互鼓勵與切磋，還讓你們忍受我多年來持續的冷笑話；還要謝謝振豐、小傅、銘彥、友駿、宏宇、胤賢、小鄧、啟風、小迷彩、小廣、阿宅、柯達、普烏、小宋、杜Q、小帥哥等這群一起在實驗室同甘共苦的可愛學弟們，有了你們讓生活更加精采！

最後，特別感謝一直支持我的家人，謝謝爸媽從小的養育和栽培，沒有你們就無法有今日的我；還有宇君，有妳一路相伴，讓我在生活、研究的路上並不孤單，給予我最溫暖、窩心的守護。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

中文摘要 .....	i
Abstract.....	iii
致謝 .....	v
Contents .....	vi
List of Tables .....	ix
List of Figures.....	xi
<b>Chapter 1 Introduction .....</b>	<b>1</b>
1.1 Background.....	1
1.2 Motivation .....	4
1.3 Overview of Unsupervised Joint Prosody Labeling and Modeling.....	5
1.3.1 Previous Works .....	5
1.3.2 Prosody Hierarchy and Prosody Tags .....	9
1.3.3 The Four Prosodic Models .....	11
1.3.4 Experimental Database.....	12
1.4 Organization of the Dissertation.....	13
<b>Chapter 2 Unsupervised Joint Prosody Labeling and Modeling .....</b>	<b>14</b>
2.1 Introduction .....	14
2.2 The Design of the Four Models .....	14
2.3 Joint Prosody Labeling and Modeling.....	19
2.3.1 Initialization .....	19
2.3.2 Iteration .....	21
2.4 Experimental Results.....	22
2.4.1 The Syllable Pitch Contour Model .....	22
2.4.2 The Break-Acoustics Model .....	26

2.4.3 The Prosodic State Model.....	28
2.4.4 The Break-Syntax Model.....	29
2.5 Analyses of the Labeled Breaks and Prosodic Constituents .....	32
2.5.1 Analyses of the Labeled Break Types .....	32
2.5.2 Analyses of Prosodic Constituents .....	40
2.5.3 Pitch Patterns of Prosodic Constituents.....	42
2.5.4 Comparison with Human Labeling.....	44
2.5.5 A Labeling Example .....	49
2.6 Conclusions .....	51

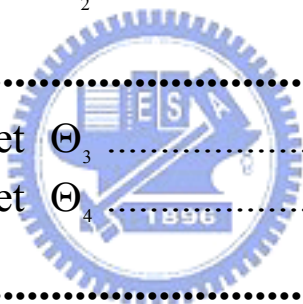
## **Chapter 3 Advanced Unsupervised Joint Prosody Labeling and Modeling.....52**

3.1 Introduction .....	52
3.2 The New Prosodic Model.....	52
3.2.1 Features and Parameters Used in the New Prosodic Model ..	52
3.2.2 Design of the New Prosodic Model.....	54
3.3 Model Training by the A-UJPLM Method.....	56
3.3.1 Initialization.....	57
3.3.2 Iteration .....	58
3.4 Experimental Results.....	59
3.3.1 The Syllable Prosodic Model.....	60
3.3.2 The Break-Acoustics Model .....	63
3.3.3 The Prosodic State Model.....	64
3.3.4 The Break-Syntax Model.....	66
3.4 Analyses of the Labeled Breaks and Prosodic Constituents .....	69
3.4.1 Comparison Between A-UJPLM and UJPLM.....	69
3.4.2 Patterns of Prosodic Constituents .....	71
3.4.3 A Labeling Example .....	75
3.5 Conclusions .....	77

## **Chapter 4 An Application to Prosody Generation for TTS.....78**

4.1 Introduction .....	78
4.2 The Proposed Break Prediction Method.....	80
4.2.1 Linguistic Features.....	80
4.2.2 Prediction Methods .....	82

4.3 Prosodic Feature Prediction.....	87
4.4 Conclusions .....	91
<b>Chapter 5 Conclusions and Future Works.....</b>	<b>92</b>
5.1 Conclusions .....	92
5.2 Future Works .....	94
<b>Bibliography .....</b>	<b>95</b>
<b>Appendix A.....</b>	<b>102</b>
<b>Appendix B.....</b>	<b>104</b>
<b>Appendix C.....</b>	<b>105</b>
C.1 The question set $\Theta_1$ .....	105
C.2 The question set $\Theta_2$ .....	106
<b>Appendix D.....</b>	<b>109</b>
D.1 The question set $\Theta_3$ .....	109
D.2 The question set $\Theta_4$ .....	109
<b>Publication List .....</b>	<b>111</b>



## List of Tables

Table 1.1: The content of the Sinica Treebank corpus .....	13
Table 2.1: The APs (log- $F_0$ levels, $\beta_p(1)$ ) and the distribution ( $P(p)$ ) of the 16 prosodic states.....	23
Table 2.2: Statistics of break types labeled for 121 prefixes and 195 suffixes.....	33
Table 2.3: Statistics of break types labeled for the DE words.....	34
Table 2.4: Statistics of break types labeled for the word sets of Ng, Di, and T. ....	36
Table 2.5: Statistics of break types labeled for word set of VE.....	36
Table 2.6: Statistics of break types labeled for word sets of Caa and Cb.....	38
Table 2.7: Statistics of break types labeled for word sets of P07 and P21.....	40
Table 2.8: Statistics of three types of prosodic constituents. Value in parentheses denotes standard deviation.....	40
Table 2.9: Count of short PPh instances with respect to the existence of PM at their two endings.....	42
Table 2.10: Correlations between unsupervised and human labeled breaks .....	46
Table 2.11: Distances measuring the difference between two acoustic feature distributions that belong to different break indices labeled by the same method: (a) the proposed method, and (b) human labeling. Upper and lower triangular matrices represent KL2 distances for pause duration and normalized pitch jump, respectively.....	49
Table 3.1: The notations of prosodic tags, prosodic features and linguistic features..	54
Table 3.2: APs of five tones.....	60
Table 3.3: APs of prosodic states.....	62
Table 3.4: TREs of the prosodic modelings for syllable pitch contour, duration and energy level w.r.t. the use of different combinations of affecting factors. ....	63
Table 3.5: Cooccurrence matrix for the break types labeled by A-UJPLM and by UJPLM.....	69
Table 3.6: Statistics of three types of prosodic constituents. Value in parentheses	

denotes standard deviation.....	70
Table 3.7: Cooccurrence matrix of break tags labeled by A-UJPLM and human.....	71
Table 3.8: Total residual errors (TREs) w.r.t. the use of different combinations of affecting factors for pitch/duration/energy level modeling .....	75
Table 4.1: Summary of linguistic features used and their abbreviations.....	81
Table 4.2: The confusion matrix of the target and predicted break types (%) using the baseline all-in-one CART-based method for (a) the inside and (b) outside tests. ....	83
Table 4.3: The confusion matrix of the break prediction for the baseline method evaluated using 3 broad classes of break: (a) The inside and (b) outside tests. (NB: non-break, MiB: minor break, MB: major break) .....	83
Table 4.4: The confusion matrix of target and predicted reduced three classes break types using the two-stage approach: (a) inside test (b) outside test.....	85
Table 4.5: The confusion matrix of target and predicted seven break types using the two-stage approach: (a) inside test and (b) outside test .....	85
Table 4.6: The confusion matrix of target and predicted reduced three classes break types using the Markov model: (a) inside test (b) outside test .....	86
Table 4.7: TREs of the prosodic feature prediction results. ....	89
Table 4.8: TREs of the prosodic feature prediction using correct break labels.....	89
Table B.1: The contextual linguistic features considered in this study. Note that the notations of POS symbols follow Ref. [72]. ....	104

# List of Figures

Figure 1.1: A commonly agreed and used prosody hierarchy structure that consists of four layers, including, syllable layer (SYL), prosodic word layer (PW), prosodic phrase layer (PP), and intonation phrase (IP). (Note: this figure is excerpted and modified from Ref. [6]).....	2
Figure 1.2: A conceptual prosody hierarchy of Mandarin speech proposed by Tseng <i>et al.</i> in Ref. [8].....	9
Figure 2.1: The relationship of observed syllable pitch contour with its APs.....	17
Figure 2.2: The decision tree for initial break type labeling.....	20
Figure 2.3: The plot of total log-likelihood versus iteration number. ....	22
Figure 2.4: The APs of five tones .....	23
Figure 2.5: The (a) forward and (c) backward coarticulation patterns, $\beta_{B,tp}^f$ and $\beta_{B,tp}^b$ , for $B_0$ (point line), $B_1$ (solid line), and $B_4$ (dashed line); and the (b) onset and (d) offset patterns, $\beta_{B_b,t}^f$ and $\beta_{B_e,t}^b$ , for $B_b$ and $B_e$ . Here $tp = (i, j)$ and $t = i$ or $j$ .....	25
Figure 2.6: The <i>pdfs</i> of (a) pause duration and (b) energy-dip level for the root nodes of these 6 break types. Numbers in ( ) denote the mean values.....	26
Figure 2.7: The decision trees of the break-acoustics model for (a) $B_4$ , (b) $B_3$ , (c) $B_2-2$ , (d) $B_2-1$ , and (e) $B_1$ . The numbers in a bracket denote average pause duration in ms (left), energy-dip level in dB (middle) and sample count (right) of the associated node. Solid line indicates positive answer to the question and dashed line indicates negative answer.....	27
Figure 2.8: The most significant prosodic state transitions for (a) $B_0$ , $B_1$ and $B_2-1$ , and (b) $B_2-2$ , $B_3$ and $B_4$ . Here, the number in each node represents the index of the prosodic state. Note that bold and thin lines denote the primary and secondary state transitions, respectively. ....	28
Figure 2.9: The decision tree of the break-syntax model. The bar plot associated with a node denotes the distributions of these six break types ( $B_0$ , $B_1$ , $B_2-1$ , $B_2-2$ , $B_3$ , $B_4$ , from left to right) and the number is the total sample count of the node.....	30
Figure 2.10: The more detailed structures of sub-trees of (a) $T_3$ , (b) $T_4$ , (c) $T_5$ and (d) $T_6$ . Solid line indicates positive answer to the question and dashed line indicates negative answer. ....	31



Figure 2.11: Histograms of lengths for BG/PG, PPh, and PW.....	41
Figure 2.12: The log- $F_0$ patterns of (a) BG/PG, (b) PPh, and (c) PW. The special symbol “□” in (a) indicates the ending syllable of a log- $F_0$ pattern.....	44
Figure 2.13: The histograms of length of the prosodic constituents formed by (a) the human labelers and (b) the proposed methods. The numbers in () represent the average length of prosodic constituents. ....	46
Figure 2.14: The histograms of (a) pause duration (in sec) and (b) normalized pitch jump (in log- $F_0$ ) for syllable-juncture instances belonging to sub-groups with different break-index pairs labeled by the two methods.....	48
Figure 2.15: An example of the automatic prosody labeling. (a) Syntactic trees with prosodic tags: upper case $B$ and lower case $b$ for break-index labeled by our method and the human labeler, respectively; and (b) syllable log- $F_0$ means: observed (open circle) and prosodic state+global mean (close circle). Solid/dash/dot lines represent $B_3/B_2-1/B_2-2$ respectively. The utterance is “ <i>yi-ju</i> (according to) <i>xing-zheng-yuan</i> (the Executive Yuan) <i>zhu-ji-chu</i> (Directorate-General of Budget, Accounting and Statistics) <i>de</i> (DE) <i>tong-ji</i> (statistics), <i>shi-yue-fen</i> (October) <i>yi</i> (1st) <i>dao</i> (to) <i>er-shi-ri</i> (20th), <i>wo-guo</i> (our country) <i>chu-kou</i> (export) <i>ji</i> (and) <i>jin-kou</i> (import) <i>jin-e</i> (the amount of money) <i>bi-qi</i> (in comparison with) <i>qu-nian</i> (last year) <i>tong-qi</i> (the same period) <i>jun</i> (both) <i>you</i> (to have some) <i>zeng-jia</i> (increase).”.....	50
Figure 3.1: The decision tree for initial break type labeling.....	57
Figure 3.2: The plot of total log-likelihood versus iteration number. ....	59
Figure 3.3: Decision tree analysis of duration APs of base-syllable type. Number in () represents the average length (ms) of the APs in the leaf node. Solid line indicates positive answer to the question and dashed line indicates negative answer.....	61
Figure 3.4: Decision tree analysis of energy-level APs of final. Number in () represents the average energy level (dB) of the APs in the leaf node. Solid line indicates positive answer to the question and dashed line indicates negative answer. ....	61
Figure 3.5: Distributions of normalized prosodic features and the APs of prosodic states (vertical lines). ....	62
Figure 3.6: The <i>pdfs</i> of (a) pause duration, (b) energy-dip level for the root nodes, (c) normalized pitch jump, (d) normalized duration lengthening factor 1 and (e) normalized duration lengthening factor 2 of these seven break types. Numbers in () denote the mean values.....	64
Figure 3.7: The most significant pitch prosodic state transitions, $P(p_n   p_{n-1}, B_{n-1})$ , for each break types. Notice that the darker lines represent the more primary prosodic state transitions.....	65

Figure 3.8: The most significant duration prosodic state transitions, $P(q_n   q_{n-1}, B_{n-1})$ , for each break types. Notice that the darker lines represent the more primary prosodic state transitions. ....	65
Figure 3.9: The most significant energy prosodic state transitions, $P(q_n   q_{n-1}, B_{n-1})$ , for each break types. Notice that the darker lines represent the more primary prosodic state transitions. ....	66
Figure 3.10: The decision tree of the break-syntax model. The bar plot associated with a node denotes the distributions of these six break types ( $B_0, B_1, B_2-1, B_2-2, B_3, B_4$ , from left to right) and the number is the total sample count of the node.....	67
Figure 3.11: The more detailed structures of sub-trees of (a) $T_3$ , (b) $T_4$ , (c) $T_6$ and (d) $T_5$ . Solid line indicates positive answer to the question and dashed line indicates negative answer. ....	68
Figure 3.12: Histograms of lengths for BG/PG, PPh and PW. ....	70
Figure 3.13: The log- $F_0$ patterns of BG/PG, PPh and PW. ....	73
Figure 3.14: The duration patterns of BG/PG, PPh and PW. ....	73
Figure 3.15: The energy level patterns of BG/PG, PPh and PW. ....	74
Figure 3.16: An example of the automatic prosody labeling by A-UJPLM. Upper, middle and lower panels represent observed (open circle) and prosodic state+global mean (solid diamond) of syllable log- $F_0$ means, syllable duration and syllable energy level, respectively. The utterance is “ <i>yi-ju</i> (according to) <i>xing-zheng-yuan</i> (the Executive Yuan) <i>zhu-ji-chu</i> (Directorate-General of Budget, Accounting and Statistics) <i>de</i> (DE) <i>tong-ji</i> (statistics), <i>shi-yue-fen</i> (October) <i>yi</i> (1st) <i>dao</i> (to) <i>er-shi-ri</i> (20th), <i>wo-guo</i> (our country) <i>chu-kou</i> (export) <i>ji</i> (and) <i>jin-kou</i> (import) <i>jin-e</i> (the amount of money) <i>bi-qi</i> (in comparison with) <i>qu-nian</i> (last year) <i>tong-qi</i> (the same period) <i>jun</i> (both) <i>you</i> (to have some) <i>zeng-jia</i> (increase).” .....	76
Figure 4.1: The proposed prosody generation method. ....	79
Figure 4.2: All-in-one CART for break prediction.....	82
Figure 4.3: A block diagram of the two-stage break prediction method. ....	84
Figure 4.4: An example of the prosodic feature prediction by the A-UJPLM-based approach. The panels from up to bottom represent, respectively, syllable log- $F_0$ means, syllable duration, syllable energy level and inter-syllable pause duration. Solid lines, open circles, and closed circles denote, correspondingly, the original features, the predicted features using predicted breaks, and the predicted features using correct break labels. Vertical dash lines represent erroneous major/minor break prediction boundaries while vertical solid lines represent correct ones. Notice that	

break labels in () represent erroneous breaks predicted.....90



# Chapter 1 Introduction

## 1.1 Background

The term prosody refers to certain inherent suprasegmental properties that carry melodic, timing, and pragmatic information of continuous speech, encompassing accentuation, intonation, rhythm, speaking rate, prominences, pauses, and attitudes or emotions intended to express. Prosodic features are physically encoded in the variations in pitch contour, energy level, duration, and silence of spoken utterances. Prosodic studies have indicated that these prosodic features are not produced arbitrarily, but rather realized after a hierarchically organized structure which demarcates speech flows into domains of varying lengths by boundary or break cues such as pre- and post-boundary lengthening, pitch and energy change, pauses, etc. Therefore, prosodic structure in English, for example, functions to set up syntagmatic contrasts to mark a prosodic word, an intermediate phrase, or an intonational boundary [1-3]. On the other hand, the prosodic structure of Mandarin Chinese also parses continuous speech into different prosodic constituents by breaks that reflect different levels of Chinese linguistic processing: phonetic, lexical, syntactic, and pragmatic. As a result, successive words with related prosodic feature variations are aggregated to form prosodic phrases, and contiguous prosodic phrases are, in turn, integrated to form prosodic phrases of a higher level.

Many literatures on Chinese prosody have shown that the prosody of Mandarin speech can be organized into hierarchical structures [4-7]. Figure 1.1 displays a commonly agreed and used prosody hierarchy structure consists of four layers, including, from the lowest layer to the highest one, syllable layer, prosodic word layer, prosodic phrase layer (or intermediate phrase), and intonation phrase. As far as the major prosodic information relevant to each of the layers is concerned, given that Mandarin is a monosyllabic and tonal language, where each syllable with its inherent tone contains a lexical meaning, and each tone carries a lexically contrastive role, the features of every syllabic tone of an utterance are the most important prosodic information for the lowest layer; besides, tone along with syllable constituents affects syllable duration and energy level as well. As for the second prosodic layer, a

prosodic word refers to di-syllabic and multi-syllabic words or phrases composed of words syntactically and semantically closely related or most frequently collocated, so the words or phrases are uttered as a single unit as in 霧(wu) “fog” + 的(de) + 形成(xing-cheng) “to form” (the formation of fog). As for the third prosodic layer, prosodic phrase is composed of one or several prosodic words and it usually ends with a perceptible but unobvious break. Finally, intonation phrase is at the top layer of the Mandarin prosodic structure. It determines the pitch contour of the intonation of a sentence containing one or several prosodic phrases and it ends with an obvious break. Basically, the four-layer prosodic structure interprets the pitch and duration variations of syllable well for sentential utterances. Some recent studies [8,9] proposed to integrate prosodic phrases into prosodic phrase groups to interpret the contributions of higher-level discourse information to the wider-range and larger variations on the prosodic features of utterances of long texts. In the science of speech processing, to model prosody is to exploit a framework or a computational model to represent a hierarchy of prosodic phrases of speech and to describe its relationship with the syntactic structure of the associated text.



Figure 1.1: A commonly agreed and used prosody hierarchy structure that consists of four layers, including, syllable layer (SYL), prosodic word layer (PW), prosodic phrase layer (PP), and intonation phrase (IP). (Note: this figure is excerpted and modified from Ref. [6])

In the past, many prosody modeling methods have been proposed for various applications, including generation of prosodic information for text-to-speech (TTS) [10-12], segmentation of untranscribed speech into sentences or topics [13-15], generation of punctuations from speech [16-18], detection of interrupt points in spontaneous speech [13,19-21], automatic speech recognition (ASR) [22-28], and so forth. It can be found from those prosody modeling studies that four main issues have been intensively addressed. The first one is concerning representing a hierarchical prosodic phrase structure indirectly by tags marking important prosodic events.

Among various prosodic events explored in the relevant literature [29-35], break type and tone pattern are the most important ones: the break types of all word boundaries can determine the hierarchical prosodic phrase structure of an utterance, and the tonal patterns of all syllables/words can indicate the accented syllables/words of an utterance, and may specify the pitch contour patterns of the prosodic constituents. Several prosody representation systems have been proposed in the past. They include ToBI (Tones and Breaks Indices, a standard prosody transcription system for American English utterances) [29], PROSPA [31], INTSINT [32], and TILT [33]. Among them, ToBI and its modifications to other languages, such as Pan-Mandarin ToBI [34] and C-ToBI [35], are most popular conventions for Mandarin Chinese prosodic tagging. The second main issue is about realizing the constituents of a hierarchical prosodic phrase structure by using prosodic feature patterns. This is mainly used in TTS for the generation of prosodic information from prosodic tags. A common approach is to use a multi-component representation model to superimpose several prototypical contours of multi-level prosodic phrases for each prosodic feature [36-38]. In Ref. [36], three components of sentence-specific contours, word-specific contours, and tone-specific contours are superimposed to form the synthesized contours of pitch and syllable duration for Mandarin TTS.

The third main issue is relating to exploring the relationship between prosodic tags (or boundary types) and the acoustic features surrounding the associated word juncture. Patterns of pause duration, pitch, and energy around word junctures are modeled for each prosodic tag or boundary type to help speech segmentation [13-15], topic identification [15], punctuation generation [16-18], interrupt point detection [13,19-21], and ASR [22-28] based on word-based features. The last issue is upon modeling the relationship between prosodic structure and syntactic structure. It is known that prosodic structure is closely related to syntactic structure although they are not identical. Usually, only the relationship between a prosodic tag, such as break or prominence, and contextual linguistic features of syntactic structure is built. A good break-syntax model should be very useful in predicting breaks of various levels from input text for TTS. Main methods of building a break-syntax model for TTS are hierarchical stochastic model [39,40], N-gram model [41], classification and regression tree (CART) [40,42-45], Markov model [46], artificial neural networks [47], maximum entropy model [48-51], etc. In the popular Markov model-based

approach, emission probabilities can be generated by CART [44] or maximum entropy model [51].

## 1.2 Motivation

Prosody modeling has been proved to be useful in above-mentioned applications, and the most commonly adopted approach by the previous studies is a supervised one to construct prosodic model from an annotated speech database with tags marking prosodic events being pre-labeled manually. However, the supervised prosody modeling based on human labeling unavoidably arises such problems as diseconomy due to labeler training and manual labeling labor, and inter-labelers' and intra-labeler's inconsistency caused by individual subjectivity and fatigue during long time labeling, respectively. This inconsistency may mislead prosody modeling to obtain erroneous results, and hence lead to unwanted degradation of modeling performance. Even in the studies where prosody labeling can be automatically done by machine, their model is still trained with a manually-annotated speech corpus [52-57], so the performance of machine labeling is still subject to the quality of human prosody labeling.

To tackle the problems arising from the supervised prosody modeling with manual labeling, this dissertation presents a new unsupervised approach of prosody modeling to jointly perform prosody modeling and labeling for Mandarin speech based on an unlabeled speech database. It is an extension of the previous work by Chen *et al.* [58,59] which will be introduced in Section 1.3. The basic idea of this work is to properly model the observed features and then let the modeled-features objectively determine prosodic tags by themselves rather than by human perception with audio and visual aid in conventional prosody labeling works. The task automatically determines two types of prosodic tags for all utterances of a corpus and to build four prosodic models simultaneously. The two types of prosodic tags are: (1) the break types of inter-syllable locations (or syllable junctures) which can be used to demarcate the constituents of a hierarchy of Mandarin speech prosody; and (2) the prosodic states of syllables which can be used to construct the pitch contour, syllable duration and energy level patterns of the prosodic constituents. The four prosodic models are introduced to describe the various relationships between the two types of prosodic tags and all available information sources including acoustic prosodic

features and syntactic structure features. Three advantages of the proposed method can be found. First, prosody modeling and labeling are accomplished jointly and automatically without using human-labeled training corpus. Second, all information sources, including acoustic and linguistic features, are systematically used (via introducing the four prosodic models) in the prosody labeling. We therefore expect that the result of the prosodic labeling is more consistent than that done by human, which will in turn make the four prosodic models more accurate. Third, the four prosodic models constructed address all the four main issues of prosody modeling discussed above. So they are useful models and may be directly used or extended to be used in those applications mentioned above.

## **1.3 Overview of Unsupervised Joint Prosody Labeling and Modeling**

Since the proposed unsupervised joint prosody labeling and modeling method is an extension of the previous works by Chen *et al.* [58,59], these previous works will be introduced in Subsection 1.3.1 to give a clearer concept of the prosodic states defined. Then the prosody hierarchy adopted in this study and its relationship to the defined prosody tags, i.e. break types and prosodic states, are described in Subsection 1.3.2. In Subsection 1.3.3, we present the general concept of the four prosodic models which are the core of this dissertation. Lastly, the database used in our experiments is introduced in Subsection 1.3.4.

### **1.3.1 Previous Works**

Two statistical prosody models for Mandarin speech using unlabeled speech corpora were proposed by Chen *et al.* [58,59]. These two models consider several affecting factors on the variations in syllable pitch contour and syllable duration, respectively, including lexical tones, initial-final or base-syllable type, and prosodic state. Here, prosodic state is conceptually defined as the state in a prosodic phrase and used as a substitution for the effects from high-level linguistic features, such as a word, a phrase or a syntactic tree. Prosodic states are also assumed to account for the prosodic variation contributed by para-linguistic features, such as intention, attitude and style of the speaker, and even by non-linguistic features, such as physical and



emotional conditions of the speaker. They therefore treat the high-level linguistic features, para-linguistic features and non-linguistic features as high-level affecting factors on prosodic variation. On the other hands, low-level affecting factors refer to some syllable-level linguistic features which represent intrinsic characteristics of Mandarin prosody, such as lexical tones and base-syllable type, and so forth. For the syllable duration model, a companding factor (CF) is hence defined to control the compression/increase or stretch/increase of syllable duration/pitch associated with each of the above low- and high-level affecting factors. Based on the assumption that all CFs are combined multiplicatively or additively, the multiplicative and additive syllable duration models are expressed respectively by

$$Z_n = X_n \gamma_{t_n} \gamma_{y_n} \gamma_{j_n} \gamma_{l_n} \gamma_{s_n} \quad (1.1)$$

and

$$Z_n = X_n + \gamma_{t_n} + \gamma_{y_n} + \gamma_{j_n} + \gamma_{l_n} + \gamma_{s_n} \quad (1.2)$$

where  $Z_n$  and  $X_n$  are the observed and normalized durations of the  $n$ -th syllable;  $\gamma_x$  represents duration CF of the affecting factors  $x$ ;  $t_n$ ,  $y_n$ ,  $j_n$ ,  $l_n$  and  $s_n$  respectively represent the lexical tone, duration prosodic state, base-syllable type, utterance, and speaker of the  $n$ -th syllable; and the residual  $X_n$  is modeled by a normal distribution with mean  $\mu$  and variance  $v$ , i.e.  $p(Z_n, y_n | \lambda) = N(Z_n; \mu \gamma_{t_n} \gamma_{y_n} \gamma_{j_n} \gamma_{l_n} \gamma_{s_n}, v \gamma_{t_n}^2 \gamma_{y_n}^2 \gamma_{j_n}^2 \gamma_{l_n}^2 \gamma_{s_n}^2)$ . Notice that the prosodic state is treated as a latent variable hence the Expectation-Maximization (EM) algorithm is introduced to train the multiplicative or additive syllable duration models based on the maximum likelihood (ML) criterion. After training, each syllable can be labeled a prosodic state index by

$$y_n^* = \max_{y_n} p(y_n | Z_n, \lambda) \quad (1.3)$$

Then, the CF sequence of prosodic state  $\{\gamma_{y_n^*}\}$  of each utterance can represent the syllable-duration variation of the utterance primarily resulted from high-level linguistic features.

Based on the same idea, the syllable pitch mean and shape models are respectively expressed by

$$Y_n = X_n + \beta_{i_n} + \beta_{pt_n} + \beta_{ft_n} + \beta_{i_n} + \beta_{f_n} + \beta_{p_n} \quad (1.4)$$

and

$$\mathbf{Z}_n = \mathbf{X}_n + \mathbf{b}_{tc_n} + \mathbf{b}_{q_n} + \mathbf{b}_{s_n} + \mathbf{b}_{i_n} + \mathbf{b}_{f_n} \quad (1.5)$$

where  $Y_n/\mathbf{Z}_n$  and  $X_n/\mathbf{X}_n$  are observed and normalized pitch mean/shape of the  $n$ -th syllable;  $\beta_x/\mathbf{b}$  represents the pitch mean/shape CF of the affecting factors  $x$ ;  $pt_n$ ,  $ft_n$ ,  $i_n$ ,  $f_n$ ,  $p_n$ ,  $tc_n$  and  $q_n$  are, correspondingly, previous lexical tone, following lexical tone, initial type, final type, pitch mean prosodic state, tone combination and pitch shape prosodic state of the  $n$ -th syllable. The training and labeling of the pitch mean/shape models are similar to the way in syllable duration modeling.

The main purpose of using prosodic state to replace conventional high-level linguistic information is to decompose the effects of low-level and high-level linguistic features on speech prosody. Through this modeling approach, some unsolved problems, such as the inconsistency between prosodic and syntactic structures, the ambiguity of word segmentation and word chunking for Mandarin Chinese, can be avoided. Hence, this modeling scheme can more focus on modeling the global effect of mapping high-level linguistic features to the prosodic state and break indices, since interference caused by low-level linguistic feature has been properly removed. The following are some key observations and conclusions of the proposed models evaluated by a speech corpus consisted of paragraphic utterance of five speakers:

1. The variances of syllable duration, pitch mean and pitch shape were greatly reduced as the observed prosodic features are normalized with the CFs for considered affecting factors.
2. The quantitative influence of each affecting factor is directly obtained from their corresponding CFs.

3. The obtained CFs for low-level linguistic features generally agreed with the prior knowledge of Mandarin prosody.
4. By investigating the relationship between the labeled prosodic states and their associated texts, the prosodic states labeled seemed to be linguistically meaningful. For example, the prosodic states with larger CFs are usually labeled on the last syllable of a sentence illustrating syllable duration lengthening effect; pattern of a prosodic phrase is more apparent when it is represented by a sequence of pitch mean prosodic state CFs than when observed in original pitch mean.
5. Experiments on prosody generation for Mandarin TTS system showed that the hybrid-regression model that normalized the observed prosodic features with CFs for syllable level linguistic features in advanced achieved a better prediction result than a conventional regression method that take observed prosodic features and all levels of linguistic features as targets and inputs. The results implied the proposed models can properly decompose the influences of high-level and low-level linguistic features on prosody.
6. A simple rule-based break labeling method is proposed. Large and medium sudden low-to-high pitch prosodic state transitions indicated minor and major breaks boundaries.

As discussed above, the two models proposed by Chen *et al.* could generate linguistically meaningful prosodic state tags and they can give a better representation of prosodic phrase patterns. However, a well-defined prosody hierarchy is not considered in these previous studies. Besides, the relationship between prosodic states and high-level linguistic features are still untouched. Some important acoustic features related to prosodic breaks are also not incorporated in those models. We therefore intend the new proposed unsupervised joint prosody labeling and modeling method in this dissertation to address those missing research fields mentioned above based on the previous works by Chen *et al.*

### 1.3.2 Prosody Hierarchy and Prosody Tags

Recently, Tseng *et al.* [8] proposed to integrate contiguous prosodic phrases into prosodic phrase groups to interpret the contributions of higher-level discourse information to the wider-range and larger variations in syllable pitch and duration of long utterances in paragraphs. Figure 1.2 displays the hierarchical prosodic phrase grouping (HPG) model of Mandarin speech proposed by Tseng. It is a five-layer structure. The first three layers in the hierarchy are the same as those of the four-layer prosodic structure introduced in Section 1.1, which are referred to as Syllable (SYL), Prosodic Word (PW), and Prosodic Phrase (PPh) in the system of Tseng *et al.*, respectively. The fourth layer, Breath Group (BG), is formed by combining a sequence of PPhs, and a sequence of BGs, in turn, constitutes the fifth layer, Prosodic Phrase Group (PG). The above five prosodic units are delimited by different type of the six breaks proposed by Tseng *et al.* Firstly,  $B_0$  and  $B_1$  are defined for SYL boundaries within PW. Here,  $B_0$  represents reduced syllabic boundary and  $B_1$  represents normal syllabic boundary. Usually no identifiable pauses exist for both  $B_0$  and  $B_1$ . Secondly,  $B_4$  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. Thirdly,  $B_2$  and  $B_3$  are perceivable boundaries defined for PW and PPh boundaries, respectively.

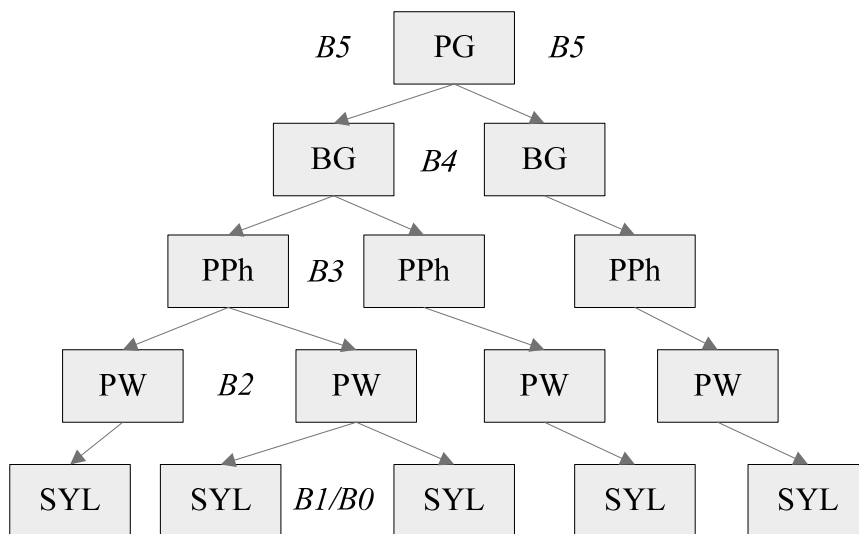


Figure 1.2: A conceptual prosody hierarchy of Mandarin speech proposed by Tseng *et al.* in Ref. [8].

In this dissertation, we adopt the prosodic structure of Tseng *et al.* because our speech database also consists of long Mandarin utterances of paragraphs. However, we modify the break type labeling scheme of HPG model by dividing  $B2$  into two types,  $B2-1$  and  $B2-2$ , and combining  $B4$  and  $B5$  into one denoted simply by  $B4$ . Here,  $B2-2$  represents syllabic boundary of  $B2$  perceived by pause, while  $B2-1$  is  $B2$  with  $F0$  movement. The reason of dividing  $B2$  into  $B2-1$  and  $B2-2$  is due to the difference of their acoustic cues to be modeled. On the contrary, the combination of  $B4$  and  $B5$  is owing to the similarity of their acoustic characteristics. So, the break-type tags used is in  $\Lambda = \{B0, B1, B2-1, B2-2, B3, B4\}$ . These six break-type tags can be used to delimit four types of prosodic units: SYL, PW, PPh, and BG/PG. These four units are the constituents of our hierarchical prosodic structure.

To further specify the four-layer prosodic structure, a representation of its constituents using prosodic features is needed. Two main approaches of representation can be considered. One is direct representation approach to represent each individual prosodic constituent by multiple prototypical patterns for each prosodic feature of syllable pitch contour, duration, or energy level [8,10,11]. The other is indirect representation approach by using some tags which carry the information of prosodic constituents and are treated as hidden, i.e. the prosodic states. Due to the following two reasons, we do not adopt direct representation approach in the prosody modeling and labeling study. First, the technique of direct representation approach is still not mature enough to produce a good direct representation for the hierarchy of Mandarin speech prosody. The modeling errors, defined as the ratio of mean square errors of direct representations to the variances of the raw data, are still as high as about 30% for the multi-layer representations of syllable duration, and energy using the HPG model [8]. Second, a good direct representation is not easy to be realized for the case of joint prosody modeling and labeling using an unlabeled speech corpus in which the prosodic structures of all utterances are not well determined in advance. Degeneration may occur because break labeling errors may produce inaccurate representation patterns of prosodic constituents, which in turn may cause more break labeling errors to occur. Instead, we adopt an indirect representation approach to employ the defined prosodic states discussed in Subsection 1.3.1 to represent the aggregative contributions of the constituents of the upper three layers on syllable pitch level. Similar to the definition of the previous works by Chen *et al.*, the

prosodic state tag is defined as a quantized and normalized syllable pitch level, duration and energy level with the effects from base-syllable or final types, the current tone and the two nearest neighboring tones being properly eliminated. So it carries mainly the prosodic information of the upper three layers of the prosodic structure, i.e. PW, PPh, and BG/PG. We call it prosodic state to roughly mean the state in a prosodic phrase (PW, PPh, or BG/PG). Two advantages of using the prosodic-state tag can be found. First, the tag is defined for each individual syllable so that the effect of a labeling error is limited to the current syllable only. No degeneration in the joint prosody modeling and labeling process will occur. Second, the tag carries the full prosodic information in the upper three layers of the prosodic structure. In experimental results, we will show the capability of the prosodic-state tag on constructing the pitch contour patterns of PW, PPh, and BG/PG. It is worth to note that the pitch prosodic state, duration prosodic state and energy prosodic state are correspondingly defined for syllable pitch mean, syllable duration, and syllable energy level variations of high-level prosodic constituents ( i.e. PW, PPh, and PG/BG). In this dissertation, for simplicity, we first only consider the pitch prosodic state in unsupervised joint prosody labeling and modeling. Then the duration prosodic state and the energy prosodic state are added to the proposed method to perform an advanced unsupervised joint prosody labeling and modeling.

### **1.3.3 The Four Prosodic Models**

The four prosodic models are designed to model the prosody hierarchy illustrated in Subsection 1.3.2 and perform unsupervised joint prosody labeling and modeling given with both acoustic and linguistic features. Two types of acoustic features can be considered. One is the prosodic features which carry the information of prosodic constituents. Primary features of this type include syllable pitch contour, syllable duration, and syllable energy level. Another is the acoustic features used to specify the break type of syllable juncture. Primary features of this type include pause duration and energy-dip level of syllable juncture, energy and pitch jumps across syllable juncture, lengthening factor of syllable duration, etc. The linguistic features used span a wide range from syllable level, word level to syntactic tree level.

The first model, referred to as the syllable prosodic model, describes the

variations in syllable prosodic features, including syllable pitch contours, duration and energy level, controlled by several major affecting factors, such as syllable-level linguistic features and prosodic tags. The next one, referred to as the break-acoustics model, describes the relationship between the break type of a syllable juncture and nearby acoustic features, such as pause duration and energy-dip level of syllable juncture. The third one describes the relationship between the break type of a syllable juncture and contextual linguistic features. It is referred to as the break-syntax model. Finally, the last model describes the relationship between the prosodic states of syllables and the break types of neighboring syllable junctures, and is referred to as the prosodic state model. We can then regard the proposed unsupervised joint prosody labeling and modeling method which is based on the four prosodic models as a clustering problem. With proper initializations of break types and prosodic states, a designed sequential optimization training algorithm is conducted to iteratively estimate parameters of the four prosodic models, and find all prosodic tags using an unlabeled speech corpus.

### 1.3.4 Experimental Database

An unlabeled read Mandarin speech database was used to evaluate the proposed unsupervised joint prosody labeling and modeling method. The database contained 425 utterances with 56237 syllables uttered by a female professional announcer in a sound-proof booth. All speech signals were digitally recorded in a form of 16kHz sampling rate and 16-bit resolution. Its associated texts were all short paragraphs composed of several sentences selected from the Sinica Treebank Version 3.0 [60]. There are six files in the Sinica Treebank Version 3.0 as listed in Table 1.1. All the texts used in this study were extracted from the “news.check” file. Those texts were automatically parsed and manually checked. The tone and base-syllable type of each syllable were transcribed by a linguistic processor with a 130,000-word lexicon and then manually error-corrected. All syllable segmentation and F0 detection were first done automatically using the Hidden Markov Model Toolkit (HTK) [61] and WaveSurfer [62], respectively, and then error corrected manually. The database is further divided into two parts: a training set of 379 utterances with 52192 syllables and a test set of 46 utterances with 4801 syllables.

Table 1.1: The content of the Sinica Treebank corpus

File name	Content
news.check, travel.check	News papers, books, or internet articles
ko.check, ev.check	Elementary school text books
oral.check	Text from phonetic balanced speech
sino.check	Text from Taiwan Panorama

## 1.4 Organization of the Dissertation

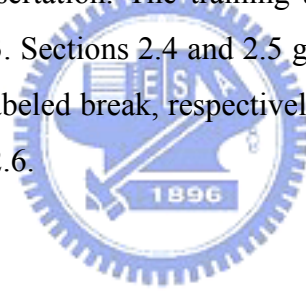
The dissertation is organized as follows. Chapter 1 gives the background, motivation, review of related previous works and description on the experimental databases used in this study. Chapter 2 presents the proposed unsupervised joint prosody labeling and modeling method which employs four prosodic models to describe relationship between prosodic tags, acoustic features and associated linguistic features. For simplicity we only consider the modeling of syllable pitch contour in this chapter and will extend the modeling to include the other two syllable prosodic features, i.e. syllable duration and syllable energy level, in Chapter 3. The experimental results on the training set of the Sinica Treebank corpus are discussed. Then, an extension to joint modeling of syllable pitch contour, duration and energy level is presented in Chapter 3. An application of the proposed model to prosody generation for TTS is discussed in Chapter 4. Some conclusions and related future research topics are given in the last chapter.



# Chapter 2 Unsupervised Joint Prosody Labeling and Modeling

## 2.1 Introduction

The proposed method first treats the problem as a model-based prosody labeling problem to define four prosodic models to describe various relationships between the prosodic tags to be labeled and the available information sources of acoustic and syntactic features. It then extends the formulation for the joint prosody labeling and modeling problem and applies a sequential optimization procedure to jointly label prosodic tags and estimate the model parameters using an unlabeled speech corpus. This chapter is organized as follows. Section 2.2 presents the four prosodic models which is the core of this dissertation. The training algorithm for the four prosodic models is given in Section 2.3. Sections 2.4 and 2.5 give the experimental results and the detail analysis of model-labeled break, respectively. Lastly, some conclusions and remarks are given in Section 2.6.



## 2.2 The Design of the Four Models

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

$$\mathbf{T}^* = \arg\max_{\mathbf{T}} P(\mathbf{T}|\mathbf{A}, \mathbf{L}) = \arg\max_{\mathbf{T}} P(\mathbf{T}, \mathbf{A}|\mathbf{L}) \quad (2.1)$$

Two types of prosodic tags which carry the information of prosodic structure of Mandarin speech are considered in this study. One is the break type of syllable juncture. A set of six break types, defined in Subsection 1.3.2, is used. It is denoted as  $\{B0, B1, B2-1, B2-2, B3, B4\}$ . These six break types are used to define a hierarchy of speech prosody comprising four constituents of SYL, PW, PPh, and BG/PG. Another is the prosodic state of syllable defined as a quantized and normalized syllable pitch level with the effects of the current tone and the two nearest neighboring tones being

properly eliminated. As discussed in Subsection 1.3.2, it is an indirect representation of the prosodic constituents to carry the pitch level information of PW, PPh, and BG/PG. So, **T** can be refined to comprise a break-type sequence **B** and a prosodic state sequence **p**.

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

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

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

$$P(\mathbf{T}, \mathbf{A} | \mathbf{L}) = P(\mathbf{B}, \mathbf{p}, \mathbf{sp}, \mathbf{pd}, \mathbf{ed} | \mathbf{l}, \mathbf{t}) = P(\mathbf{sp}, \mathbf{pd}, \mathbf{ed} | \mathbf{B}, \mathbf{p}, \mathbf{l}, \mathbf{t}) P(\mathbf{B}, \mathbf{p} | \mathbf{l}, \mathbf{t}) \quad (2.2)$$

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

Since the break type tag sequence,  $\mathbf{B}$ , has already carried the prosodic cues related to syllable junctures, we therefore assume that the observed syllable-based acoustic feature,  $\mathbf{sp}$ , and the juncture-based acoustic features, ( $\mathbf{pd}, \mathbf{ed}$ ), are independent as  $\mathbf{B}$  is given. So we split  $P(\mathbf{sp}, \mathbf{pd}, \mathbf{ed} | \mathbf{B}, \mathbf{p}, \mathbf{l}, \mathbf{t})$  into two terms:

$$P(\mathbf{sp}, \mathbf{pd}, \mathbf{ed} | \mathbf{B}, \mathbf{p}, \mathbf{l}, \mathbf{t}) \approx P(\mathbf{sp} | \mathbf{B}, \mathbf{p}, \mathbf{l}, \mathbf{t}) P(\mathbf{pd}, \mathbf{ed} | \mathbf{B}, \mathbf{p}, \mathbf{l}, \mathbf{t}) \quad (2.3)$$

Here  $P(\mathbf{sp} | \mathbf{B}, \mathbf{p}, \mathbf{l}, \mathbf{t})$  is a syllable pitch contour model describing the variation in syllable pitch contour controlled by ( $\mathbf{B}, \mathbf{p}, \mathbf{l}, \mathbf{t}$ ) and  $P(\mathbf{pd}, \mathbf{ed} | \mathbf{B}, \mathbf{p}, \mathbf{l}, \mathbf{t})$  is a break-acoustics model describing the acoustic cues of syllable junctures for different break types. The syllable pitch contour model is realized using a modified version of the syllable pitch contour model proposed previously by Chen *et al* [58]. It models the pitch contour of each syllable separately and considers four main affecting factors, including the current prosodic state  $p_n$ , the current tone  $t_n$ , and the coarticulations from the two nearest neighboring tones,  $t_{n-1}$  and  $t_{n+1}$ , conditioned, respectively, on the break types,  $B_{n-1}$  and  $B_n$ , of the syllable junctures on both sides. Specifically, the model is expressed by

$$P(\mathbf{sp} | \mathbf{B}, \mathbf{p}, \mathbf{l}, \mathbf{t}) \approx P(\mathbf{sp} | \mathbf{B}, \mathbf{p}, \mathbf{t}) \approx \prod_{n=1}^N P(\mathbf{sp}_n | p_n, B_{n-1}, t_{n-1}^{n+1}) \quad (2.4)$$

where

$$\mathbf{sp}_n = \mathbf{sp}_n^r + \boldsymbol{\beta}_{t_n} + \boldsymbol{\beta}_{p_n} + \boldsymbol{\beta}_{B_{n-1}, t_{n-1}}^f + \boldsymbol{\beta}_{B_n, p_n}^b + \boldsymbol{\mu} \quad \text{for } 1 \leq n \leq N \quad (2.5)$$

is the observed pitch contour of  $n$ -th syllable (referred to as *syllable*  $n$  hereafter) represented by the first four orthogonally-transformed parameters of syllable log- $F_0$  contour [63];  $B_{n-1}^n = (B_{n-1}, B_n)$ ;  $t_{n-1}^{n+1} = (t_{n-1}, t_n, t_{n+1})$ ;  $\mathbf{sp}_n^r$  is the normalized (or residual) version of  $\mathbf{sp}_n$ ;  $\boldsymbol{\beta}_x$  represents the affecting pattern (AP) of affecting factor  $x$ . Here AP means the effect of a factor on increase or decrease of the observed syllable pitch contour vector  $\mathbf{sp}_n$ .  $\boldsymbol{\beta}_{t_n}$  and  $\boldsymbol{\beta}_{p_n}$  are the APs of affecting factors  $t_n$  and  $p_n$ ,

respectively;  $tp_n$  is tone pair  $t_n^{n+1}=(t_n, t_{n+1})$ ;  $\beta_{B_{n-1}, tp_{n-1}}^f$  and  $\beta_{B_n, tp_n}^b$  are the APs of forward and backward coarticulation contributed from *syllable n-1* and *syllable n+1*, respectively; and  $\mu$  is the AP of global mean. For taking care of utterance boundaries, two special break types,  $B_b$  and  $B_e$ , are assigned to the two ending locations of all utterances, i.e.,  $B_0 = B_b$  and  $B_N = B_e$ ; and two special APs of coarticulation,  $\beta_{B_b, t_1}^f = \beta_{B_0, tp_0}^f$  and  $\beta_{B_e, t_N}^b = \beta_{B_N, tp_N}^b$ , are accordingly adopted to represent the effects of utterance onset and offset, respectively.  $\beta_{p_n}$  is set to have nonzero value only in its first dimension in order to restrict the influence of prosodic state merely on the log- $F0$  level of the current syllable. Figure 2.1 displays the relationship of  $\mathbf{sp}_n$  with these affecting factors. By assuming that  $\mathbf{sp}_n^r$  is zero-mean and normally distributed, i.e.  $N(\mathbf{sp}_n^r; \mathbf{0}, \mathbf{R})$ , we have

$$P(\mathbf{sp}_n | p_n, B_{n-1}^n, t_{n-1}^{n+1}) = N(\mathbf{sp}_n; \beta_{t_n} + \beta_{p_n} + \beta_{B_{n-1}, tp_{n-1}}^f + \beta_{B_n, tp_n}^b + \mu, \mathbf{R}) \quad \text{for } 1 \leq n \leq N \quad (2.6)$$

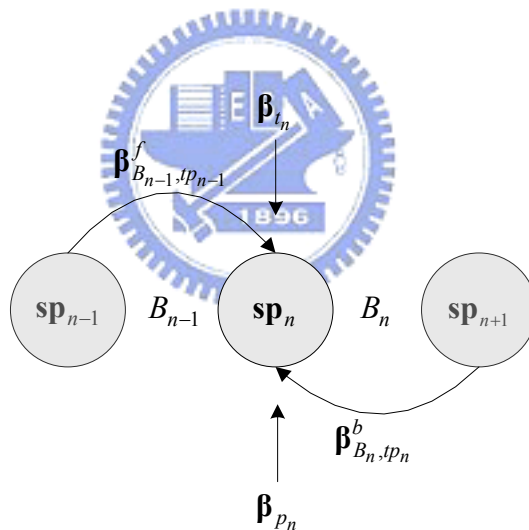


Figure 2.1: The relationship of observed syllable pitch contour with its APs.

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

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

$$P(\mathbf{pd}, \mathbf{ed} | \mathbf{B}, \mathbf{p}, \mathbf{l}, \mathbf{t}) \approx P(\mathbf{pd}, \mathbf{ed} | \mathbf{B}, \mathbf{l}) \approx \prod_{n=1}^{N-1} P(pd_n, ed_n | B_n, \mathbf{l}_n) \quad (2.7)$$

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

$$P(pd_n, ed_n | B_n, \mathbf{l}_n) = g(pd_n; \alpha_{B_n, \mathbf{l}_n}, \beta_{B_n, \mathbf{l}_n}) N(ed_n; \mu_{B_n, \mathbf{l}_n}, \sigma_{B_n, \mathbf{l}_n}^2) \quad (2.8)$$

In this study,  $g(pd_n; \alpha_{B_n, \mathbf{l}_n}, \beta_{B_n, \mathbf{l}_n})$  and  $N(ed_n; \mu_{B_n, \mathbf{l}_n}, \sigma_{B_n, \mathbf{l}_n}^2)$  are concurrently generated by the decision tree method for each break type.

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

$$P(\mathbf{B}, \mathbf{p} | \mathbf{l}, \mathbf{t}) \approx P(\mathbf{B}, \mathbf{p} | \mathbf{l}) = P(\mathbf{p} | \mathbf{B}, \mathbf{l}) P(\mathbf{B} | \mathbf{l}) \approx P(\mathbf{p} | \mathbf{B}) P(\mathbf{B} | \mathbf{l}) \quad (2.9)$$

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

$$P(\mathbf{p} | \mathbf{B}) \approx P(p_1) \left[ \prod_{n=2}^N P(p_n | p_{n-1}, B_{n-1}) \right] \quad (2.10)$$

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

$$P(\mathbf{B} | \mathbf{l}) = \prod_{n=1}^{N-1} P(B_n | \mathbf{l}_n). \quad (2.11)$$

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

## 2.3 Joint Prosody Labeling and Modeling

A sequential optimization procedure based on the ML criterion is proposed to jointly label the prosodic tags for all utterances of the training corpus and to estimate the parameters of the four prosodic models. It is divided into two main parts: initialization and iteration. The initialization part determines initial prosodic tags of all utterances and estimates initial parameters of the four prosodic models by a specially designed procedure. The iteration part first defines an objective likelihood function for each utterance by

$$Q = \left( \prod_{n=1}^N P(\mathbf{sp}_n | p_n, B_{n-1}, t_{n-1}^{n+1}) \right) \left( P(p_1) \prod_{n=2}^N P(p_n | p_{n-1}, B_{n-1}) \right) \left( \prod_{n=1}^{N-1} (P(pd_n, ed_n | B_n, \mathbf{I}_n) P(B_n | \mathbf{I}_n)) \right). \quad (2.12)$$

It then applies a multi-step iterative procedure to update the labels of prosodic tags and the parameters of the four prosodic models sequentially and iteratively. In the following subsections, we discuss the sequential optimization procedure in detail.

### 2.3.1 Initialization

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

#### (a) Initial labeling of break indices

The initial break index of each syllable juncture is determined by a decision tree (see Figure 2.2) designed based on a prior knowledge about break labeling/modeling gained in previous studies [8,28,40,52-58,64,65]. It is known that pause duration is the most important acoustic cue to specify breaks. Most word junctures with PM have long pauses so that they are most likely labeled as major break, or in our case  $B3$  and  $B4$ . On the other hand, most intra-word syllable junctures have very short pause duration so that they are generally labeled as non-break, or in our case  $B0$  and  $B1$ . Moreover,  $B0$  represents tightly coupled syllable juncture so that it is distinguished

from  $B1$  by having very short pitch pause duration and high energy-dip level. In-between these extreme situations, non-PM inter-word junctures with medium pause duration and with medium pitch jump are likely labeled as  $B2-2$  and  $B2-1$ , respectively. By using the prior knowledge, we develop the algorithms to determine all thresholds of the decision tree ( $Th1 \sim Th6$ ) in a systematic way to avoid doing it manually or by trial and error. Detail of the algorithms is given in Appendix A.

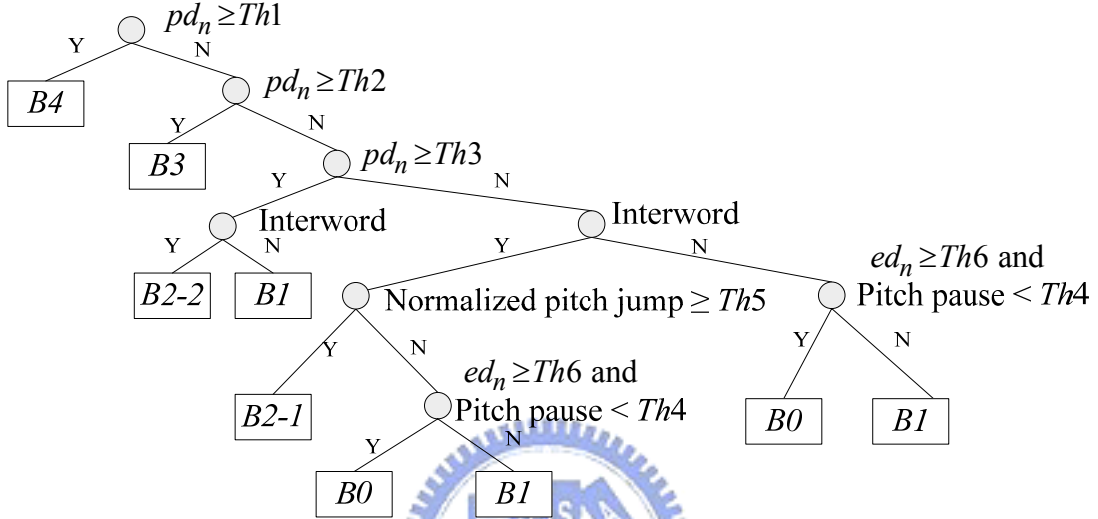


Figure 2.2: The decision tree for initial break type labeling.

**(b) Estimation of the initial parameters of the four prosodic models and prosodic-state indices**

The initializations of the break-acoustics model and the break-syntax model can be done independently with initial break indices of all syllable junctures being given. We realize them by the CART algorithm [66]. For the initialization of the break-acoustics model, the CART algorithm with the node splitting criterion of maximum likelihood gain is adopted to classify pause duration  $pd_n$  and energy-dip level  $ed_n$  for each break type  $B$  according to a question set  $\Theta_1$  derived from the contextual linguistic features  $\mathbf{I}_n$ . Each leaf node represents the product of a gamma distribution  $g(pd_n; \alpha_{B, \mathbf{I}_n}, \beta_{B, \mathbf{I}_n})$  and a normal distribution  $N(ed_n; \mu_{B, \mathbf{I}_n}, \sigma_{B, \mathbf{I}_n}^2)$ . For the initialization of the break-syntax model  $P(B_n | \mathbf{I}_n)$ , a decision tree is built by using another question set  $\Theta_2$  derived also from  $\mathbf{I}_n$  to classify break types. The details of  $\mathbf{I}_n$ ,  $\Theta_1$ , and  $\Theta_2$  used in this study are listed in Appendixes B, C.1, and C.2, respectively.

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

### 2.3.2 Iteration

The iteration is a multi-step iterative procedure listed below:

- Step 1:* Update the APs of five tones  $\boldsymbol{\beta}_t$  with all other APs being fixed.
- Step 2:* Update the APs of coarticulation  $\{\boldsymbol{\beta}_{B,tp}^f, \boldsymbol{\beta}_{B,tp}^b, \boldsymbol{\beta}_{B_b,t_1}^f$  and  $\boldsymbol{\beta}_{B_e,t_N}^b\}$  with all other APs being fixed, and then update  $\mathbf{R}$ .
- Step 3:* Re-label the prosodic state sequence of each utterance by using the Viterbi algorithm so as to maximize  $Q$  defined in Eq. (2.12). Then, update the APs of prosodic state  $\boldsymbol{\beta}_p$ , the prosodic state model  $P(\mathbf{p}|\mathbf{B})$  and  $\mathbf{R}$ .
- Step 4:* Re-label the break type sequence of each utterance by using the Viterbi algorithm so as to maximize  $Q$ . Then, update the prosodic state model  $P(\mathbf{p}|\mathbf{B})$  and  $\mathbf{R}$ .
- Step 5:* Re-construct the decision trees to update  $P(pd_n, ed_n | B_n, \mathbf{I}_n)$  and  $P(B_n | \mathbf{I}_n)$  by the CART algorithm using the question sets  $\Theta_1$  and  $\Theta_2$ , respectively.
- Step 6:* Repeat *Steps 1* to *5* until a convergence is reached.



## 2.4 Experimental Results

The experiment was conducted on the training set which consisted of 379 utterances with in total 52192 syllables. The number of prosodic states was properly set to be 16 because the root mean squared error (RMSE) of VQ saturated when the number of prosodic states was greater than 16. As shown in Figure 2.3, the sequential optimization procedure took 69 iterations to reach a convergence. Following is examinations and interpretations of the parameters of the four prosodic models.

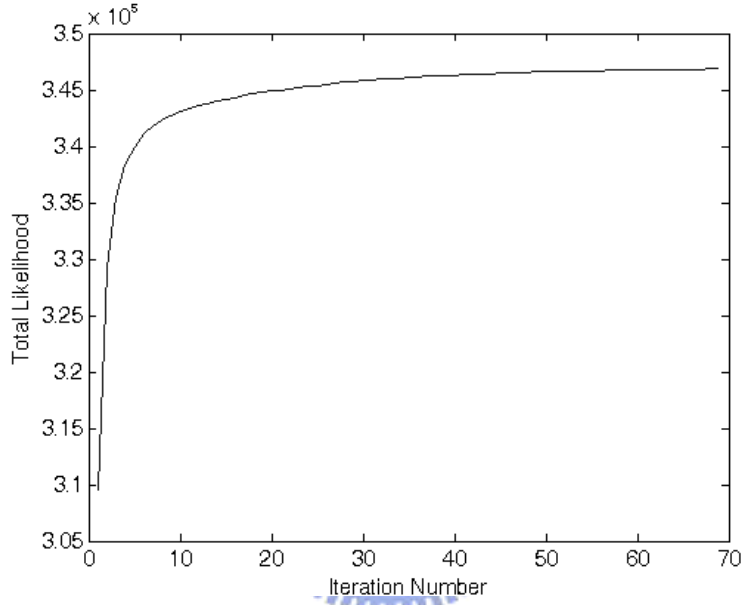


Figure 2.3: The plot of total log-likelihood versus iteration number.

### 2.4.1 The Syllable Pitch Contour Model

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

$$\mathbf{R}_{\mathbf{sp}} = \begin{bmatrix} 565.4 & 23.9 & -25.6 & -0.5 \\ 23.9 & 90.5 & 9.7 & -8.2 \\ -25.6 & 9.7 & 17.8 & -0.9 \\ -0.5 & -8.2 & -0.9 & 5.0 \end{bmatrix} \times 10^{-4} \Rightarrow \mathbf{R}_{\mathbf{sp}^r} = \begin{bmatrix} 3.5 & 0.2 & -0.2 & 0.0 \\ 0.2 & 31.9 & 2.6 & -1.5 \\ -0.2 & 2.6 & 11.1 & 0.6 \\ 0.0 & -1.5 & 0.6 & 3.7 \end{bmatrix} \times 10^{-4}$$

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

Figure 2.4 displays the APs of five tones. We find from the figure that the APs of the first four tones conformed well to the standard tone patterns found by Chao [67]. As for tone 5, its low dipping pattern resembles the pattern of tone 3 to some degree. This also matched the finding in the previous study about tone 5 [68].

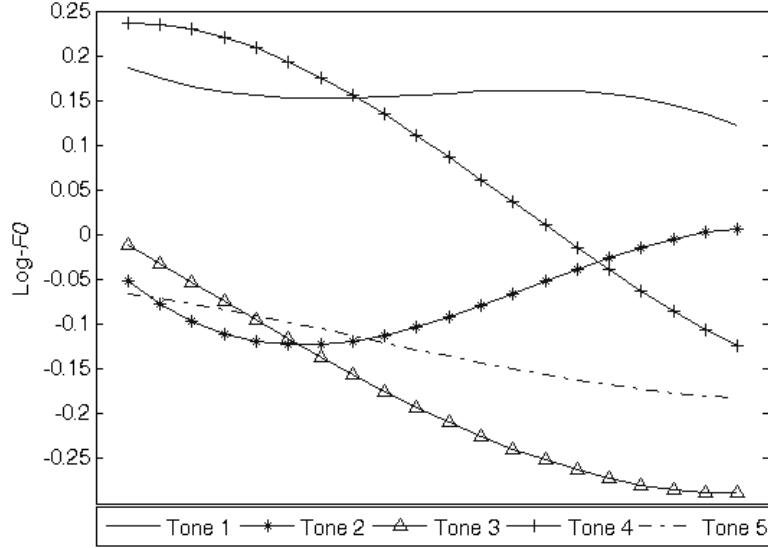


Figure 2.4: The APs of five tones

Table 2.1 displays the APs (log- $F0$  levels) and the distribution of the 16 prosodic states. It can be seen from Table 2.1 that these 16 log- $F0$  levels spanned widely to cover the whole dynamic range of log- $F0$  variation with lower indices of prosodic state corresponding to lower log- $F0$  levels; and the prosodic states distributed normally with relatively few located at the two extremes of high and low prosodic states.

Table 2.1: The APs (log- $F0$  levels,  $\beta_p(1)$ ) and the distribution ( $P(p)$ ) of the 16 prosodic states.

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

Figures 2.5(a) and 2.5(c) display the APs of forward and backward coarticulations,  $\beta_{B,tp}^f$  and  $\beta_{B,tp}^b$ , for the three break types of  $B0$ ,  $B1$ , and  $B4$ . These three break types were chosen on purpose to show extreme cases of inter-syllable

coarticulation: *B0* for tightly coupling, *B1* for normal coupling, and *B4* for no coupling. Some interesting phenomena can be observed from the figure. First, it can be seen from Figure 2.5(a) that most APs of forward coarticulation for *B0* and *B1*,  $\beta_{B0,tp}^f$  and  $\beta_{B1,tp}^f$ , were bended in their beginning parts. These bendings were to compensate the level mismatch between the beginning and ending parts of the log-*F0* contours of the tone pairs for highly coarticulated preceding and current syllables, so as to make their log-*F0* contours be concatenated more smoothly. For example, the upward bending at the beginning parts of  $\{\beta_{B,tp}^f | tp = (1,2), (1,3), (2,2), (2,3), (1,5)\}$  were due to *H-L* mismatches, while the downward bending at the beginning parts of  $\{\beta_{B,tp}^f | tp = (3,1), (3,4), (5,1), (5,4), (4,1), (4,4)\}$  corresponded to *L-H* mismatches. Similarly, it can be observed from Figure 2.5(c) that the ending parts of the APs of backward coarticulation for *B0* and *B1*,  $\beta_{B0,tp}^b$  and  $\beta_{B1,tp}^b$ , were bended. But the degrees of their upward and downward bendings were generally smaller. This conformed to the observation reported in Ref. [69] that the carry-over effect on the syllable *F0* contour influenced by the preceding syllable is much larger than the anticipation effect caused by the following syllable. Second, it can be found from Figures 2.5(a) and 2.5(c) that most APs of forward and backward coarticulations for *B4* with the same current tone looked similar and hence were nearly independent of their respective preceding and succeeding tones. This showed that the inter-syllable coarticulation across a *B4* break was relatively low as compared with those of *B0* and *B1*. Moreover, many APs of forward and backward coarticulations for *B4* were downward bended in their beginning and ending parts, respectively. They exhibited the onset and offset phenomena at the beginning and ending syllables of BG/PG. Furthermore, we find from Figures 2.5(b) and 2.5(d) that most utterance initial and final patterns,  $\beta_{B_e,t}^f$  and  $\beta_{B_e,t}^b$ , looked very similar to those of  $\beta_{B4,tp}^f$  and  $\beta_{B4,tp}^b$ , respectively, to show the same onset and offset phenomena at the two types of utterance boundaries. We also find that  $\beta_{B_e,3}^b$  and  $\beta_{B_e,5}^b$  were two exceptional patterns which had lower levels. These probably resulted from the total relaxation of pronunciation at the utterance ending for these two tones. Third, it can be found from Figure 2.5(c) that the APs of  $\beta_{B0,(3,3)}^b$  and  $\beta_{B1,(3,3)}^b$  were upward bended drastically in their ending parts. As combining with the AP of tone 3 shown in Figure 2.4, these

bendings would make the integrated log- $F_0$  patterns of the first syllable in a (3,3) tone pair change from middle-falling tone-3 shape to middle-rising tone-2 shape to fulfill the well-known 3-3 tone *sandhi* rule which says that the first tone 3 of a 3-3 tone pair will change to a tone 2. On the contrary, we find that the pattern  $\beta_{B4,(3,3)}^b$  did not bend upward. This showed that the 3-3 tone *sandhi* rule did not apply when the syllable juncture was a  $B4$ . Last, we made some comments to the APs of forward and backward coarticulation for  $B2-1$ ,  $B2-2$  and  $B3$ . Basically, the APs of  $B2-1$  and  $B3$  resembled to those of  $B4$  but with smaller upward and downward bendings, and  $B2-2$  had similar patterns to those of  $B1$  but with smaller upward and downward bendings.

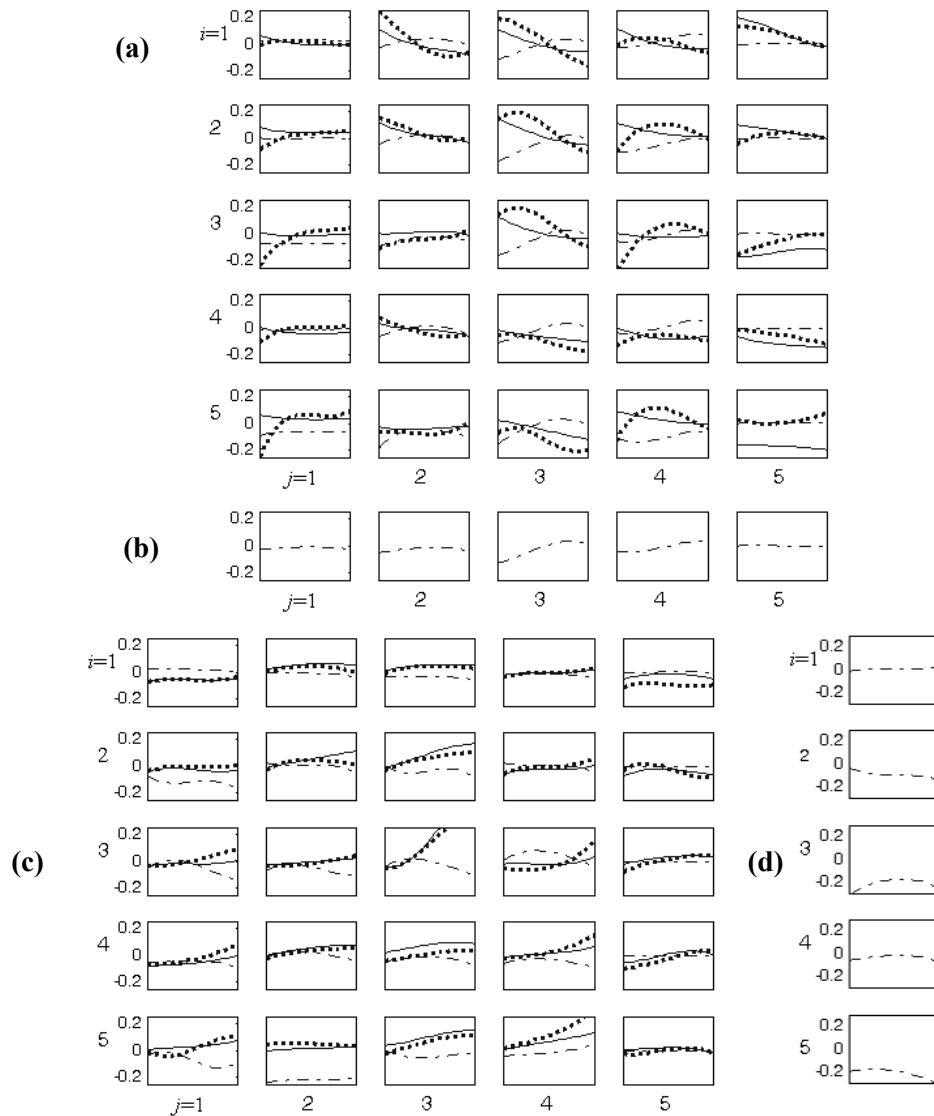


Figure 2.5: The (a) forward and (c) backward coarticulation patterns,  $\beta_{B,tp}^f$  and  $\beta_{B,tp}^b$ , for  $B0$  (point line),  $B1$ (solid line), and  $B4$ (dashed line); and the (b) onset and (d) offset patterns,  $\beta_{B_b,t}^f$  and  $\beta_{B_e,t}^b$ , for  $B_b$  and  $B_e$ . Here  $tp = (i, j)$  and  $t = i$  or  $j$ .

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

## 2.4.2 The Break-Acoustics Model

The two break-acoustics models,  $g(pd_n; \alpha_{B_n, I_n}, \beta_{B_n, I_n})$  and  $N(ed_n; \mu_{B_n, I_n}, \sigma_{B_n, I_n}^2)$ , were built by the decision tree method using the question set  $\Theta_1$ . One decision tree was constructed for each break type. Figure 2.6 displays the distributions of pause duration and energy-dip level for the root nodes of these six break types. It can be found from the figure that the break types of higher level were generally associated with longer pause duration and lower energy-dip level.  $B0$  had very short pause duration and wide-spread energy-dip level with very high mean value.  $B1$  and  $B2-1$  had similar distributions of short pause durations and wide-spread high energy-dip level.  $B2-2$  had medium long pause duration and medium high energy-dip level. Both  $B3$  and  $B4$  had wide-spread long pause duration and low energy-dip level. These conformed to the prior knowledge about break types [4,8,70,71].

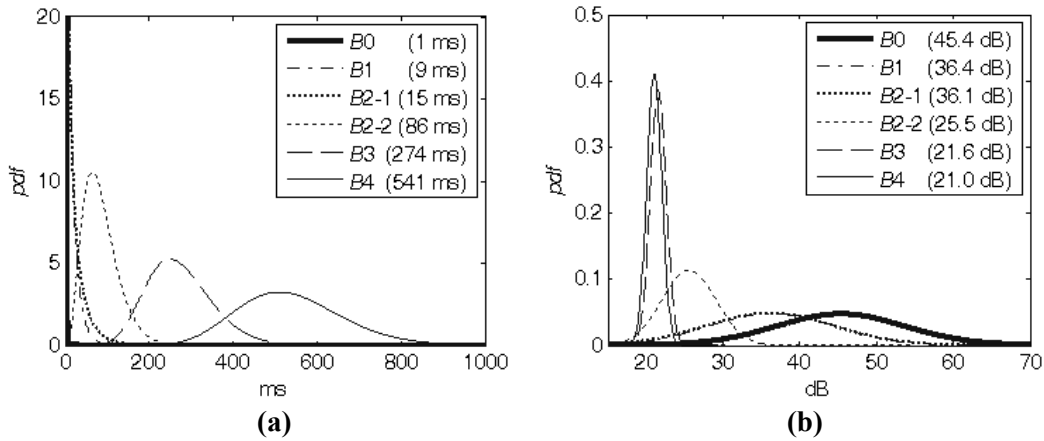


Figure 2.6: The *pdfs* of (a) pause duration and (b) energy-dip level for the root nodes of these 6 break types. Numbers in () denote the mean values.

To further examine the model, we show its decision trees for the five break types of  $B4$ ,  $B3$ ,  $B2-2$ ,  $B2-1$  and  $B1$  in Figure 2.7. It is noted here that no tree split for

$B0$  due to the relative uniformity on the acoustic prosodic features of its samples. Generally, the questions used to split trees of higher-level break types ( $B4$  and  $B3$ ) tended to be related to higher-level syntactic features, such as PM ( $Q_{1.2.4}$ ) and syntactic phrase size ( $Q_{3.1.3}$ ,  $Q_{3.3.2}$ ,  $Q_{3.3.11}$ , and  $Q_{3.3.18}$ ). On the contrary, the questions of lower-level phonetic features ( $Q_{1.1}$ ,  $Q_{1.3}$ , and  $Q_{1.4}$ ) tended to split trees of lower-level break types ( $B1$  and  $B2-1$ ).

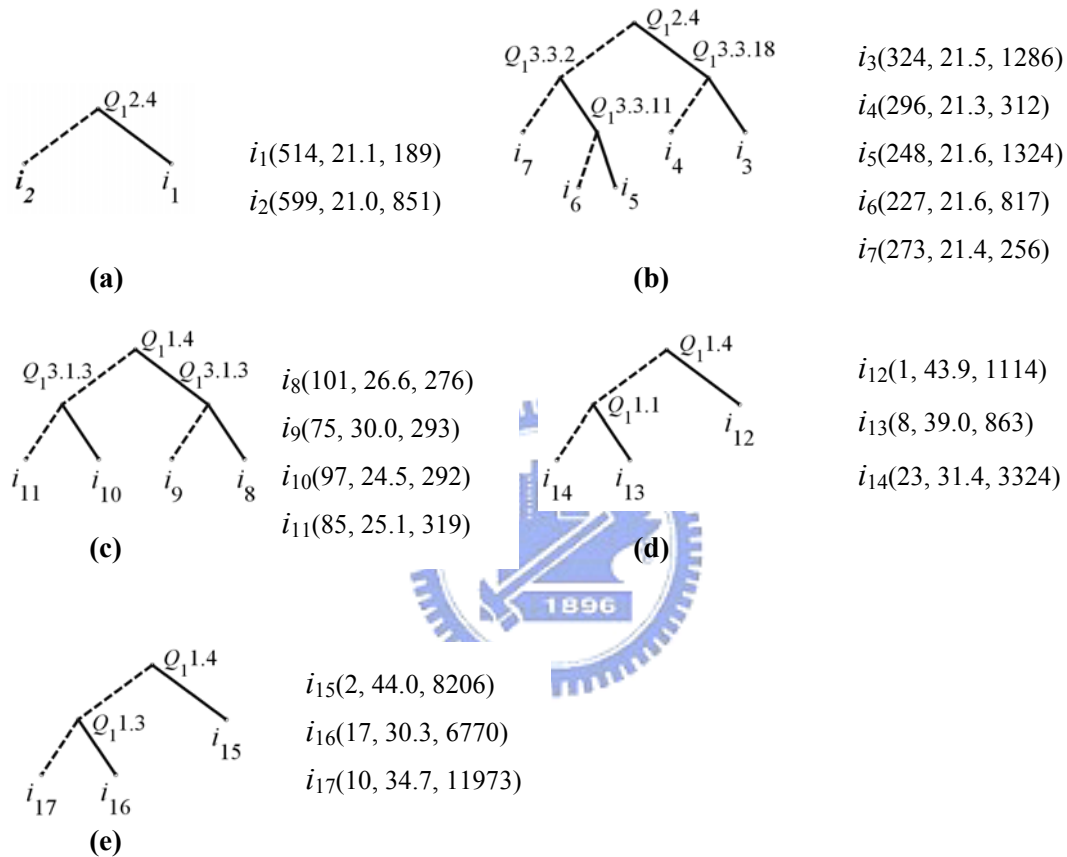


Figure 2.7: The decision trees of the break-acoustics model for (a)  $B4$ , (b)  $B3$ , (c)  $B2-2$ , (d)  $B2-1$ , and (e)  $B1$ . The numbers in a bracket denote average pause duration in ms (left), energy-dip level in dB (middle) and sample count (right) of the associated node. Solid line indicates positive answer to the question and dashed line indicates negative answer.

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

### 2.4.3 The Prosodic State Model

We then examined the prosodic state model. Figure 2.8 displays some most significant transitions of  $P(p_n | p_{n-1}, B_{n-1})$  for six break types. For  $B0$  and  $B1$ , the general high-to-low, nearby-state transitions showed that the syllable log- $F0$  level declined slowly within PWs. We also find that some low-to-high, nearby-state transitions occurred within PWs of low pitch level. This demonstrated the sustaining phenomenon of the log- $F0$  trajectory at the ending part of some PPhs.

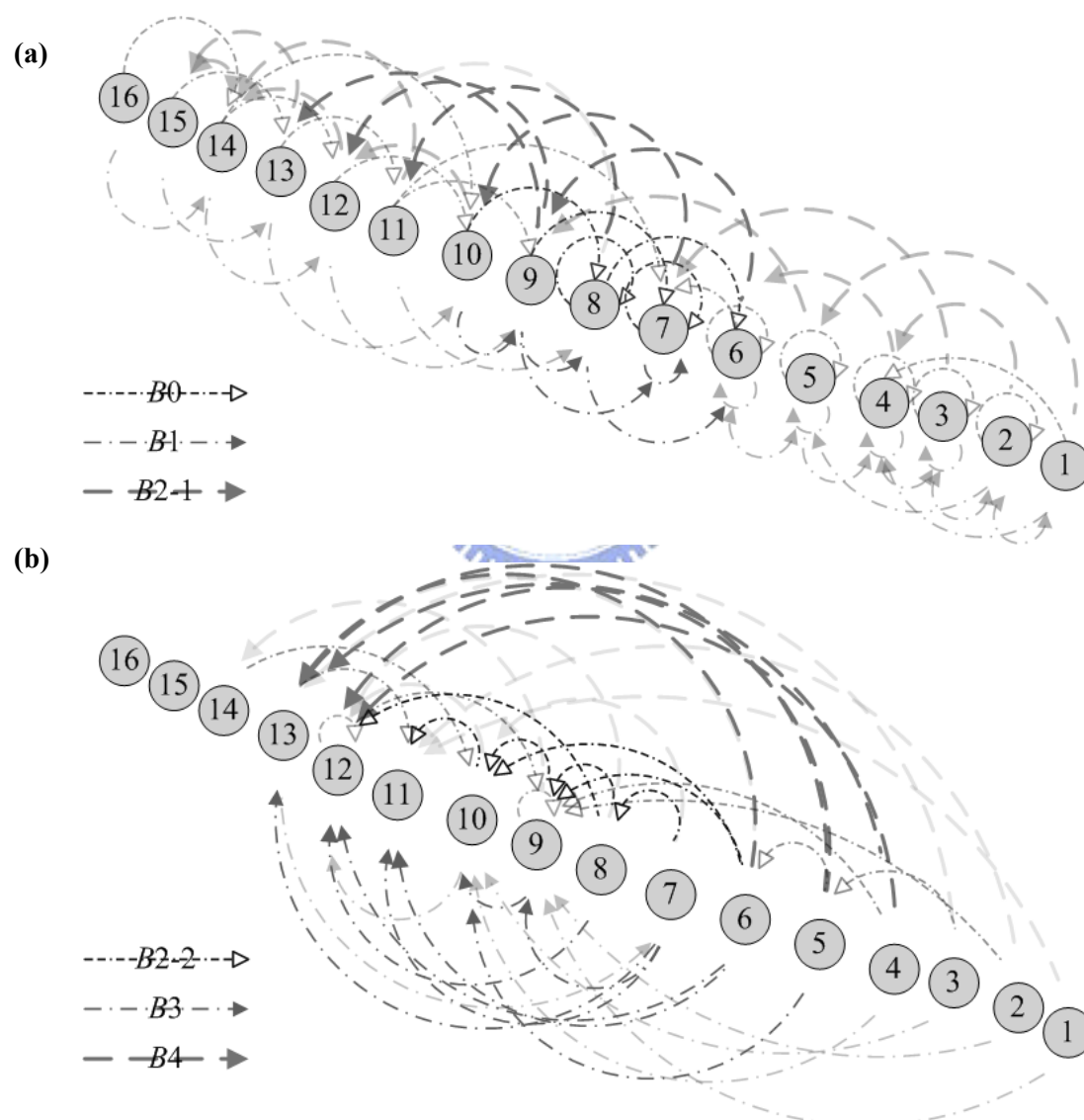


Figure 2.8: The most significant prosodic state transitions for (a)  $B0$ ,  $B1$  and  $B2-1$ , and (b)  $B2-2$ ,  $B3$  and  $B4$ . Here, the number in each node represents the index of the prosodic state. Note that bold and thin lines denote the primary and secondary state transitions, respectively.

For *B2-2*, it had both high-to-low and low-to-high state transitions. For *B2-1*, *B3*, and *B4*, their low-to-high state transitions showed clearly the phenomena of syllable log-*F0* level resets across PWs, PPhs, and BG/PGs. Comparing with these clear log-*F0* level resets, the resets of *B2-2* were insignificant. Combining the results shown in Figures 2.6 and 2.8, we find that *B2-1* and *B2-2* had different acoustic characteristics: *B2-1* had significant log-*F0* reset with very short pause duration, while *B2-2* had longer pause duration with low or no log-*F0* reset.

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

#### 2.4.4 The Break-Syntax Model

The break-syntax model  $P(B_n | I_n)$  was built by the decision tree method using the question set  $\Theta_2$ . Figure 2.9 displays the decision tree of the break-syntax model. The tree was divided into four sub-trees, *T3-T6*, by the three questions of  $Q_2$ .2.1.1 (PM?),  $Q_2$ .2.1.3 (minor PM?) and  $Q_2$ .1.3 (intra-word?). It can be seen from the figure that the root node of sub-tree *T3*, which corresponded to syllable juncture with minor PM, was mainly composed of *B3* and *B4*. Similarly, the root nodes of sub-trees *T4* and *T5*, corresponding to major PM and intra-word syllable juncture, were mainly composed of *B4* and *B0/B1*, respectively. Due to the fact that the break-type constituents of both *T4* and *T5* were pure, they had very simple tree structures. On the contrary, sub-tree *T6* was a miscellaneous collection of all other types of syllable juncture without PM. So, it had the most complex tree structure.



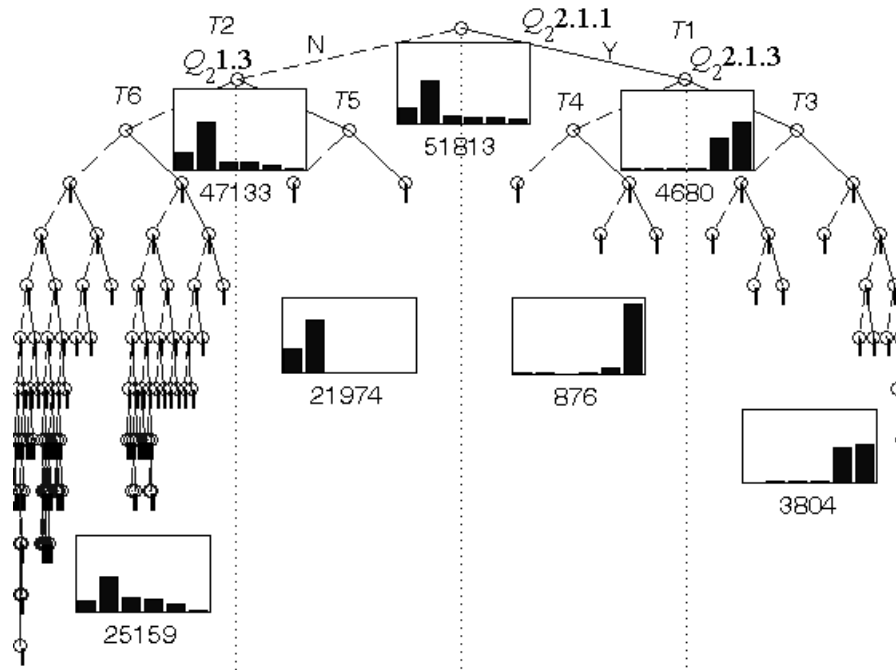


Figure 2.9: The decision tree of the break-syntax model. The bar plot associated with a node denotes the distributions of these six break types ( $B_0$ ,  $B_1$ ,  $B_{2-1}$ ,  $B_{2-2}$ ,  $B_3$ ,  $B_4$ , from left to right) and the number is the total sample count of the node.

Figure 2.10 displays the more detailed structures of these four sub-trees up to the fourth layer. From Figures 2.10(a) and 2.10(b), we find that nodes in  $T_3$  and  $T_4$  were mainly split by questions related to high-level linguistic features such as  $Q_{2.3.3.19}$  (Is the length of the following syntactic phrase/sentence greater than 6?) and  $Q_{2.3.3.29}$  (Is the length of the preceding syntactic phrase/sentence greater than 7?). As shown in Fig. 2.10(c),  $T_5$  had two leaf nodes split by  $Q_{2.1.1}$  (Does the following syllable have a null initial or initial in  $\{m, n, l, r\}$ ?). The set associated with positive answer was mainly composed of  $B_0$ , while another set was mainly composed of  $B_1$ . As shown in Figure 2.10(d),  $T_6$  was constructed by questions related to features of various levels, including  $Q_{2.1.1}$ ,  $Q_{2.2.4.18}$  (Is the preceding word “DE”?),  $Q_{2.2.3.2}$  (Is the preceding word a function word?),  $Q_{2.3.3.24}$  (Is the length of the preceding syntactic phrase greater than 2?), and so on. We also find from Figure 2.10 that the purities of the break-type constituents were high for leaf nodes of  $T_4$  and  $T_5$ , medium high for nodes of  $T_3$ , and relatively low for most nodes of  $T_6$ . This implies that it is difficult to correctly label (or predict) the break types of syllable junctures other than intra-word and those with major PM by the break-syntax model using only linguistic features without the help of acoustic cues.

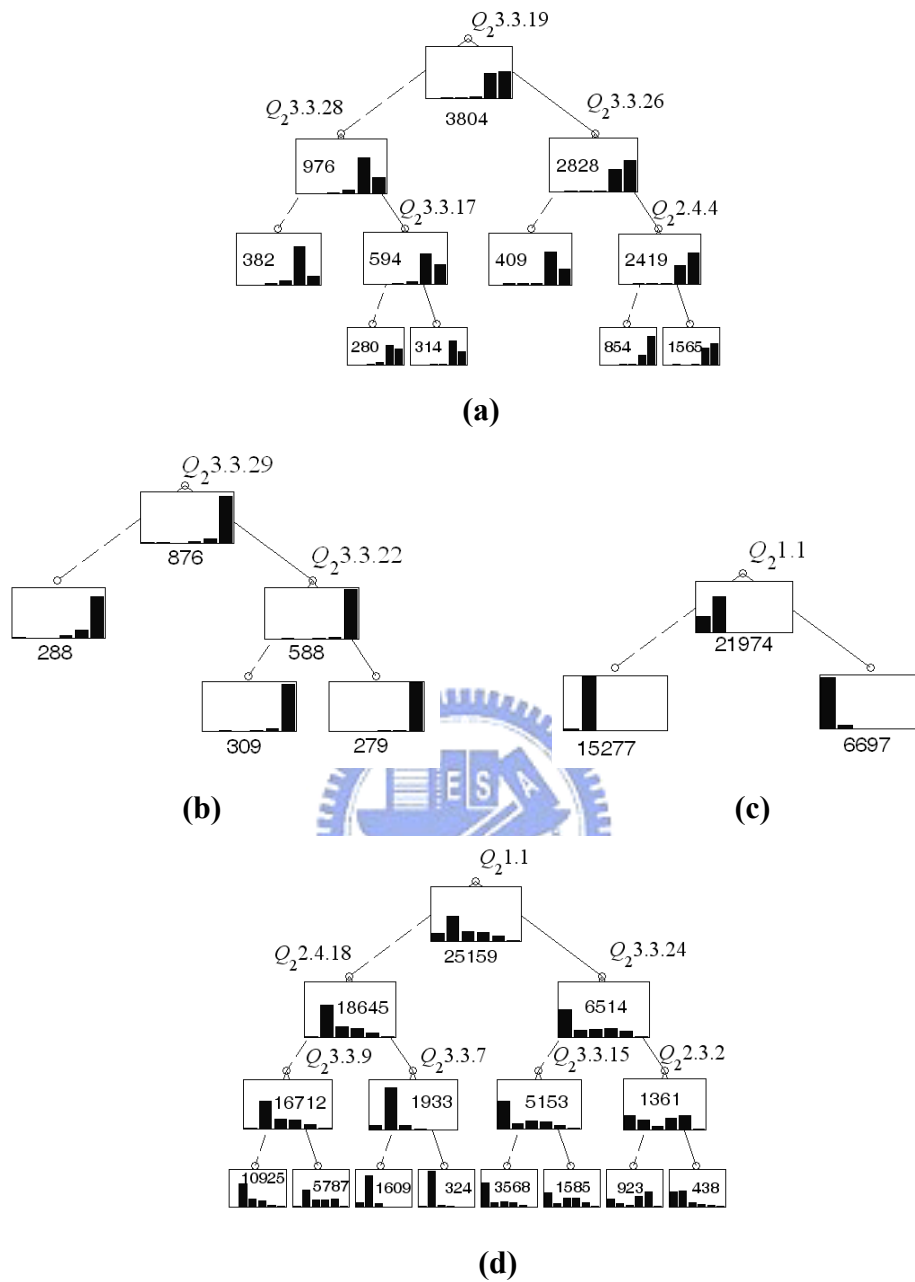


Figure 2.10: The more detailed structures of sub-trees of (a)  $T_3$ , (b)  $T_4$ , (c)  $T_5$  and (d)  $T_6$ . Solid line indicates positive answer to the question and dashed line indicates negative answer.

## 2.5 Analyses of the Labeled Breaks and Prosodic Constituents

To further evaluate the performance of the proposed method, explorations in the relationships between prosodic breaks and linguistic features of texts, the length of prosodic constituents, and the general pitch patterns of prosodic constituents obtained in our method were described in Subsections 2.5.1 through 2.5.3. Besides, to further verify the labeling outcomes generated by our models, a comparison conducted between human labeling and our labeling was given in Subsections 2.5.4 and 2.5.5.

### 2.5.1 Analyses of the Labeled Break Types

Since the purpose of announcers' broadcasting is to propagate information accurately to the audience relying exclusively on their audio perception, our well-trained informant skillfully manipulated as many segmental and prosodic cues as possible, such as clear and precise articulation, strategic variations in the fundamental frequency, volume, syllable length, and types of breaks. These prosodic information carried in the utterance speech, in turn, reflects the informant's mental grammar, his/her Mandarin linguistic competence that determines when to form a semantically appropriate word chunk, a prosodic phrase, or a larger unit, and hence where and how long a break in an utterance should be so that the informant's speech would sound natural, informative, and attention attracting to the audience.

As a research based on our informant's speech data rich in the Mandarin prosodic cues, our break-type-labeling model also can generate appropriate break types consistent with native speakers' psychological reality. To verify this point, we examined the relationship between some special groups of words/morphemes and their concurring break types that both our break-type-labeling model and the ordinary Mandarin native speakers would consistently produce. These special groups of words/morphemes include (1) affix morpheme; (2) DE; (3) Ng, Di, and T; (4) VE; (5) Caa and Cb; and (6) P [72]. The results are discussed in more detail as follows.

#### 1. *Set of Affix Morpheme*

It is well-known that prefixes and suffixes are bound morphemes that attach to their preceding or following heads to form units of complex words. Since the resultant form after combining the head and the affix is a unit, it is reasonable to predict that

the breaks at the boundaries between the head and the affix tend to fall in *B0* or *B1* types. These phenomena were observed in our corpus. We found that some Mandarin Chinese mono-syllabic prefixes, such as *bu-* “不 not, un-, dis-, in-,” *ke-* “可 -able,” *wu-* “無 not, -less, un-,without,” etc. [67,73], tend to join the following roots to form legitimate words as in *bu-li* “不利 unfavorable,” *bu-fang-bian* “不方便 inconvenient,” *ke-wu* “可惡 detestable,” *ke-sing* “可行 feasible,” *wu-sian* “無限 limitless,” *wu-shuang* “無雙 unparalleled.” Similarly, by attaching mono-syllabic suffixes, such as *-bian* “邊 side,” *-zhe* “者 -er, -or,” *-hua* “化 -ize,” etc., to the preceding roots, we can derive complex words as in *lu-bian* “路邊 roadside, curb,” *he-bian* “河邊 riverside,” *zuo-zhe* “作者 author, writer,” *sing-zhe* “行者 religious practitioner,” *gung-yie-hua* “工業化 industrialize,” *min-zhu-hua* “民主化 democratize.”

Table 2.2 lists the statistics of the break types labeled for the syllable boundaries of 121 prefixes and 195 suffixes. It can be seen from the table that 79.6% of the post-syllable boundaries of these 121 prefixes and 98.5% of the pre-syllable boundaries of these 195 suffixes were labeled as *B0* or *B1*. These prosodic findings reflect the fact that morphologically the combination of head and affix generates a lexical unit, and thus the break between them is determined to be the break type of intra-PW category by our method. The results were also consistent with some rules found in Refs. 64, 65, and 68.

Table 2.2: Statistics of break types labeled for 121 prefixes and 195 suffixes.

Labeled break type		<i>B0</i>	<i>B1</i>	<i>B2-1</i>	<i>B2-2</i>	<i>B3</i>	<i>B4</i>	Total count
Prefix	Pre-boundary	94	1289	460	545	193	5	2586
	Post-boundary	584	1475	344	178	5	0	2586
Suffix	Pre-boundary	1046	2466	31	20	3	0	3566
	Post-boundary	307	1479	272	482	568	458	3566

## 2. Word Set of DE

The words in the DE set particularly refer to *de*, *zhe*, and *di*, which serve multi-functions including a possessive marker, an adjective marker, and an adverbial marker [72]. They are characterized by the fact that a DE word can combine with a wide range of preceding syntactic constituents to form a possessive adjective as in an noun phrase (NP)-*de* structure: *xue-sheng-de quan-li* “學生的權力 students’ right,” to derive an adjective phrase as in a verb phrase (VP)-*de* structure: *se-siang-zhe qin*

“思鄉之情 nostalgia,” or to function as an adverbial phrase as in a DM-de structure: *ke-ren yi-bo-bo-di yong-jin-dien-lai* “客人一波波地湧進店來 guest were flocking to the shop.” Despite the variety of the preceding constituent, a DE word, similar to a suffix, builds closer connection with its preceding constituent to form a larger syntactic unit; consequently, it is predictable that the break at the DE words’ pre-boundary position tends to fall into *B0* and *B1*, which means a pause is hardly to be perceived at this juncture. It is also reasonable to infer that due to a looser connection between the DE words and the following constituent, less *B0* and *B1* would occur at the post-boundary position.

The statistics in Table 2.3 indicates that the distribution of the break types labeled by our model just conformed with our anticipation; while 92.3% pre-boundary breaks of the DE words were *B0* and *B1*, only 65% post-syllable boundaries of the DE words fell into the same types, which suggests that for the DE words, the majority of the neighboring breaks are unperceivable, and in most cases only at the post-boundary position can perceivable breaks be sensed. This result also matched the findings in Refs. 64, 65, and 68. Two examples are given below for illustration:

**Ex.1:** ... 。因為(because, Cbaa) 女性(women, Nad) 在(in, P21) 社會(society, Nac) *B1* 的(DE) *B1* 地位(position, Nad) 提高(raise, VC2) , ...  
 (... Because women’s social status has been improved, ...)

**Ex.2:** ... 。目前(now, Nddc) 我(I, Nhaa) 是 (am, V\_11) 三十一歲(thirty one year old, DM) *B1* 的(DE) *B2-2* 單身(single, VH11) 女郎(woman, Nab) ,  
 (... Now, I am a thirty-one-year-old single woman.)

Table 2.3: Statistics of break types labeled for the DE words.

Labeled break type	<i>B0</i>	<i>B1</i>	<i>B2-1</i>	<i>B2-2</i>	<i>B3</i>	<i>B4</i>	Total count
Pre-boundary	168	1600	146	1	0	0	1915
Post-boundary	210	1035	331	294	41	4	1915

### 3. Word Sets of Ng, Di, and T

Ng, Di, and T represent the word sets of Mandarin Chinese localizers, aspectual adverbs, and particles [72], respectively. The distinctive shared feature of these sets of words is that almost all the words are no longer than two syllables in length and that when combining with other syntactic constituent to form a larger phrase, they are all positioned at the end of the derived phrase, such as *san-tian hoNg* “三天後 three days

latter,” *kai-hui dang-zhung*<sub>Di</sub> “開會當中 while the meeting is being held,” and *bu-qu le-ma*<sub>T</sub> “不去了嗎? Not going?.” Due to the characteristic of being post-positioned in a phrase, these words are inclined to be incorporated with their preceding constituents, and predictably barely any pauses can be perceived at the pre-boundary position.

The statistic results listed in Table 2.4 indicate that our model’s break-labeling performance just exactly met our expectation. As high as 94%, 93%, and 87% of the pre-syllable boundaries of the words in this category were labeled as *B0* or *B1*. Three examples are given below:

**Ex.3:** ... , 牠(it, Nhaa) 跟(and, Caa) 婦人(woman, Nab) 和(and, Caa) 相命(tell one's fortune, VB11) 者(person, Nab) B1 之間(between, Ng) B3 並(Dbb) 無(no, VJ) 勾串(conspire, Nv1) 作弊(cheat, Nv4) , ...  
(... There is no conspiracy to cheat among it, the woman and the fortune teller.)

**Ex.4:** ... , 忘記(forget, VK1) B0 了(Di) B1 婚後(after marriage, Ndc) 現實(realistic, VH11) 的(DE) 環境(environment, Nac) 。 Be  
(... have forgotten the realistic environment after marriage.)

**Ex.5:** ... , 因為(because, Cbaa) 我(I, Nhaa), 永遠(always, Dd) 有(have, V\_2) 看(read, VC2) 不完(have no limit, VC2) 的(DE) 書(book, Nab) B1 啊(“ah”, Tc) B4 !  
(... for I always have books to read!)

On the other hand, it is also interesting to find that for the breaks at the post-syllable boundaries, 67% and 96% of them were labeled as *B3* /*B4* especially for the Ng-set and T-set words, respectively. Further investigation reveals that most of the longer breaks were caused by a following PM, an index representing the occurrence of a detectable pause. Besides, because the T-set words are phrasal or sentential final particles and hence are highly likely to be followed by a PM, a much higher ratio of *B3*/*B4* could be found. Two examples are given in the following:

**Ex.6:** *Bb* 對(to, P31)人類(human, Naeb) B0 來說(in some sense, Ng) B3 , ...  
(...To human beings, ...)

**Ex.7:** ... , 想(think, VE2) 辦法(idea, Nac) 解決(to solve, VC2) B0 而已(nothing but, Tb) B4 ! ...  
(...just figured out the solution ...)

Table 2.4: Statistics of break types labeled for the word sets of Ng, Di, and T.

Labeled break type		<i>B0</i>	<i>B1</i>	<i>B2-1</i>	<i>B2-2</i>	<i>B3</i>	<i>B4</i>	Total count
Ng	Pre-boundary	97	420	19	12	0	2	550
	Post-boundary	26	81	17	58	245	123	550
Di	Pre-boundary	107	83	12	1	0	2	205
	Post-boundary	30	68	36	41	11	19	205
T	Pre-boundary	89	84	14	11	0	0	198
	Post-boundary	0	5	1	2	22	168	198

#### 4. Word Set of VE

VE represents a class of transitive verbs that take a sentence as the object, such as *ren-wei* “認為 to suppose/think/believe (that),” *gan-dao* “感到 to feel (that),” *biao-she* “表示 to show/indicate/mean/suggest (that),” etc [72]. It is evident that since the message carried in a sentential object, compared to a NP object for example, demands longer time to process mentally before being accurately expressed, a longer pause is reasonably anticipated to occur after a VE verb for information operation. Based on the statistic results listed in Table 2.5, on the whole 72% post-word boundaries of the VE verbs were labeled as breaks with distinctly audible pauses, namely *B2-1*, *B2-2*, *B3*, or even *B4*, another quite favorable evidence that the break types labeled by our model were consistent with the pause duration people usually take in their utterances. A typical example is given in Ex. 8.

Table 2.5: Statistics of break types labeled for word set of VE.

Labeled break type	<i>B0</i>	<i>B1</i>	<i>B2-1</i>	<i>B2-2</i>	<i>B3</i>	<i>B4</i>	Total count
Post-boundary	63	177	99	108	159	234	840

**Ex.8:** ...。當局(government, Nad) *B1* 說(proclaim, VE2) *B3* 他(he, Nhaa) 違反(violate, VJ1) 市政府(city government, Ncb) 一戶(one, DM) 人家(family, Nab) 不能(can not, Dbab) 儲存(store, VC33) 二千五百雙(2,500, DM) 鞋子(shoes, Nab) 的(DE) 規定(rule, Nac), ...  
( ...City government proclaimed that he violated the rule that no more than 2,500 pairs of shoes were allowed to be stored in a family.... )

However, it cannot be neglected that no less than 28% post-word boundaries of the VE verbs were labeled as *B0* or *B1*, implying that seemingly our model still

generated quite a few unexpected break types for the VE verbs. Further observation of the data, nevertheless, found two main reasons to account for this discrepancy of labeling. First, besides a sentential object, part of the VE verbs could also take a NP object, so the breaks occurring before a NP object were predictably shorter than before a sentential object. A typical example is given below:

**Ex.9:** ... , 因(because, Cbaa) 你(you, Nhaa) 可(can, Dbab) 從(from, P19) 過去(past, Ndda) 經驗(experience, Nac) 中(in the course of, Ng) B2-2 檢討(examine, VE2) B1 成敗(success and failure, Nad) , ...

(...because you can examine your success and failure based on the past experience...)

The other reason for the occurrence of *B0/B1* after a VE verb is that to express attitudinal, temporal, spatial, or manner information about a VE verb, a small word from the DE, Di, Ng, or T sets (such as *de, zhe, le, guo*, etc.) was attached to the verb, and this attachment and the close connection between the small word and the VE verb caused no need to pause at the juncture. However, the originally expected long pause (*B3/B4*) after the VE verb did not actually disappear; it was retained and only lagged behind to occur after the VE verb, for instance:

**Ex.10:** ... , 以(with, P11) 行動(action, Nad) 說明(prove, VE2) B1 了(Di) B3 他(he, Nhaa) 在乎(care, VK1) 妳(you, Nhaa) 的(DE) 感覺(feeling, Nac) 與(and, Caa) 期望(expectation, Nac) 。 ...

(...his actions have already proved that he cares about your feeling and expectation...)

## 5. Word Sets of Caa and Cb

Caa and Cb are two subcategories of Mandarin conjunctions, representing conjunctive conjunctions and correlative conjunctions [72], respectively. In the case of Caa, the arguments linked by the Caa conjunctions are words or phrases of identical syntactic categories and are usually associated in their meaning as in *feng<sub>N</sub> he<sub>Caa</sub> yu<sub>N</sub>* “風和雨 wind and rain,” *re<sub>VH</sub> hia-shi<sub>Caa</sub> leng<sub>VH</sub>* “熱還是冷 hot or cold,” *si<sub>Neu</sub> zhi<sub>Caa</sub> shi<sub>Neu</sub> sui<sub>Nf</sub>* “四至十歲 from four to ten years old,” and the like. Upon observation, we found that people usually tend to take a longer pause at pre-word boundary than at the post-word context, forming a sensible rhythmic variation and hence facilitating message delivery.



The statistics of the labeling results in Table 2.6 informs us that 90% of the Caa pre-boundary breaks were not shorter than *B2-2*, while, on the contrary, 98% of the post-word breaks were not longer than *B2-2*, a labeling outcome verifying our observation of the Caa words' neighboring breaks; that is, longer pauses tended to occur at the boundary between the preceding argument and the conjunction. The results matched some findings in Ref. 40. One example is given below for illustration:

**Ex.11:** ... ◦ 生活(life, Nad) 緊張(stress, VH21) B3 與(and, Caa) B1 結婚(marriage, Nv4) 延後(put off, VC2) 的(DE) 問題(problem, Nac) , ...  
 (...the problems of stressful life and postponed marriage....)

On the other hand, the Cb conjunctions function to join two clauses – a syntactic unit much larger than Caa's arguments – into a compound sentence, and therefore have higher potential to be preceded or followed by a PM in written texts to delimit the domain of a clause or a sentence; in read speech the occurrence of a PM elicits the announcer to take a longer pause to index a message transition or a piece of new message is coming. Our statistic results show that in the case of Cb conjunctions 80% of the pre-word boundaries and 20% of the post-word boundaries were labeled as *B3/B4*, which means much more PMs occurred before Cb conjunctions than afterward. One typical example is given below for demonstration:

**Ex.12:** ... , B4 因為(because, Cbaa) B2-1 學歷(academic credential, Nad) 並(Dbb) 非(not, VG2) 擇偶(choose spouse, VA4) 的(DE) 絕對(absolute, A) 條件(condition, Nac) ◦  
 (... , because the academic credentials are not absolute conditions of choosing spouse.)

Table 2.6: Statistics of break types labeled for word sets of Caa and Cb.

Labeled break type		<i>B0</i>	<i>B1</i>	<i>B2-1</i>	<i>B2-2</i>	<i>B3</i>	<i>B4</i>	Total count
Caa	Pre-boundary	5	32	1	127	214	26	405
	Post-boundary	52	104	157	85	7	0	405
Cb	Pre-boundary	61	46	23	39	168	512	849
	Post-boundary	135	284	166	95	150	19	849

## 6. Word set of *P*

*P* represents the class of Chinese prepositions, which precede a required argument and together play several semantic roles and indicate various relationships

such as time, location, tool, purpose, etc. Although Chinese Knowledge Information Processing (CKIP) categorizes prepositions into 65 types [73], only 13 types are active in the Sinica Treebank corpus. As for the adjacent pause of a preposition, it is reasonable to expect that due to the close connection of a preposition and its following argument, the pause at the post-word boundary tends to be short. For convenience of illustration, only *ba/jiang* “把,將” (labeled as P07) and *zai* “在” (labeled as P21), two typical and most frequently used prepositions, are selected out as the representative examples for discussion.

The statistic results in Table 2.7 show that on the whole for both *ba/jiang* “把,將” and *zai* “在” about 90% of the post-word boundaries were labeled as breaks no longer than *B2-1* (a break type caused by a pitch jump instead of lengthened pause duration), which indicates that the pauses at this juncture were either unperceivable or tending to be very short, again another confirmation of our model’s sound labeling job. Besides, a closer look at the distribution of break type percentages reveals that as high as 49% and 69% of the post-word breaks were *B2-1* for *ba/jiang* “把,將” and *zai* “在”, respectively. This statistics reflected our informant’s idiosyncratic style of articulating prepositional phrase; namely, besides leaving no pauses (Ex. 13), she often made a pitch jump between a preposition and the following argument to cause a sensible short pause (Ex. 14).

**Ex.13:** ,B4 把(P07) B1 孩子(children, Nab) 當做(regard as, VG1) 一塊(a piece of, DM) 璞玉(uncut jade, Nab) ,  
(...treat children as unpolished jade....)

**Ex.14:** 曾(at one time, Dd) 探詢(investigate, VE2) 他(he, Nhaa) B3 在(among, P21) B2-1 過去(past, Ndda) 眾多(a large number of, VH11) 的(DE) 著作(works, Nab) 中(among, Ng)  
(...has investigated that among a large number of works he wrote...)

On the other hand, as far as the labeling at the pre-word boundary is concerned, most labels were either *B1* or *B3/B4*; that is, 46% and 41% of the labels were *B3/B4* and 44% and 49% of them were *B1* for *ba/jiang* “把,將” and *zai* “在,” respectively, which suggests that our informant either took quite a long pause or just no pause at the pre-word position. To explain this phenomenon, further examination on the data containing these two prepositions revealed that the informant’s long breaks (*B3/B4*)

before a preposition were contributed by a left PM (Ex. 15), and in the remained cases she usually took no pause at this position (Ex. 16).

**Ex.15:** ... , *B4* 將(P07) *B2-1* 卷證(document, Nab) 移送(transfer, VC32) 桃園 (Taoyan, Nca) 地院(district court, Ncb) 審理(process, VC2) 。  
( ...transfer the documents to Taoyan District Court to process....)

**Ex.16:** , 協會(association, Nac) 就(an auxiliary confirming, Dd) 設(establish, VC33) *B1* 在(at, P21) *B1* 他(his, Nhaa) 家(home, Ncb) 。  
( ...The association is established at his home....)

Table 2.7: Statistics of break types labeled for word sets of P07 and P21.

Labeled break type		<i>B0</i>	<i>B1</i>	<i>B2-1</i>	<i>B2-2</i>	<i>B3</i>	<i>B4</i>	Total count
P07	Pre-boundary	0	39	0	9	32	9	89
	Post-boundary	8	38	34	9	0	0	89
P21	Pre-boundary	1	168	12	24	88	53	346
	Post-boundary	27	79	208	28	4	0	346

## 2.5.2 Analyses of Prosodic Constituents

Based on the break type labeling, we can divide the syllable sequence of each utterance into three types of prosodic constituents (i.e., PW, PPh, and BG/PG) to form a four-layer prosodic structure. Statistics in Table 2.8 shows that the average lengths for these three types of prosodic constituents are, respectively, 3.17 syllables or 1.85 lexical words (LWs) for PWs; 6.98 syllables, 4.02 LWs, or 1.69 PWs for PPhs; 16.69 syllables, 9.62 LWs, 4.07 PWs, or 1.94 PPhs for BG/PGs.

Table 2.8: Statistics of three types of prosodic constituents. Value in parentheses denotes standard deviation.

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

According to the histograms displayed in Figure 2.11, the length of each of these three prosodic constituents spans, respectively, from 1 to 12 syllables for PWs, from 1 to 33 syllables for PPhs, and from 1 to 99 syllables for BG/PGs. Besides, the

histograms also reveal that quite a few PPhs and BG/PGs, whose average lengths are supposed to be about 6.98 and 16.69 syllables, respectively, are nevertheless no longer than three syllables in length. Further investigation into these oddly short PPhs and BG/PGs indicates that the main reason lies in several special structure patterns of these constituents that require a long pause to highlight their prominence for successful information processing. First of all, in the case of short BG/PGs, defined as a sequence of syllables bounded by a *B4* on both sides, many of the particularly short BG/PGs actually consisted of a mono-syllabic subject and VE verb, which, as discussed in Subsection 2.5.1, due to its sentential object was tending to be followed by a long break up to *B4*; accordingly, bounded by a *B4* on both sides, the structure pattern of a subject plus a VE verb, both mono-syllabic in length, could generate as many short BG/PGs as possible.

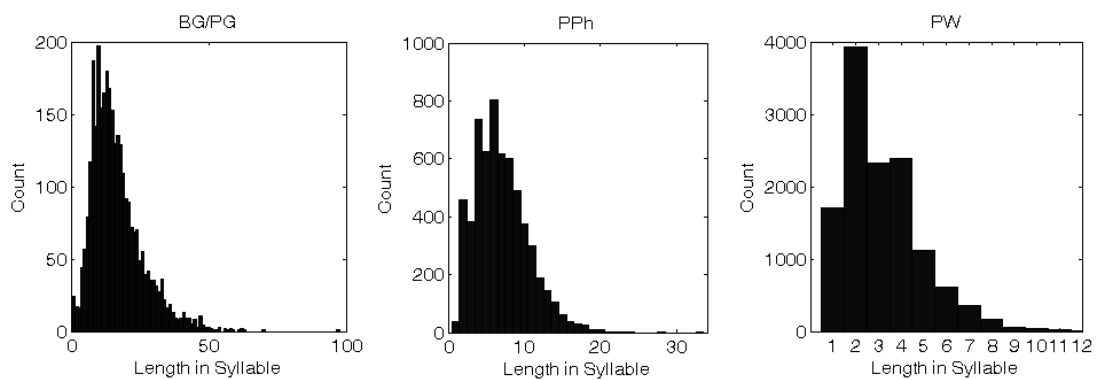


Figure 2.11: Histograms of lengths for BG/PG, PPh, and PW.

As for the cases of short PPhs, defined as a sequence of syllables delimited by (1) a *B3* at both sides or (2) a *B3* and a *B4* at each side, respectively, most of the *B3*s or *B4*s bounding the very short PPhs were actually caused by the existence of PMs that cued long pause duration. Table 2.9 shows the statistic results of the short PPh instances with respect to the existence of PMs at their two endings. As shown in the table, 66% of one-syllable PPhs were bounded by PMs on both sides, and most of them were numbers that were used to enumerate events. On the other hand, in the case of two- or three-syllable PPhs, on the whole about 84% of them were delimited at least by a left-sided PM, which means that the majority of these PPhs occurred at the beginning of a sentence. In terms of the internal structure of the two-syllable PPhs, 91% of them were bi-syllabic LWs functioned to express transitional relationships

like contrast, comparison, reinforcement, or addition. As for the three-syllable PPhs, the structure of them were either a topicalized tri-syllabic noun, or any phrasal structure composed of two smaller syntactic elements as in a subject-VE structure (*wo ren-wui* “我認為 I suppose”), a preposition-noun structure (由重慶 *you Chung-qing* “from Chung-qing”), a noun-localizer structure (*hun-zhan zhong* “混戰中 in the scuffle”), etc., and the long pauses adjacent to these PPhs were, on the informant’s part, strategies to cause prominent stress on these short phrases, and on the audience’s part, offered the listeners longer time to process and catch the information with least distortion.

Table 2.9: Count of short PPh instances with respect to the existence of PM at their two endings.

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

### 2.5.3 Pitch Patterns of Prosodic Constituents

We then explored the log- $F_0$  patterns of the three prosodic constituents of PW, PPh, and BG/PG. First, we extracted the prosodic state patterns from the observed pitch contour,  $\mathbf{sp}_n$ , by eliminating the influence of the current tone, the coarticulations from the two nearest neighboring tones, and the global mean, i.e.,

$$pm_n = \mathbf{sp}_n(1) - \beta_{t_n}(1) - \beta_{B_{n-1}, p_{n-1}}^f(1) - \beta_{B_n, p_n}^b(1) - \mu(1) \quad \text{for } 1 \leq n \leq N \quad (2.13)$$

where  $\mathbf{x}(1)$  denotes the first dimension of vector  $\mathbf{x}$ . A sequence of  $pm_n$  delimited by B2-1/B2-2/B3/B4 at both sides is regarded as a prosodic state pattern formed by integrating the log- $F_0$  mean patterns of the three prosodic constituents we considered. A model of prosodic state pattern is therefore defined by

$$pm_n = pm_n^r + \beta_{PW_n} + \beta_{PPh_n} + \beta_{BG/PG_n} \quad (2.14)$$

where  $pm_n^r$  is the residual of log- $F_0$  mean at syllable  $n$ ;  $\beta_{PW_n}$ ,  $\beta_{PPh_n}$  and  $\beta_{BG/PG_n}$

are the log- $F0$  patterns of PW, PPh and BG/PG, with  $PW_n=(i,j)$ ,  $PPh_n=(i,j)$  and  $BG/PG_n=(i,j)$  denoting that *syllable n* is located at the *j*th place of an *i*-syllable PW, PPh and BG/PG, respectively. The model was trained by a sequential optimization procedure. After well-training, the variances of  $\mathbf{sp}_n(1)$ ,  $pm_n$  and  $pm_n^r$  were  $565.4 \times 10^{-4}$ ,  $359.1 \times 10^{-4}$ , and  $191.2 \times 10^{-4}$ , respectively. Hence, the total residual error (TRE), which is the percentage of sum-squared residue over the observed sum-squared log- $F0$  mean, is about 33.8% by the current representation.

Figure 2.12 displays the patterns of  $\beta_{PW_n}$ ,  $\beta_{PPh_n}$ , and  $\beta_{BG/PG_n}$  with different lengths. It is noted that only the patterns calculated using more than 20 instances of prosodic state patterns are displayed because we want to know their general log- $F0$  patterns. It can be found from Figure 2.12(a) that all  $\beta_{BG/PG}$  had declining patterns with dynamic range spanning approximately from -0.1 to 0.1. Moreover, most of them had short ending resets. From Figure 2.12(b), we find that short  $\beta_{PPh}$  had rising-falling patterns, while long  $\beta_{PPh}$  had rising-falling-sustaining-falling patterns. Moreover, they had smaller dynamic range spanning approximately in [-0.07, 0.07]. Lastly, we find from Figure 2.12(c) that short  $\beta_{PW}$  showed high-falling patterns, while long  $\beta_{PW}$  showed falling-sustaining-falling patterns. Their dynamic range spanned approximately from -0.1 to 0.1.

From above analyses, we find that the prosodic-state tags possess rich information to represent the high-level prosodic constituents of the four-layer prosodic structure defined in this study. All these three types of log- $F0$  patterns generally agree with the findings of previous studies on intonation patterns of Mandarin speech [58,68,74,75]. The superposition patterns  $\beta_{PPh} + \beta_{BG/PG}$ , and all these three patterns ( $\beta_{PW}$ ,  $\beta_{PPh}$  and  $\beta_{BG/PG}$ ) resembled the intonation patterns reported in the studies of Tseng and co-workers [9,76-78] and the study of Chen *et al.* [36], respectively. Furthermore, with this prosodically meaningful finding, these quantitative prosodic constituent patterns combining with the APs of tone and coarticulation (i.e.,  $\beta_t$  and  $\beta_{B,tp}^f / \beta_{B,tp}^b$ ) can be used in Mandarin TTS to generate pitch contour if all break type can be properly predicted from the input text. However, due to the fact that the errors of the current representation are still high, a further

study to explore a more efficient representation is worthwhile doing in the future.

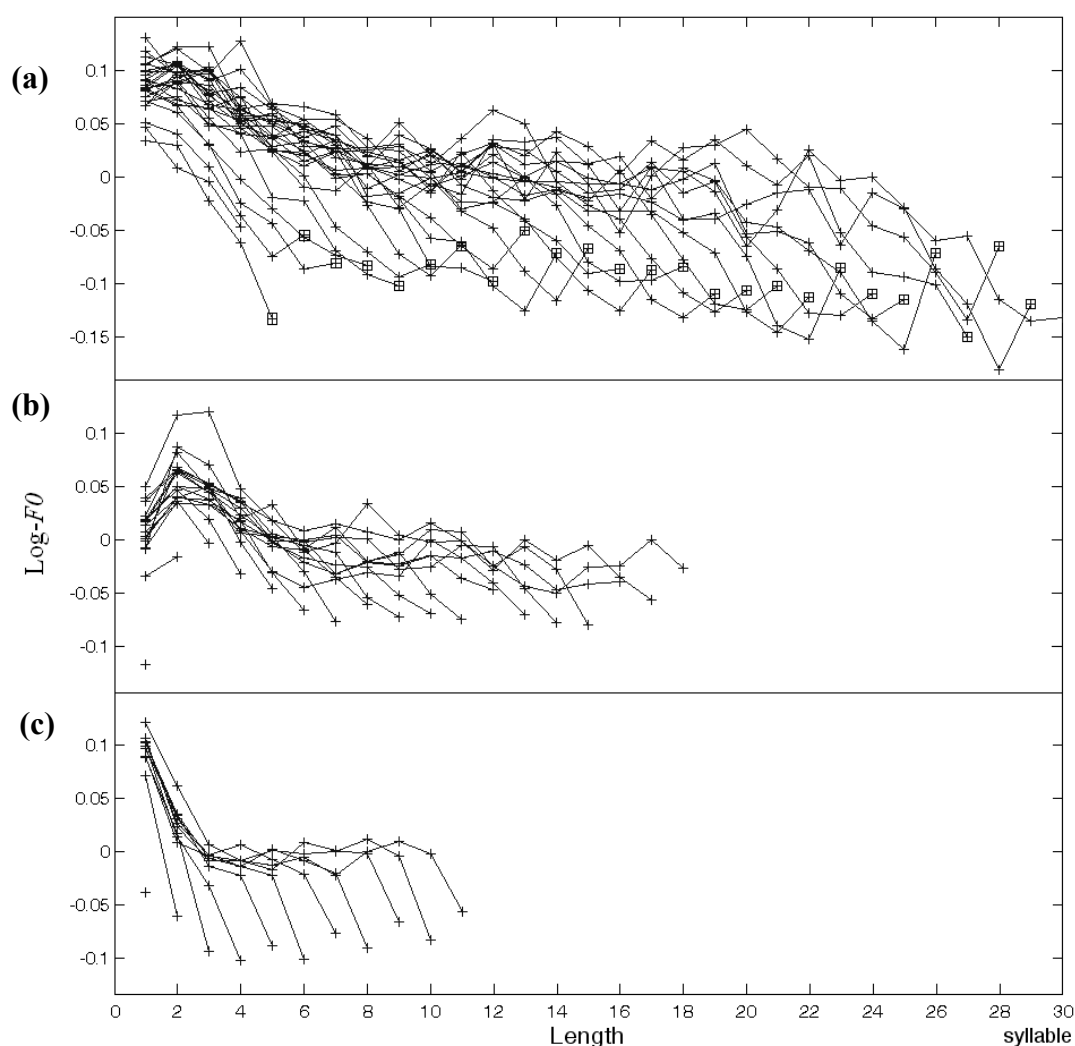


Figure 2.12: The log- $F_0$  patterns of (a) BG/PG, (b) PPh, and (c) PW. The special symbol “□” in (a) indicates the ending syllable of a log- $F_0$  pattern.

## 2.5.4 Comparison with Human Labeling

To further evaluate the performance of break labeling of the proposed method, a part of the Sinica Tree-Bank corpus used in this study was labeled cooperatively by two experienced labelers working in the Phonetics Laboratory, Department of Foreign Languages and Literatures of National Chiao Tung University. The annotated dataset consisted of 42 utterances with 5326 syllables. The labeling system used was a ToBI-like one developed by the laboratory, which represents the Mandarin speech prosody by a four-layer structure containing syllable, PW, intermediate phrase, and intonation phrase. These four prosodic constituents are delimited by four break types of  $b_1$ ,  $b_2$ ,  $b_3$ , and  $b_4$ , respectively. Here  $b_1$  represents an implicit non-break index,  $b_2$

is a perceivable break index for PW boundary,  $b_3$  is a minor-break index, and  $b_4$  is a major-break index.

Table 2.10 displays the correlation matrix of the break indices labeled by the two methods. It can be found from Table 2.10 that 97.8% of human-labeled  $b_4$ s, i.e., major breaks, were labeled as break indices of phrase or utterance boundaries (i.e.,  $B_3$ ,  $B_4$ , or  $B_e$ ) in our method, and 96.5% of  $b_1$ s, i.e., non-breaks, were labeled as indices of SYL boundaries within PW (i.e.,  $B_0$  or  $B_1$ ). This indicates that the two labeling methods were consistent for the two extreme cases of non-break and major break. It is also observed from the table that  $b_3$ s mainly (73.6%) corresponded to break indices  $\geq B_2-2$ , suggesting that the intermediate phrase boundaries in manual labeling, defined and perceived by the labelers as a minor-break, were, to quite a certain extent, consistently judged as a clearly perceived short pause ( $B_2-2$ ) or medium pause ( $B_3$ ) in our labeling. However, in the cases of  $b_2$ , 69.7% of them, defined as perceivable breaks, inconsistently corresponded to non-breaks ( $B_0$  or  $B_1$ ) in our scheme. To account for such inconsistency, a statistics on the internal morphological and syntactic structures of the PWs delimited by  $B_2$  and  $b_2$  shows that (1) while as high as nearly 69.3% of PW-LW correspondence occurred in the human labeling, 40.0% of such correspondence was found in our method, and (2) while 41.2% of the PWs labeled by our method was cases of compound words or long phrases composed of at least four syllables, only 2.2% of the PWs in the similar types was judged by the labelers. This significant discrepancy in the demarcation of PWs between these two methods suggests that labelers, though trained to listen to the prosodic cues with visual aids of graphic user interface to label the breaks, tended to subjectively treat LWs as PWs or as pronunciation units rather than objectively and exclusively relied on the actual prosodic features in prosodic labeling. This inclination obviously resulted in shorter average lengths of prosodic constituents in human labeling. Figure 2.13 displays the histograms of length of the prosodic constituents formed by the two labeling methods. It can be found from the figure that the average lengths of PWs, PPhs, and BG/PGs labeled by our method were indeed longer than human-labeled PWs, intermediate phrases, and intonational phrases, respectively.



Table 2.10: Correlations between unsupervised and human labeled breaks

		Human				
		<i>b1</i>	<i>b2</i>	<i>b3</i>	<i>b4</i>	total
Unsupervised						
<i>B0</i>		836	207	9	0	1052
<i>B1</i>		1970	726	70	0	2766
<i>B2-1</i>		81	313	53	1	448
<i>B2-2</i>		20	93	227	12	352
<i>B3</i>		0	0	137	260	397
<i>B4</i>		0	0	4	265	269
<i>B<sub>e</sub></i>		0	0	0	42	42
total		2907	1339	500	580	5326

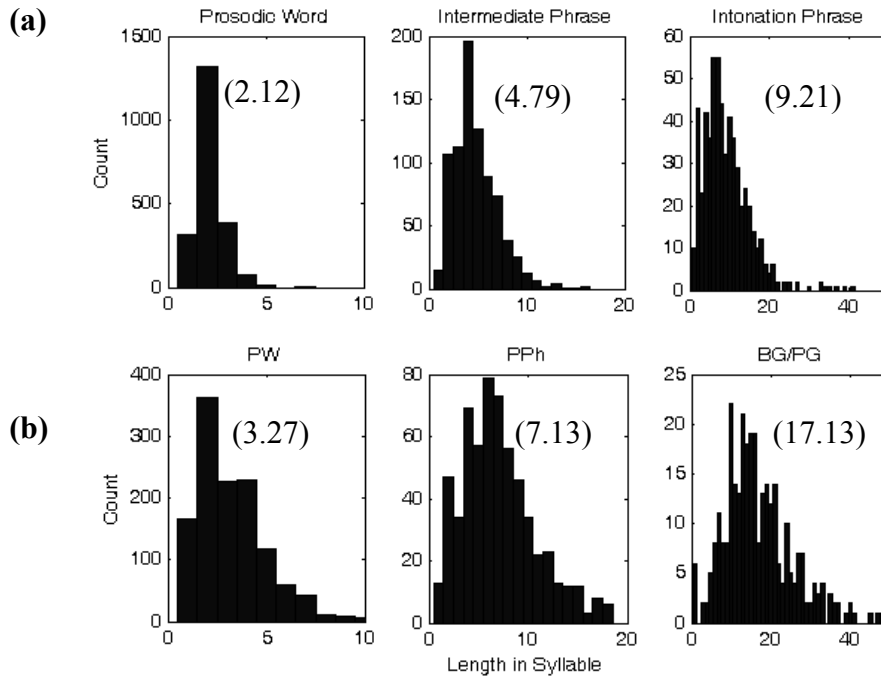


Figure 2.13: The histograms of length of the prosodic constituents formed by (a) the human labelers and (b) the proposed methods. The numbers in ( ) represent the average length of prosodic constituents.

From the perspective of prosodic features, it can be found from Figures 2.14(a) and 2.14(b) that the similar histograms of pause duration and normalized pitch jump in the same rows represented labeling consistency in our method, while the distinguishable histograms in the same columns expressed labeling inconsistency in

human labeling. Furthermore, Table 2.11 displays symmetric Kullback-Leibler distances (KL2) [79] for the two break labeling methods to measure the difference between two acoustic feature distributions that belong to different break indices labeled by the same method. It can be found from Table 2.11 that the KL2 distances for the proposed unsupervised method were generally greater than those of human labeling. Moreover, we find from Table 2.11(a) that the KL2 distances of pause duration were relatively large for all break index pairs of the proposed method except  $(B1, B2-1)$ ; nevertheless the KL2 distances of normalized pitch jump for  $(B1, B2-1)$  were large. On the contrary, we find from Table 2.11(b) that the KL2 distances of both acoustic features were low for  $(b1, b2)$  of human labeling. This confirms that the six break types  $B0$ - $B4$  in our labeling have distinct characteristics of acoustic features but the break types in human labeling have less discriminated ones. Specifically,  $B4$  has very large pause duration and significant pitch reset;  $B3$  has large pause duration and pitch reset,  $B2-2$  has medium pause duration,  $B2-1$  and  $B1$  have small pause duration but  $B2-1$  has significant pitch reset and  $B0$  has almost no pause duration. This property will be advantageous to our labeling method on those prosody modeling applications using acoustic features.



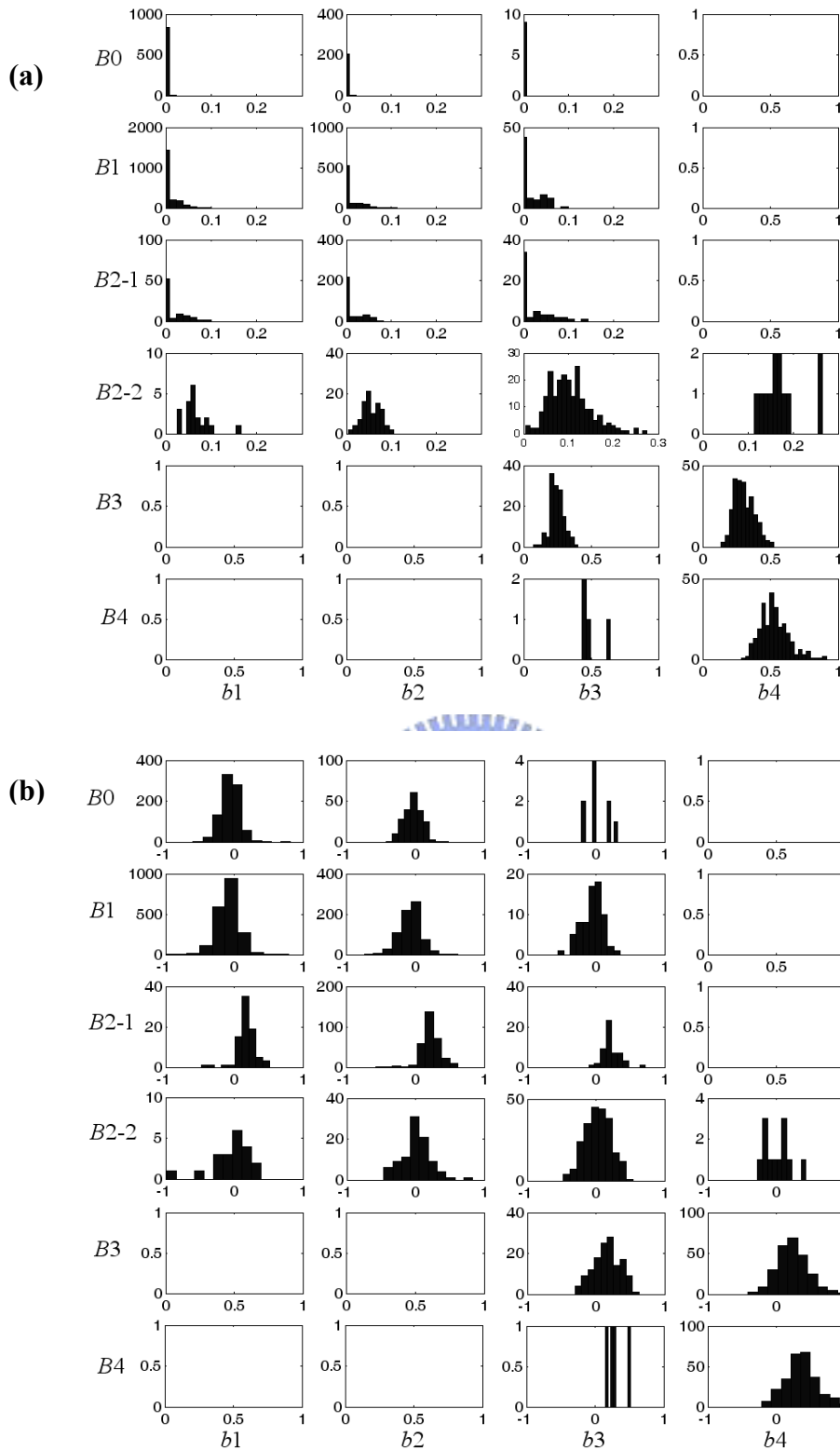


Figure 2.14: The histograms of (a) pause duration (in sec) and (b) normalized pitch jump (in log- $F_0$ ) for syllable-juncture instances belonging to sub-groups with different break-index pairs labeled by the two methods.

Table 2.11: Distances measuring the difference between two acoustic feature distributions that belong to different break indices labeled by the same method: (a) the proposed method, and (b) human labeling. Upper and lower triangular matrices represent KL2 distances for pause duration and normalized pitch jump, respectively.

(a)						
	<i>B0</i>	<i>B1</i>	<i>B2-1</i>	<i>B2-2</i>	<i>B3</i>	<i>B4</i>
<i>B0</i>		2.63	3.39	23.59	23.42	22.77
<i>B1</i>	0.19		0.16	14.21	23.28	22.66
<i>B2-1</i>	4.59	4.87		11.92	21.17	20.62
<i>B2-2</i>	0.52	0.72	2.79		13.84	18.85
<i>B3</i>	1.66	2.12	1.25	1.43		12.71
<i>B4</i>	3.69	4.18	0.36	2.50	0.88	

(b)					
	<i>b1</i>	<i>b2</i>	<i>b3</i>	<i>b4</i>	
<i>b1</i>			0.12	6.83	23.16
<i>b2</i>	0.24			6.07	22.10
<i>b3</i>	0.60	0.36			10.56
<i>b4</i>	2.05	1.20	0.82		

### 2.5.5 A Labeling Example

A typical example displaying the labeling results of the beginning part of a long utterance by the two methods is given in Figure 2.15. We first examined the labeling results of our method. From Figure 2.15(a), we find that the three PMs were labeled as two *B3* and one *B4*. One other *B3* without PM appeared at the right boundary of a nine-syllable NP. Besides, there existed five *B2-1* and four *B2-2*. They all appeared at inter-word junctures. We also find from Figure 2.15(b) that all three *B3* and five *B2-1* had clear normalized log-*F0* reset. Moreover, the curve of integrating APs of prosodic state and the global-mean of pitch level showed smoother PW patterns derived via removing the tone and coarticulation effects from the observed zigzag curve of log-*F0* mean. We then compared the results of the two labeling methods. It can be found from Figure 2.15(a) that aside from giving indices of breaks to all the above-mentioned breaks labeled by our method, human labelers gave four additional breaks to divide the nine-syllable-NP (行政院 主計處 的 統計 *xing-zheng-yuan zhu-ji-chu de tong-ji*) PW into three PWs, and the two four-syllable compound-word PWs, “進口 *jin-kou*(import) 金額 *jin-e*(the amount of money)” and “去年 *qu-nian*(last year) 同期 *tong-qi*(the same period)”, into four two-syllable words. To justify whether the deletions of these four human-labeled breaks were reasonable, we

examined the pause durations of these four word junctures and the normalized pitch patterns of the three integrated PWs. The pause durations were 12 ms, 40 ms, 22 ms, and 1ms. Obviously, they were all not significant. Besides, as seen in Figure 2.15(b) all the three normalized pitch patterns of none-syllable-NP PW and two four-syllable compound-word PWs were smooth. So the deletions of these four breaks by our method seemed reasonable.

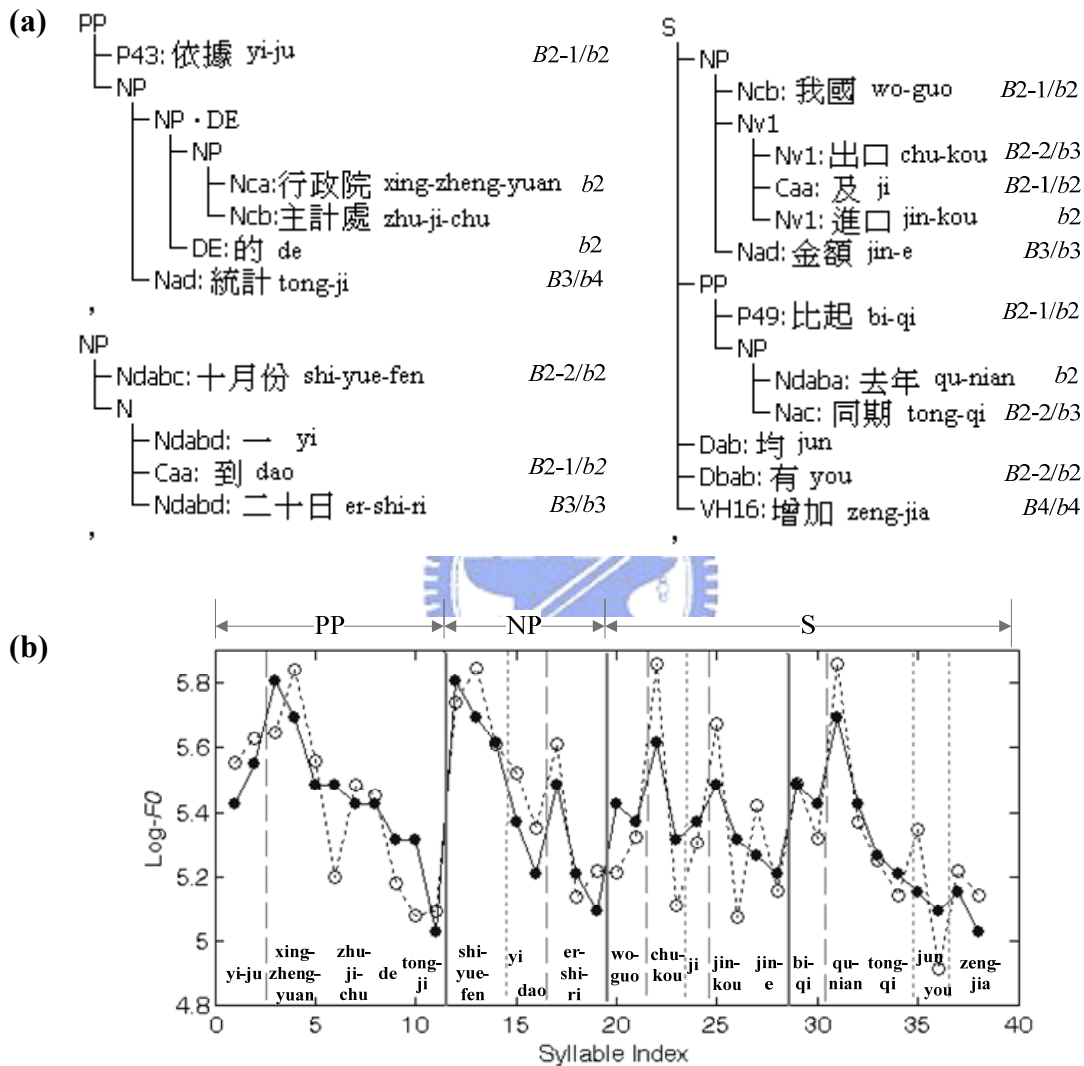


Figure 2.15: An example of the automatic prosody labeling. (a) Syntactic trees with prosodic tags: upper case *B* and lower case *b* for break-index labeled by our method and the human labeler, respectively; and (b) syllable log-*F*0 means: observed (open circle) and prosodic state+global mean (close circle). Solid/dash/dot lines represent *B*3/*B*2-1/*B*2-2 respectively. The utterance is “*yi-ju*(according to) *xing-zheng-yuan*(the Executive Yuan) *zhu-ji-chu*(Directorate-General of Budget, Accounting and Statistics) *de*(DE) *tong-ji*(statistics), *shi-yue-fen*(October) *yi*(1st) *dao*(to) *er-shi-ri*(20th), *wo-guo*(our country) *chu-kou*(export) *ji*(and) *jin-kou*(import) *jin-e*(the amount of money) *bi-qi*(in comparison with) *qu-nian*(last year) *tong-qi*(the same period) *jun*(both) *you*(to have some) *zeng-jia*(increase).”

## 2.6 Conclusions

In this chapter, a new approach of joint prosody labeling and modeling for Mandarin speech has been proposed. It first employed four prosodic models to describe the relationship of two types of prosodic tags to be labeled with the input acoustic prosodic features and linguistic features, and then used a sequential optimization procedure to determine all prosodic tags and estimate the parameters of the four prosodic models jointly using the Sinica Treebank speech corpus. Experimental results showed that the estimated parameters of the four prosodic models were able to penetratingly explore and appropriately describe the hierarchy of Mandarin prosody. First, the syllable pitch contour model was able to interpret the variation in syllable pitch contour controlled by such affecting factors as lexical tones, adjacent breaks, and prosodic state. Next, the prosodic state model was developed to clearly describe the declination effect of log- $F_0$  level within PW and the resets across PW, PPh, and BG/PG, and hence to extract the pitch patterns of each prosodic constituent. Then, the break-acoustics model could demonstrate the distinct acoustic characteristics for each of the six break types. The last model, the break-syntax model, was built to express the general relationship between the break type and the linguistic features of various levels. Besides, the performance of our models was further confirmed by the corresponding relationships found between the break indices labeled and their associated words which served as evidences to manifest the connections between prosodic and linguistic parameters, and it was also verified by our more consistent and discriminative prosodic feature distributions than those in human labeling by a quantitative comparison. In conclusion, the method we proposed to develop the joint prosody labeling and modeling for Mandarin speech was able to construct interpretive prosodic models and generate prosodic tags that were automatically and consistently labeled.

# Chapter 3 Advanced Unsupervised Joint Prosody Labeling and Modeling

## 3.1 Introduction

Motivated by the success of the unsupervised joint prosody labeling and modeling method (referred to as the UJPLM method hereafter) on the modeling of syllable pitch contour discussed in Chapter 2, we extend the study to include the other two important prosodic features, syllable duration and energy level, in this chapter. The study will jointly model syllable pitch contour, duration and energy level using the same method presented in the previous chapter. For simplicity, this extension is referred to as the advanced UJPLM (A-UJPLM) method.

## 3.2 The New Prosodic Model

In the A-UJPLM method, a new prosodic model to jointly consider the modeling of syllable pitch contour, duration, and energy level is first proposed. The new model considers more acoustic features, more prosodic tags, and more affecting factors. We discuss the new prosodic model in detail in the following subsections.

### 3.2.1 Features and Parameters Used in the New Prosodic Model

Aside from the prosodic features  $\mathbf{A}=\{\mathbf{sp},\mathbf{pd},\mathbf{ed}\}$  used in the previous study of pitch modeling, we consider more features in this study, including syllable duration sequence  $\mathbf{sd}$ , syllable energy-level sequence  $\mathbf{se}$ , normalized inter-syllable pitch jump sequence  $\mathbf{pj}$  defined by

$$pj_n = (\mathbf{sp}_{n+1}(1) - \beta_{t_{n+1}}(1)) - (\mathbf{sp}_n(1) - \beta_{t_n}(1)), \quad (3.1)$$

and normalized syllable duration lengthening factor sequences  $\mathbf{dl}$  and  $\mathbf{df}$  defined by

$$dl_n = (sd_n - \gamma_{t_n} - \gamma_{s_n}) - (sd_{n-1} - \gamma_{t_{n-1}} - \gamma_{s_{n-1}}) \quad (3.2)$$

and

$$df_n = (sd_n - \gamma_t - \gamma_s) - (sd_{n+1} - \gamma_{t_{n+1}} - \gamma_{s_{n+1}}), \quad (3.3)$$

where  $\gamma_t$  and  $\gamma_s$  represent respectively the syllable duration APs of tone and base-syllable type (to be defined latter). Hence, the acoustic feature set becomes  $\mathbf{A} = \{\mathbf{sp}, \mathbf{sd}, \mathbf{se}, \mathbf{pd}, \mathbf{ed}, \mathbf{pj}, \mathbf{dl}, \mathbf{df}\}$ . For a better presentation of these acoustic features, we divide them into three classes: syllable prosodic features  $\mathbf{X} = \{\mathbf{sp}, \mathbf{sd}, \mathbf{se}\}$ , inter-syllable prosodic features  $\mathbf{Y} = \{\mathbf{pd}, \mathbf{ed}\}$ , and differential syllable prosodic features  $\mathbf{Z} = \{\mathbf{pj}, \mathbf{dl}, \mathbf{df}\}$ .

As for prosodic tags, two new types of prosodic states, the duration prosodic state  $\mathbf{q}$  and the energy prosodic state  $\mathbf{r}$ , are introduced for the modeling of syllable duration and energy level to consider the effects contributed from high-level prosodic constituents of PW, PPh and PG/BG. Besides, a new break type  $B2-3$  is added to represent the syllabic boundary of  $B2$  with perceived lengthening of the preceding syllable. So, the complete prosodic tag set becomes  $\mathbf{T} = \{\mathbf{B}, \mathbf{PS}\}$ , where  $\mathbf{B} = \{B0, B1, B2-1, B2-2, B2-3, B3, B4\}$  is the break-type set and  $\mathbf{PS} = \{\mathbf{p}, \mathbf{q}, \mathbf{r}\}$  represents the prosodic-state tag set.

The linguistic features used in the new prosodic model are similar to those used in the old model with the following modifications. Firstly, in the syllable level, we separate base-syllable and final types from the linguistic feature  $\mathbf{L}$  because they are two important linguistic features, other than syllable tone, that seriously affect the variations of syllable duration and energy level. Secondly, all syntactic tree-level features are removed to consider that they can not be extracted reliably in practical applications. Lastly, two utterance-level normalization factors are added to consider respectively the variation in syllable duration due to the speaking rate and the variation in syllable energy level due to the recording volume. Hence, the linguistic feature  $\mathbf{L}$  is refined to include a syllable tone sequence  $\mathbf{t}$ , a base-syllable type sequence  $\mathbf{s}$ , a final type sequence  $\mathbf{f}$ , an utterance sequence  $\mathbf{u}$ , and a reduced linguistic feature set  $\mathbf{I}$ . To give a clearer picture of notations used in this study, we summarize them in Table 3.1.



Table 3.1: The notations of prosodic tags, prosodic features and linguistic features

<b>T</b> : prosodic tag	<b>B</b> : break type	<b>p</b> : pitch prosodic state
	<b>PS</b> : prosodic state	<b>q</b> : duration prosodic state
		<b>r</b> : energy prosodic state
<b>A</b> : prosodic feature	<b>X</b> : syllable prosodic feature	<b>sp</b> : syllable pitch contour
		<b>sd</b> : syllable duration
		<b>se</b> : syllable energy level
	<b>Y</b> : inter-syllabic prosodic feature	<b>pd</b> : pause duration
		<b>ed</b> : energy-dip level
	<b>Z</b> : differential prosodic features	<b>pj</b> : normalized pitch jump
		<b>dl</b> : normalized duration lengthening factor 1
		<b>df</b> : normalized duration lengthening factor 2
<b>L</b> : linguistic feature	<b>I</b> : reduced linguistic feature set	
	<b>t</b> : syllable tone sequence	
	<b>s</b> : base-syllable type sequence	
	<b>f</b> : final type sequence	
	<b>u</b> : utterance sequence	

### 3.2.2 Design of the New Prosodic Model

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

$$\begin{aligned}
 P(\mathbf{T}, \mathbf{A} | \mathbf{L}) &= P(\mathbf{A} | \mathbf{T}, \mathbf{L}) P(\mathbf{T} | \mathbf{L}) = P(\mathbf{X}, \mathbf{Y}, \mathbf{Z} | \mathbf{B}, \mathbf{PS}, \mathbf{L}) P(\mathbf{B}, \mathbf{PS} | \mathbf{L}) \\
 &\approx P(\mathbf{X} | \mathbf{B}, \mathbf{PS}, \mathbf{L}) P(\mathbf{Y}, \mathbf{Z} | \mathbf{B}, \mathbf{L}) P(\mathbf{PS} | \mathbf{B}) P(\mathbf{B} | \mathbf{L})
 \end{aligned} \tag{3.4}$$

where  $P(\mathbf{X} | \mathbf{B}, \mathbf{PS}, \mathbf{L})$  is a syllable prosodic model describing the variation in syllable prosodic features controlled by  $\mathbf{B}$ ,  $\mathbf{PS}$ , and  $\mathbf{L}$ ;  $P(\mathbf{Y}, \mathbf{Z} | \mathbf{B}, \mathbf{L})$  is a break-acoustic model describing the inter-syllable acoustic characteristics specified for different break type and surrounding linguistic features;  $P(\mathbf{PS} | \mathbf{B})$  is a prosodic state model describing the dynamics of prosodic states controlled by break types; and  $P(\mathbf{B} | \mathbf{L})$  is a break-syntax model describing the dependence of break occurrence on the surrounding linguistic features.

$P(\mathbf{X} | \mathbf{B}, \mathbf{PS}, \mathbf{L})$  is further elaborated by modeling syllable log- $F_0$  contour sequence  $\mathbf{sp}$ , syllable duration sequence  $\mathbf{sd}$ , and syllable energy level sequence  $\mathbf{se}$  separately, and assuming that their variations are controlled by five main affecting factors of lexical tone  $\mathbf{t}$ , base-syllable type  $\mathbf{s}$ , final type  $\mathbf{f}$ , utterance  $\mathbf{u}$ , prosodic state  $\mathbf{PS} = \{\mathbf{p}, \mathbf{q}, \mathbf{r}\}$ , and break  $\mathbf{B}$ , to obtain

$$\begin{aligned}
p(\mathbf{X}|\mathbf{B},\mathbf{P},\mathbf{S},\mathbf{L}) &\approx p(\mathbf{sp}|\mathbf{B},\mathbf{p},\mathbf{t})p(\mathbf{sd}|\mathbf{q},\mathbf{t},\mathbf{s},\mathbf{u})p(\mathbf{se}|\mathbf{r},\mathbf{t},\mathbf{f},\mathbf{u}) \\
&\approx \prod_{n=1}^N p(\mathbf{sp}_n|B_{n-1},p_n,t_{n-1}^{n+1}) \prod_{n=1}^N p(sd_n|q_n,t_n,s_n,u_n) \prod_{n=1}^N p(se_n|r_n,t_n,f_n,u_n)
\end{aligned} \tag{3.5}$$

where

$$p(\mathbf{sp}_n|B_{n-1},p_n,t_{n-1}^{n+1}) = N(\mathbf{sp}_n; \boldsymbol{\beta}_{t_n} + \boldsymbol{\beta}_{p_n} + \boldsymbol{\beta}_{B_{n-1},p_{n-1}}^f + \boldsymbol{\beta}_{B_n,p_n}^b + \boldsymbol{\mu}, \mathbf{R}) \tag{3.6}$$

models the variation in syllable log- $F0$  contour  $\mathbf{sp}_n$ , represented by the first four orthogonally-transformed parameters, with  $\boldsymbol{\beta}_{t_n}$ ,  $\boldsymbol{\beta}_{p_n}$ ,  $\boldsymbol{\beta}_{B_{n-1},p_{n-1}}^f$ , and  $\boldsymbol{\beta}_{B_n,p_n}^b$  denoting, respectively, the APs of tone  $t_n$ , the pitch prosodic state  $p_n$ , and the forward (carryover) and backward (anticipatory) coarticulations contributed from syllable  $n-1$  and syllable  $n+1$ , respectively, and  $\boldsymbol{\mu}$  and  $\mathbf{R}$  denoting the global mean and the covariance matrix of residual;

$$p(sd_n|q_n,t_n,s_n,u_n) = N(sd_n; \gamma_{t_n} + \gamma_{q_n} + \gamma_{s_n} + \gamma_{u_n} + \mu_d, R_d) \tag{3.7}$$

models the variation in syllable duration  $sd_n$  with  $\gamma$ 's denoting various APs, and  $\mu_d$  and  $R_d$  denoting the global mean and the variance of residual; and

$$p(se_n|r_n,t_n,f_n,u_n) = N(se_n; \alpha_{t_n} + \alpha_{r_n} + \alpha_{f_n} + \alpha_{u_n} + \mu_e, R_e) \tag{3.8}$$

models the variation in syllable energy level  $se_n$  with  $\alpha$ 's denoting various APs, and  $\mu_e$  and  $R_e$  denoting the global mean and the variance of residual.

The break-acoustic model  $P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L})$  is further elaborated by

$$P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L}) \approx P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{I}) \approx \prod_{n=1}^N p(pd_n, ed_n, pj_n, dl_n, df_n | B_n, \mathbf{I}_n) \tag{3.9}$$

where  $p(pd_n, ed_n, pj_n, dl_n, df_n | B_n, \mathbf{I}_n)$  is derived by the CART algorithm with the node splitting criterion of maximum likelihood gain. The CART algorithm jointly classifies the samples of pause duration  $pd_n$ , energy-dip level  $ed_n$ , normalized pitch jump  $pj_n$ , and normalized duration lengthening factors  $dl_n$  and  $df_n$  for each break type according to a question set derived from the contextual linguistic features  $\mathbf{I}_n$ . A joint  $pdf$  formed by the product of a gamma distribution for pause duration and four normal distributions for energy-dip level, normalized pitch jump, and the two duration lengthening factors is generated for each leaf node.

The prosodic state model  $p(\mathbf{PS}|\mathbf{B})$  is further divided into three sub-models for the three types of prosodic states and expressed by

$$p(\mathbf{PS}|\mathbf{B}) \approx p(\mathbf{p}|\mathbf{B})p(\mathbf{q}|\mathbf{B})p(\mathbf{r}|\mathbf{B}) \quad (3.10)$$

where  $p(\mathbf{p}|\mathbf{B})$ ,  $p(\mathbf{q}|\mathbf{B})$  and  $p(\mathbf{r}|\mathbf{B})$  are all represented by bigram models as

$$P(\mathbf{p}|\mathbf{B}) \approx P(p_1) \left[ \prod_{n=2}^N P(p_n | p_{n-1}, B_{n-1}) \right], \quad (3.11)$$

$$P(\mathbf{q}|\mathbf{B}) \approx P(q_1) \left[ \prod_{n=2}^N P(q_n | q_{n-1}, B_{n-1}) \right], \quad (3.12)$$

and

$$P(\mathbf{r}|\mathbf{B}) \approx P(r_1) \left[ \prod_{n=2}^N P(r_n | r_{n-1}, B_{n-1}) \right]. \quad (3.13)$$

The break-syntax model  $P(\mathbf{B}|\mathbf{L}) \approx P(\mathbf{B}|\mathbf{l})$  is elaborated in the same way as the old model discussed in Chapter 2 (see Eq. (2.11)).

### 3.3 Model Training by the A-UJPLM Method

Like the UJPLM method, the A-UJPLM method employs 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 the new prosodic model. It is divided into two main parts: initialization and iteration. The initialization part determines initial prosodic tags of all utterances and estimates initial parameters of the new prosodic model, which is composed of eight sub-models as discussed in Subsection 3.2.2, by a specially designed procedure. The iteration part first defines an objective likelihood function for each utterance by

$$Q = \left( \prod_{n=1}^N P(\mathbf{sp}_n | p_n, B_{n-1}, t_{n-1}^{n+1}) p(sd_n | q_n, t_n, s_n, u_n) p(se_n | r_n, t_n, f_n, u_n) \right) \left( P(p_1) P(q_1) P(r_1) \prod_{n=2}^N P(p_n | p_{n-1}, B_{n-1}) P(q_n | q_{n-1}, B_{n-1}) P(r_n | r_{n-1}, B_{n-1}) \right) \left( \prod_{n=1}^{N-1} (p(pd_n, ed_n, pj_n, dl_n, df_n | B_n, \mathbf{l}_n)) P(B_n | \mathbf{l}_n) \right). \quad (3.14)$$

It then applies a multi-step iterative procedure to update the labels of prosodic tags

and the parameters of the eight prosodic sub-models sequentially and iteratively. In the following subsections, we discuss the sequential optimization procedure in detail.

### 3.3.1 Initialization

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

#### (a) Initial labeling of break indices

The determination initial break index of each syllable juncture is similar to the method described in Chapter 2. As shown in Fig. 3.1, the decision rules for determining  $B4$ ,  $B3$ ,  $B2-2$  and  $B2-1$  are the same as the ones illustrated in Fig. 2.2. Non-PM inter-word junctures with apparent duration lengthening at the preceding syllable are likely labeled as  $B2-3$ . Intra-word junctures and non-PM inter-word junctures failed to be judged as  $B2-1$ ,  $B2-2$ , and  $B2-3$  are most likely to be labeled as  $B0$  or  $B1$ . The thresholds  $Th1 \sim Th6$  are set in the same way as those used in Chapter 2. The algorithms to determine  $Th7$  and  $Th8$  are given in Appendix A.

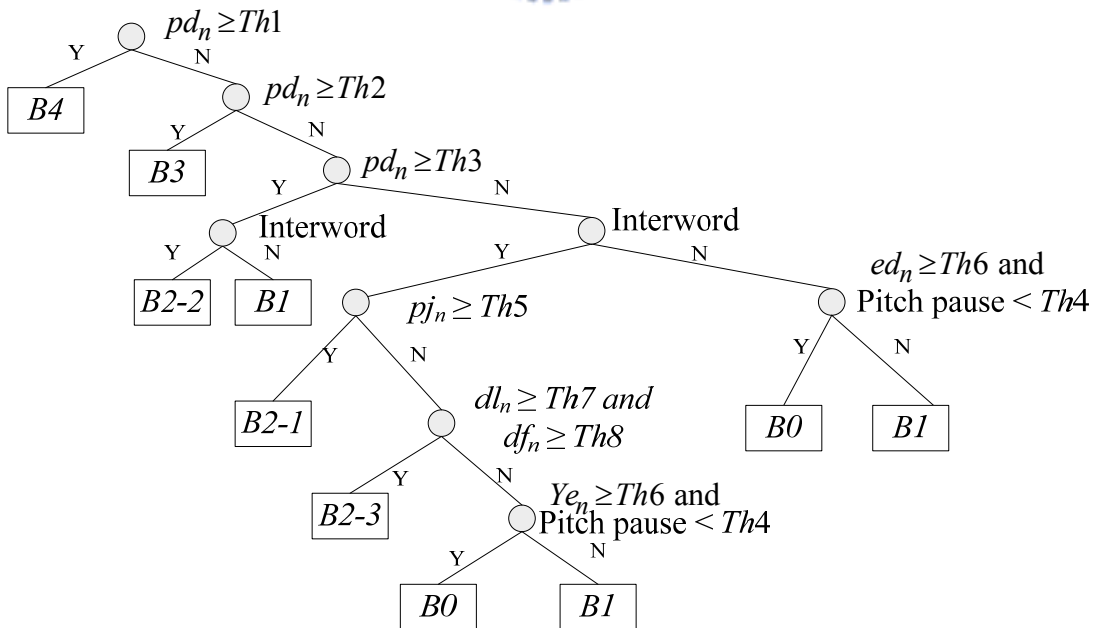


Figure 3.1: The decision tree for initial break type labeling.

## (b) Estimation of the initial parameters of the eight prosodic sub-models and prosodic-state indices

The initializations of the break-acoustics model and the break-syntax model can be done independently with initial break indices of all syllable junctures being given in previous step. We realize them by the CART algorithm with the node splitting criterion of maximum likelihood gain given the question sets  $\Theta_3$  and  $\Theta_4$ . The details of  $\Theta_3$  and  $\Theta_4$  used in this study are listed in Appendixes D.1 and D.2, respectively. The initializations of the syllable pitch contour, duration and energy level models and prosodic-state indices are integrated together and performed by a progressive estimation procedure. A progressive estimation strategy is adopted to first determine the initial APs which can be estimated most reliably and then eliminate their effects from the surface syllable prosodic features for the estimations of the remaining APs. In this study, the order of initial AP estimation is listed as follows: global mean  $\{\boldsymbol{\mu}, \mu_d, \mu_e\}$ , utterance  $\{\gamma_u, \alpha_u\}$ , five tones  $\{\boldsymbol{\beta}_t, \gamma_t, \alpha_t\}$ , base-syllable and final types  $\{\gamma_s, \alpha_f\}$ , coarticulation  $\{\boldsymbol{\beta}_{B,tp}^f, \boldsymbol{\beta}_{B,tp}^b, \boldsymbol{\beta}_{B_e,t_1}^f, \boldsymbol{\beta}_{B_e,t_N}^b\}$ , and prosodic states  $\{\boldsymbol{\beta}_p, \gamma_q, \alpha_r\}$ . Notice that the initial prosodic-state indices are assigned by performing VQs separately to the three features of pitch-level components of the residue pitch contours, the residual syllable durations and the residual syllable energy levels. The APs are set to be the corresponding codewords. Lastly, the initialization of the prosodic state sub-models  $P(\mathbf{p}|\mathbf{B})$ ,  $P(\mathbf{q}|\mathbf{B})$  and  $P(\mathbf{r}|\mathbf{B})$  are done using the labeled prosodic-state indices and break indices.

### 3.3.2 Iteration

The iteration is a multi-step iterative procedure listed below.

- Step 1:* Update the APs of utterance  $\gamma_u$  and  $\alpha_u$  with all other APs being fixed.
- Step 2:* Update the APs of five tones  $\boldsymbol{\beta}_t$ ,  $\gamma_t$  and  $\alpha_t$  with all other APs being fixed.
- Step 3:* Update the APs of coarticulation  $\{\boldsymbol{\beta}_{B,tp}^f, \boldsymbol{\beta}_{B,tp}^b, \boldsymbol{\beta}_{B_e,t_1}^f$  and  $\boldsymbol{\beta}_{B_e,t_N}^b\}$  with all other APs being fixed, and then update  $\mathbf{R}$ .

- Step 4:* Update the APs of base-syllable type and final,  $\gamma_s$  and  $\alpha_f$  with all other APs being fixed, and then update  $R_d$  and  $R_e$ .
- Step 5:* Re-label the prosodic state sequence of each utterance by using the Viterbi algorithm so as to maximize  $Q$  defined in Eq. (3.14). Then, update the APs of prosodic states  $\beta_p$ ,  $\gamma_q$  and  $\alpha_r$ , the prosodic state sub-models  $P(\mathbf{p}|\mathbf{B})$ ,  $P(\mathbf{q}|\mathbf{B})$ ,  $P(\mathbf{r}|\mathbf{B})$ ,  $\mathbf{R}$ ,  $R_d$  and  $R_e$ .
- Step 6:* Re-label the break type sequence of each utterance by using the Viterbi algorithm so as to maximize  $Q$ . Then, update  $P(\mathbf{p}|\mathbf{B})$ ,  $P(\mathbf{q}|\mathbf{B})$ ,  $P(\mathbf{r}|\mathbf{B})$ ,  $\mathbf{R}$ ,  $R_d$  and  $R_e$ .
- Step 7:* Re-construct the decision trees to update  $p(pd_n, ed_n, pj_n, dl_n, df_n | B_n, \mathbf{I}_n)$  and  $P(B_n | \mathbf{I}_n)$  by the CART algorithm using the question sets  $\Theta_1$  and  $\Theta_2$ , respectively.
- Step 8:* Repeat *Steps 1* to *7* until a convergence is reached.

### 3.4 Experimental Results

The same Treebank database was used to evaluate the A-UJPLM method. The numbers of the three types of prosodic state were all empirically set to 16. As shown in Figure 3.2, the sequential optimization procedure took 109 iterations to reach a convergence. Following is examinations and interpretations of the parameters of the 8 prosodic sub-models.

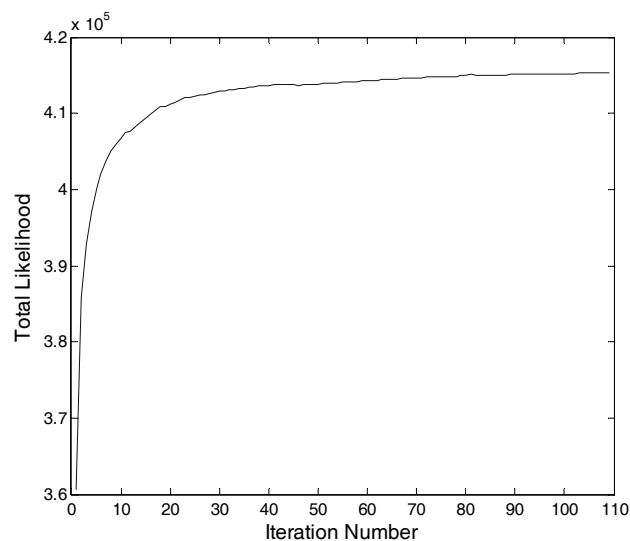


Figure 3.2: The plot of total log-likelihood versus iteration number.

### 3.3.1 The Syllable Prosodic Model

We first examined the parameters of the syllable prosodic model  $p(\mathbf{X}|\mathbf{B},\mathbf{PS},L)$ . The covariance matrices/variances of the original and residual syllable log- $F0$  contour, duration and energy level are shown below:

$$\mathbf{R}_{sp} = \begin{bmatrix} 565.4 & 23.9 & -25.6 & -0.5 \\ 23.9 & 90.5 & 9.7 & -8.2 \\ -25.6 & 9.7 & 17.8 & -0.9 \\ -0.5 & -8.2 & -0.9 & 5.0 \end{bmatrix} \times 10^{-4} \Rightarrow \mathbf{R}_{sp^r} = \begin{bmatrix} 3.8 & 0.2 & -0.2 & 0.0 \\ 0.2 & 31.9 & 2.6 & -1.5 \\ -0.2 & 2.6 & 11.1 & 0.6 \\ 0.0 & -1.5 & 0.6 & 3.7 \end{bmatrix} \times 10^{-4}$$

$$R_{sd} = 382.1 \times 10^{-5} \Rightarrow R_{sd^r} = 3.7 \times 10^{-5}$$

$$R_{se} = 48.43 \Rightarrow R_{se^r} = 0.26$$

Obviously, all components of covariance and variances of the three residuals were much smaller than their counterparts of the original features. This showed that the influences of the affecting factors considered were indeed essential to the variation of **sp**, **sd** and **se**.

Table 3.2 displays the APs of five tones. These results generally agreed with those of previous studies [58,59,67,68].

Table 3.2: APs of five tones

Tone	1	2	3	4	5
Pitch mean	0.153	-0.080	-0.175	0.088	-0.145
Syllable duration	0.012	0.015	-0.008	-0.001	-0.075
Energy level	0.367	-1.015	-1.272	1.500	-1.940

Figure 3.3 displays the decision tree analysis of the duration APs of base-syllable type. It can be found from the figure that the syllables with initial in {b, d, g} are much shorter in average than other combinations of initial-final. Generally, syllables with initial in {q, ch, c, f, h, x, sh, s, p, t, k} are longer while syllables with final of single vowel are shorter. The results generally confirmed to those of previous studies [59]. The decision tree analysis of energy-level APs of final type is shown in Figure 3.4. It can be seen from the figure that the average energy level, from large to small, are those of open, mid and close vowels. Besides, the energy level of final with medial is generally smaller than others.

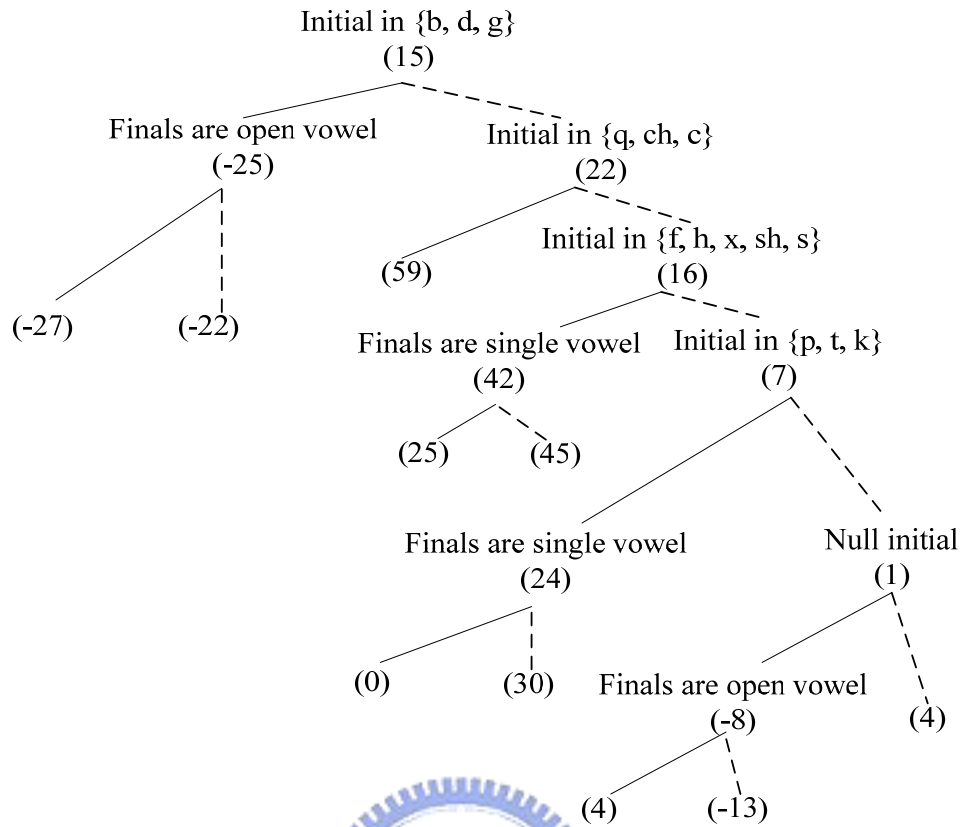


Figure 3.3: Decision tree analysis of duration APs of base-syllable type. Number in () represents the average length (ms) of the APs in the leaf node. Solid line indicates positive answer to the question and dashed line indicates negative answer.

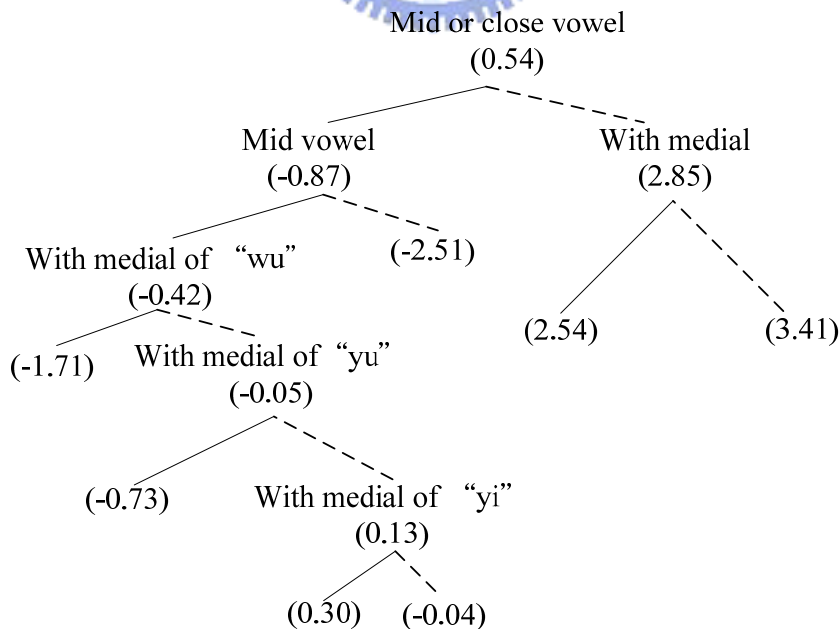


Figure 3.4: Decision tree analysis of energy-level APs of final. Number in () represents the average energy level (dB) of the APs in the leaf node. Solid line indicates positive answer to the question and dashed line indicates negative answer.



Table 3.3 displays the APs of the pitch, duration and energy prosodic states. It can be seen from Figure 3.5 that, for each of the three prosodic features, the APs of 16 prosodic states spanned widely to cover the whole dynamic range.

Table 3.3: APs of prosodic states

p/q/r	1	2	3	4	5	6	7	8
$\beta_p(1)$	-0.87	-0.58	-0.42	-0.33	-0.26	-0.20	-0.14	-0.09
$\gamma_q$	-0.12	-0.09	-0.08	-0.06	-0.05	-0.03	-0.02	-0.01
$\alpha_r$	-18.49	-13.25	-10.50	-8.40	-6.57	-4.96	-3.47	-2.12
p/q/r	9	10	11	12	13	14	15	16
$\beta_p(1)$	-0.03	0.03	0.09	0.15	0.21	0.28	0.37	0.48
$\gamma_q$	0.00	0.02	0.03	0.05	0.07	0.09	0.12	0.17
$\alpha_r$	-0.80	0.58	1.98	3.46	5.05	6.82	9.03	12.15

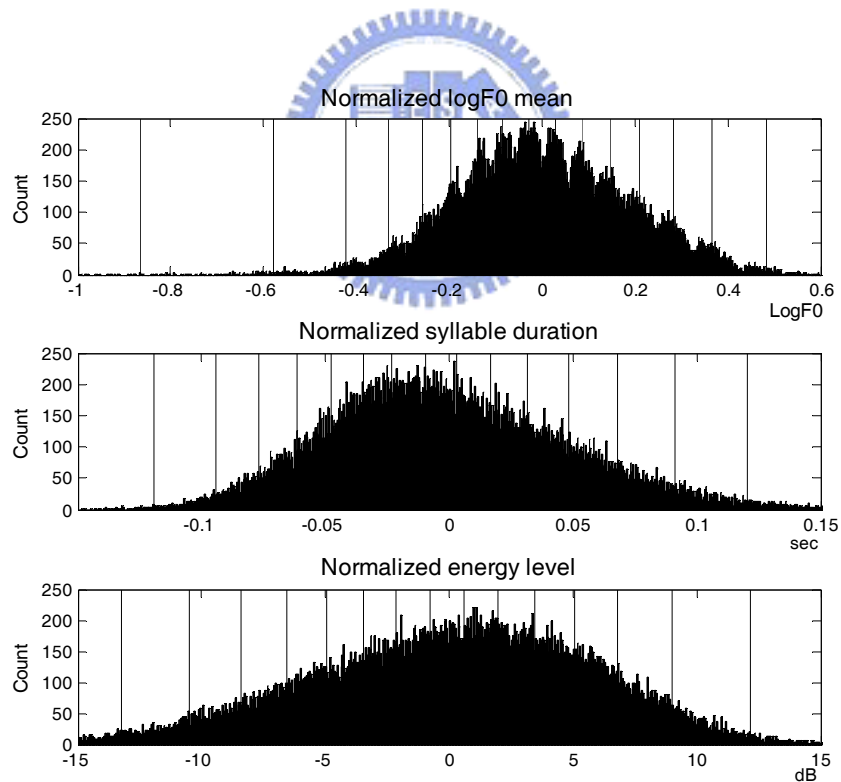


Figure 3.5: Distributions of normalized prosodic features and the APs of prosodic states (vertical lines).

Table 3.4 displays the total residual errors (TREs) of the prosodic modelings for syllable pitch contour, duration, and energy level with respect to the use of different combinations of APs. It can be seen from the table that the TREs reduced as more APs were considered and the most significant one is prosodic state. This result suggested that higher-level prosodic constituents (i.e., PW, PPh and PG/BG) may account for great amount of prosodic variations. More detail analysis of prosodic state will be given in Subsection 3.3.3.

Table 3.4: TREs of the prosodic modelings for syllable pitch contour, duration and energy level w.r.t. the use of different combinations of affecting factors.

Pitch		Duration		Energy level	
APs	TRE	APs	TRE	APs	TRE
		+ Utterance	98.8%	+ Utterance	77.8%
+ Tone	71.6%	+ Tone	88.1%	+ Tone	74.5%
+ Coarticulation	60.3%	+ Base-syllable	62.9%	+ Final	48.0%
+ Prosodic state	1.1%	+ Prosodic state	1.1%	+ Prosodic state	1.0%

### 3.3.2 The Break-Acoustics Model

Figure 3.6 displays the distributions of pause duration, energy-dip level, normalized pitch jump, and normalized duration lengthening factors for the root nodes of these seven break types. As can be seen from the figure, the break types of higher level were generally associated with longer pause duration, lower energy-dip level, greater normalized pitch jump, and larger duration lengthening factors. The distributions of pause duration and energy-dip level were similar to those obtained in the previous study shown in Fig. 2.5. Notice that *B2-3* was similar to *B1* and *B2-1* in the distributions of pause duration, and energy-dip level. *B2-1*, *B2-2*, *B3*, and *B4* had positive normalized pitch jumps in average while *B0*, *B1*, and *B2-3* had negative ones. This result illustrated the declination and reset effects of  $\log-F_0$  at intra-PW and inter-PW syllable boundaries, respectively. Normalized duration lengthening factors of *B2-2*, *B2-3*, *B3*, and *B4* were relatively larger than those of *B0*, *B1*, and *B2-1*. These distributions showed the lengthening effect for the last syllable of PW, PPh, and PG/BG.

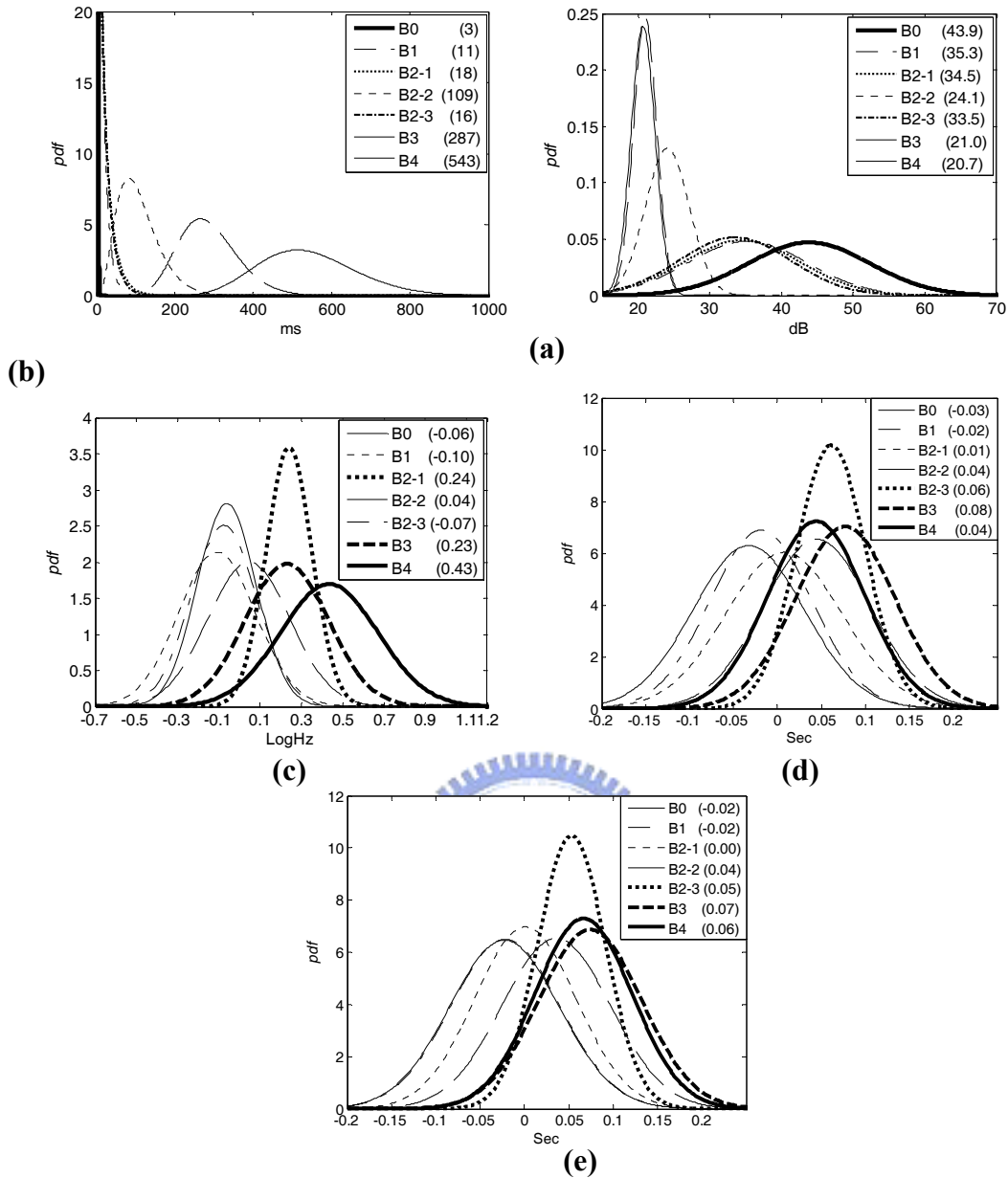


Figure 3.6: The *pdfs* of (a) pause duration, (b) energy-dip level for the root nodes, (c) normalized pitch jump, (d) normalized duration lengthening factor 1 and (e) normalized duration lengthening factor 2 of these seven break types. Numbers in () denote the mean values.

### 3.3.3 The Prosodic State Model

Figure 3.7 displays some most significant transitions of pitch prosodic state  $P(p_n | p_{n-1}, B_{n-1})$  for seven break types. It can be found that the prosodic state transitions of  $B0$ ,  $B1$ ,  $B2-1$ ,  $B2-2$ ,  $B3$  and  $B4$  generally agree with the results illustrated in Subsection 2.4.3. The transition of  $B2-3$  is similar to those of  $B0$  and  $B1$ . This implies no apparent pitch reset exists at the duration-lengthening juncture of  $B2-3$ .

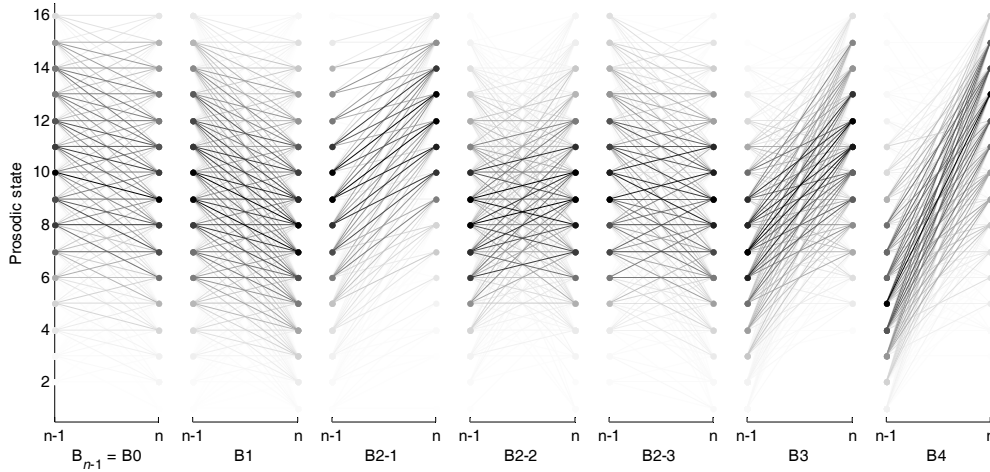


Figure 3.7: The most significant pitch prosodic state transitions,  $P(p_n | p_{n-1}, B_{n-1})$ , for each break types. Notice that the darker lines represent the more primary prosodic state transitions.

Figure 3.8 illustrates the transitions of duration prosodic state  $P(q_n | q_{n-1}, B_{n-1})$ . Generally, larger break types made more significant high-to-low transitions. It can be observed from the transitions of  $B3$  and  $B4$  that PPhs and PG/BGs usually begun with lower states and ended with higher states to manifest the significant duration lengthening effect before major break junctures. Compared with the transitions of  $B3$  and  $B4$ , those of  $B2-2$  and  $B2-3$  had less high-to-low dynamics implying less syllable duration lengthening before minor break junctures. As for  $B0$ ,  $B1$  and  $B2-1$ , they had small nearby-state transitions without preferred direction.

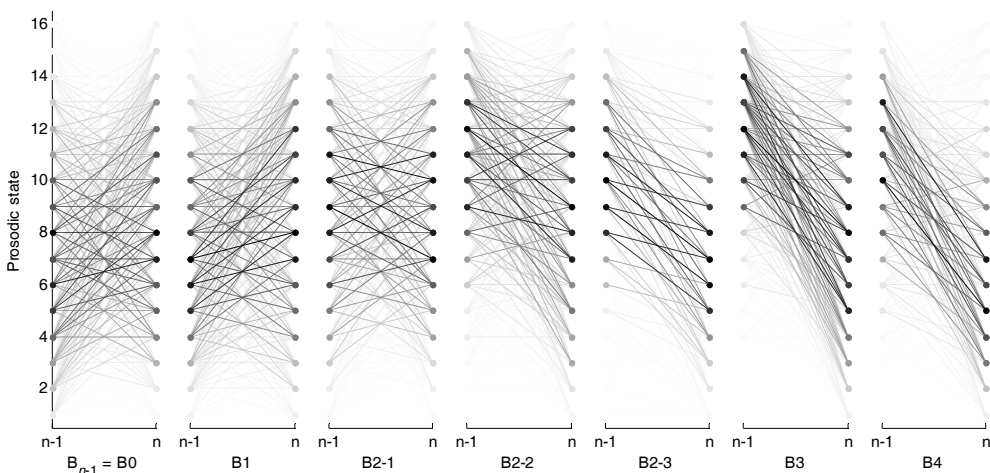


Figure 3.8: The most significant duration prosodic state transitions,  $P(q_n | q_{n-1}, B_{n-1})$ , for each break types. Notice that the darker lines represent the more primary prosodic state transitions.

The energy prosodic state transitions are shown in Figure 3.9. Apparently, low-to-high transitions were primarily found in major breaks (i.e.,  $B3$  and  $B4$ ), while high-to-low and level transitions were mainly observed in non-break and minor break. These results demonstrated the declination of energy level within a PPh or PG/BG, and the reset when restarted a PPh or PG/BG.

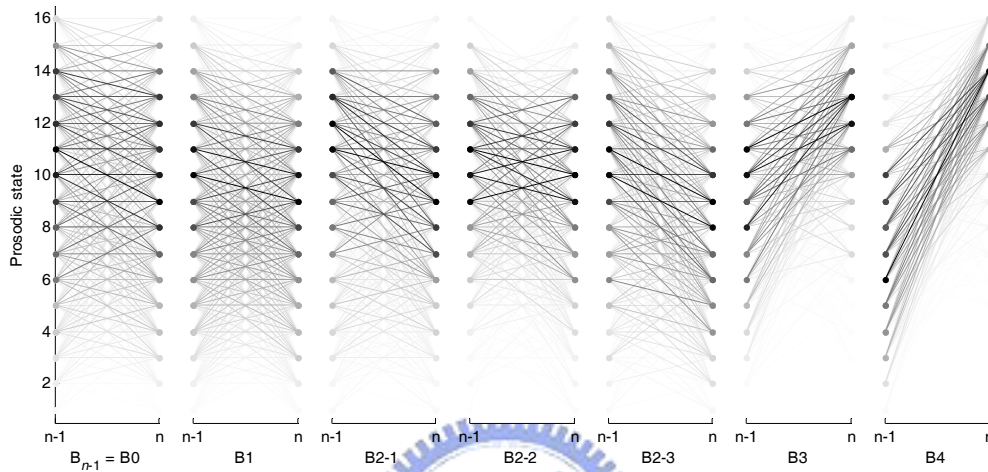


Figure 3.9: The most significant energy prosodic state transitions,  $P(q_n | q_{n-1}, B_{n-1})$ , for each break types. Notice that the darker lines represent the more primary prosodic state transitions.

### 3.3.4 The Break-Syntax Model

Figure 3.10 displays the decision tree of the break-syntax model. It can be seen from the figure that the root nodes of the two sub-trees  $T3$  and  $T4$ , which corresponded to syllable juncture with PM, were mainly composed of major break types of  $B3$  and  $B4$ .  $T4$  contained more  $B4$  because it corresponded to major PMs.  $T6$  which corresponded to intra-word was mainly composed of non-break.  $T5$  had much more complex tree structure than other sub-trees. By further analyzing the entropies of the leaf nodes in sub-trees  $T3$ -6, we find that  $T6$  had the largest entropy. This implies that it is more difficult to correctly predict the break types of non-PM inter-word junctures.

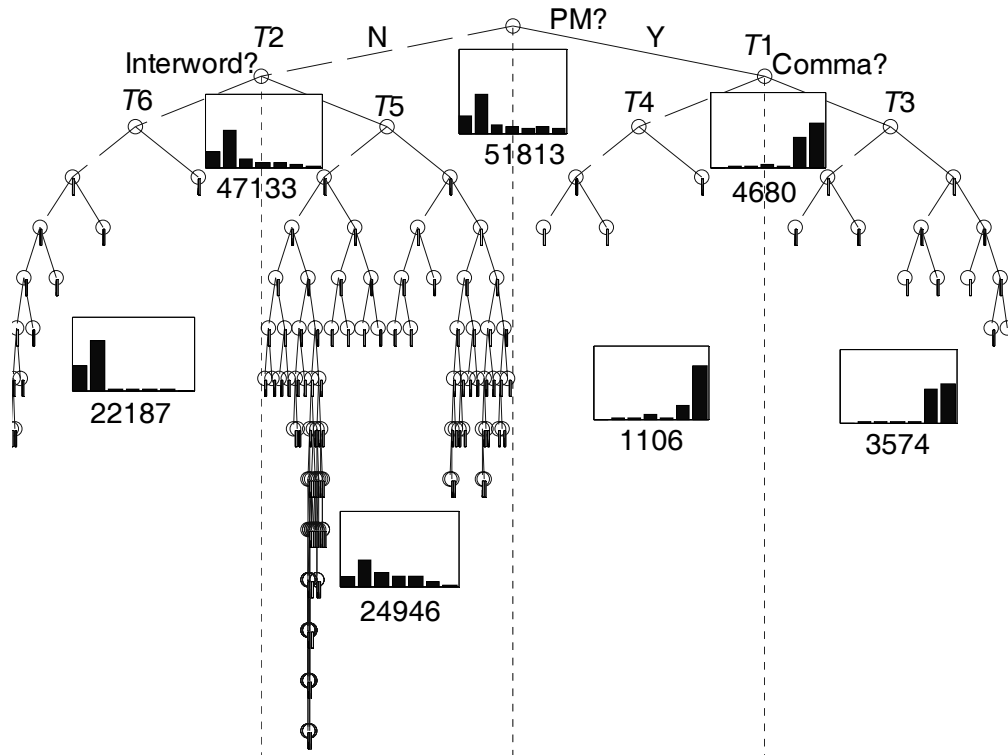


Figure 3.10: The decision tree of the break-syntax model. The bar plot associated with a node denotes the distributions of these six break types ( $B_0$ ,  $B_1$ ,  $B_{2-1}$ ,  $B_{2-2}$ ,  $B_3$ ,  $B_4$ , from left to right) and the number is the total sample count of the node.

More detailed structures of these four sub-trees up to the fourth layer are shown in Figure 3.11. It is found from Figures 3.11(a) and (b) that nodes in  $T_3$  and  $T_4$  were mainly split by questions related to sentence-level linguistic features such as  $LFS \geq 7$  (Is the length of the following sentence equal to or greater than 7?). Generally, the juncture of PM was more likely to be  $B_4$  when the previous/following sentence was long. It is also found from Figure 3.11(b) that the minor PM “dun hao” (or “、”) was likely to be labeled as  $B_3$  and  $B_{2-2}$  other than  $B_4$ . We find from Figure 3.11(c) that in  $T_6$  Type-2 intra-word junctures, which are anticipated as potential break positions, were more likely to be minor breaks than Type-1 intra-word junctures which were simply labeled as  $B_0$  and  $B_1$ . For the most complex sub-tree  $T_5$  (see Figure 3.11(d)), the labeling of non-PM inter-word juncture could be firstly discriminated as tending to  $B_0$  or  $B_1$  by the initial type of the following syllable {null initial, m, n, l, r}. The junctures with non-sonorant initial could be further discriminated as non-break by the following word with POS “DE”. This result matched with the previous finding presented in Subsection 2.5.1. It was also found that the distance to previous PM ( $DPP \geq 2$ ,  $DDP \geq 11$ ,  $DFP \geq 3$ ) and the distance to next PM ( $DFP \geq 3$ ) were used to

discriminate other sub-trees of  $T5$ . Generally speaking, a non-PM inter-word juncture had higher potential to be labeled as minor breaks as its distance to the nearby PM was longer.

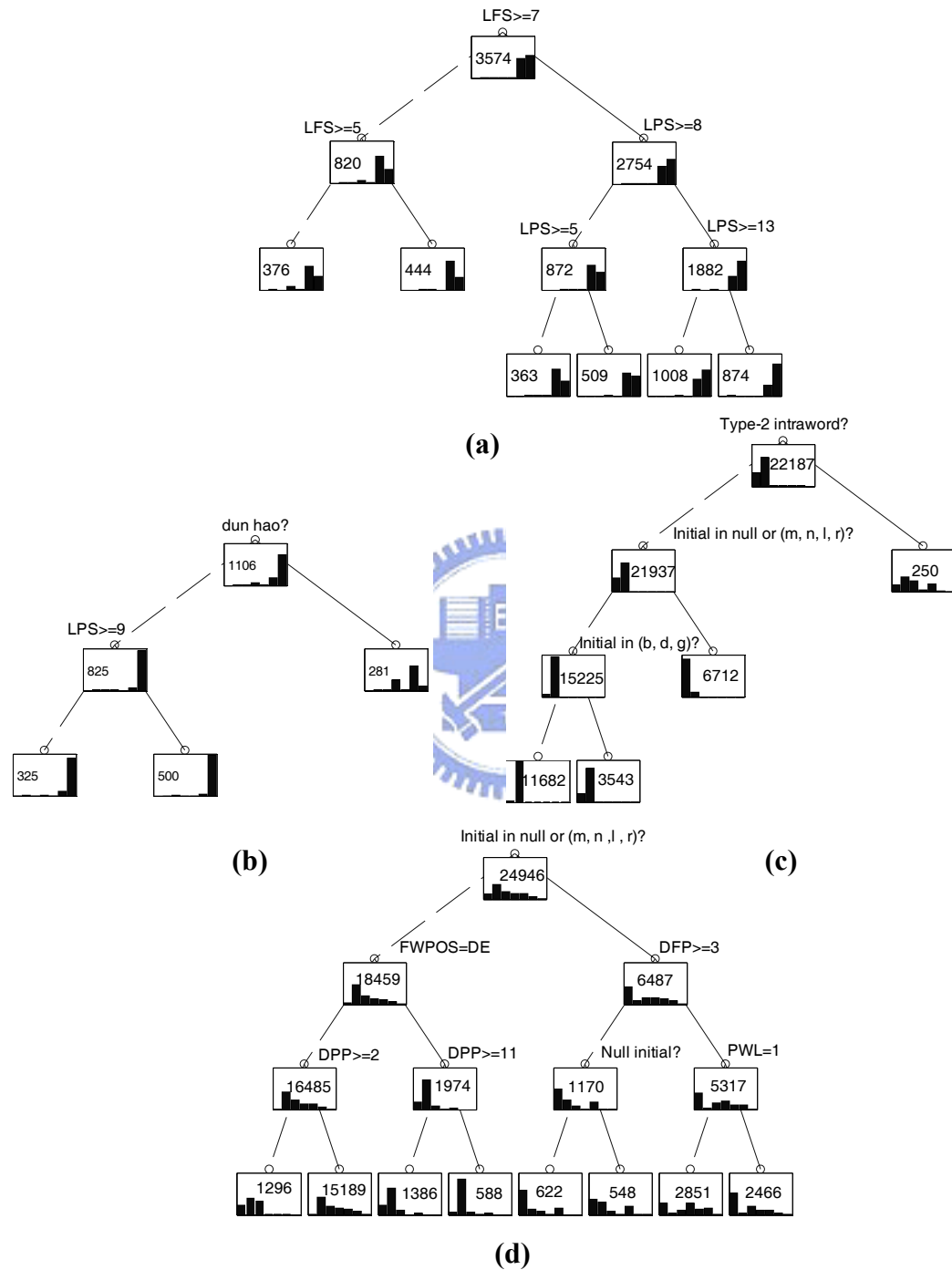


Figure 3.11: The more detailed structures of sub-trees of (a)  $T3$ , (b)  $T4$ , (c)  $T6$  and (d)  $T5$ . Solid line indicates positive answer to the question and dashed line indicates negative answer.

## 3.4 Analyses of the Labeled Breaks and Prosodic Constituents

### 3.4.1 Comparison Between A-UJPLM and UJPLM

Table 3.5 shows the cooccurrence matrix for the break types labeled by A-UJPLM and by UJPLM. Some findings from the table are discussed as follows. Firstly, the labeling results of these two methods were roughly consistent to each other. This is especially true for non-breaks and major breaks. Secondly, about 83% of the new break type  $B2-3$  labeled by A-UJPLM corresponded to the tags of  $B0$  and  $B1$  labeled by UJPLM. This implies A-UJPLM inserted more minor breaks than UJPLM by re-labeling non-breaks into  $B2-3$ . Thirdly, by more detailed analysis we find that the two most inconsistent pairs ( $B2-2, B3$ ) and ( $B1, B2-1$ ) were owing to the similarities in the distributions of their inter-syllable acoustic features.

Table 3.5: Cooccurrence matrix for the break types labeled by A-UJPLM and by UJPLM

	UJPLM	$B0$	$B1$	$B2-1$	$B2-2$	$B3$	$B4$	Count
Advanced UJPLM								
$B0$	<b>82.7</b>	16.3	1.0	0	0	0	0	10311
$B1$	3.0	<b>94.3</b>	1.1	1.6	0	0	0	23892
$B2-1$	3.8	11.6	<b>76.2</b>	8.3	0.1	0	0	4812
$B2-2$	0	3.9	2.7	<b>76.4</b>	17.0	0	0	3464
$B2-3$	<b>16.4</b>	<b>66.8</b>	5.8	10.9	0	0	0	3065
$B3$	0	0	0.1	3.2	<b>87.7</b>	9.0	0	3549
$B4$	0	0	0	0	10.6	<b>89.3</b>	0	2720

Figure 3.12 displays the histograms of length for the three high-level prosodic constituents of BG/PG, PPh and PW. Compared with Fig. 2.10, we find that both histograms of BG/PG and PPh looked similar for these two methods, while the histogram of PW for A-UPJLM shrank significantly. Table 3.6 shows the statistics of length for the three prosodic constituents. As can be seen from the table, the average length of PW (2.8 syllables) was shorter than that of UJPLM (3.17 syllables, see Table 2.8) due to the insertions of  $B2-3$ s. The average length of PPh (7.46 syllables)



was longer than that of UJPLM (6.98 syllables) due to the substitutions of  $B_3$ s with  $B_2$ -2s. The average length of PG/BG (16.85 syllables) was slightly longer than that of UJPLM (16.69 syllables).

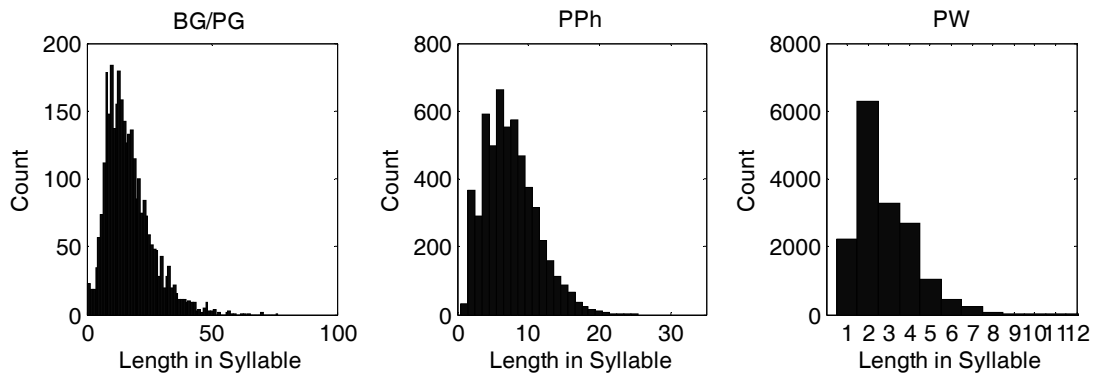


Figure 3.12: Histograms of lengths for BG/PG, PPh and PW.

Table 3.6: Statistics of three types of prosodic constituents. Value in parentheses denotes standard deviation.

Average length in	Prosodic constituent		
	PW	PPh	BG/PG
Syllable	2.80(1.40)	7.46(3.66)	16.85(9.38)
LW	1.64(0.84)	4.30(2.29)	9.75(5.44)
PW	1.00	2.31(1.79)	5.28(3.38)
PPh	X	1.00	1.77(1.66)

Table 3.7 displays the cooccurrence matrix of the break indices labeled by A-UJPLM and by human. Basically, the cooccurrence matrix is similar to that of UJPLM shown in Table 2.10. It is found that 94.7% of human-labeled  $b_4$ s, i.e. major breaks, were labeled as break indices of phrase or utterance boundaries (i.e.  $B_3$ ,  $B_4$  or  $B_e$ ) in A-UJPLM; and 94.4% of  $b_1$ s, i.e. non-breaks, were labeled as indices of SYL boundaries within PW (i.e.  $B_0$  or  $B_1$ ). It is also observed that  $b_3$ s still corresponded mainly to break indices of short or medium pause, i.e.  $B_2$ -2 and  $B_3$ . The most significant difference lies in the tags of  $b_2$ . It is found that many  $b_2$ s were inconsistently labeled to non-break tags of  $B_0$  and  $B_1$ . By adding the new break type  $B_2$ -3, A-UJPLM reduced the inconsistency rate from 69.7% for UJPLM to 53.5%.

Table 3.7: Cooccurrence matrix of break tags labeled by A-UJPLM and human

Human						
Unsupervised	<i>b1</i>	<i>b2</i>	<i>b3</i>	<i>b4</i>	total	
<i>B0</i>	884	167	5	0	1056	
<i>B1</i>	1860	549	31	1	2441	
<i>B2-1</i>	81	361	66	1	509	
<i>B2-2</i>	17	54	212	29	312	
<i>B2-3</i>	65	208	52	0	325	
<i>B3</i>	0	0	129	242	371	
<i>B4</i>	0	0	5	265	270	
<i>B<sub>e</sub></i>	0	0	0	42	42	
Total	2907	1339	500	580	5326	

### 3.4.2 Patterns of Prosodic Constituents

To explore the general patterns of syllable pitch contour, duration, and energy level for high-level prosodic constituents of PW, PPh, and BG/PG, we first extract the prosodic state patterns from the observed syllable prosodic features (i.e.  $\mathbf{sp}_n$ ,  $sd_n$  and  $se_n$ ) by eliminating the influences of utterance, current tone, coarticulations from the two nearest neighboring tones, base-syllable type, final type, and the global mean, i.e.,

$$\mathbf{pm}_n = \mathbf{sp}_n - \beta_{t_n} - \beta_{B_{n-1}, t_{p_{n-1}}}^f - \beta_{B_n, t_{p_n}}^b - \mu \quad \text{for } 1 \leq n \leq N \quad (3.15)$$

$$dm_n = sd_n - \gamma_{u_n} - \gamma_{t_n} - \gamma_{s_n} - \mu_d \quad \text{for } 1 \leq n \leq N \quad (3.16)$$

$$em_n = se_n - \alpha_{u_n} - \alpha_{t_n} - \alpha_{f_n} - \mu_e \quad \text{for } 1 \leq n \leq N \quad (3.17)$$

Sequences of  $\mathbf{pm}_n$ ,  $dm_n$  and  $em_n$  delimited by *B2-1/B2-2/B3/B4* at both sides are regarded as prosodic state patterns formed by integrating the log-F0/syllable duration/energy level patterns of the three prosodic constituents we considered. Three superposition models for prosodic state patterns are therefore defined by

$$\mathbf{pm}_n = \mathbf{pm}_n^f + \beta_{PW_n} + \beta_{PPh_n} + \beta_{BG/PG_n} \quad (3.18)$$

$$dm_n = dm_n^f + \gamma_{PW_n} + \gamma_{PPh_n} + \gamma_{BG/PG_n} \quad (3.19)$$

$$em_n = em_n^f + \alpha_{PW_n} + \alpha_{PPh_n} + \alpha_{BG/PG_n} \quad (3.20)$$

where  $\mathbf{pm}_n^r$ ,  $dm_n^r$  and  $em_n^r$  are respectively the residuals of log- $F_0$ , syllable duration and syllable energy level at *syllable*  $n$ ;  $\beta_x$ ,  $\gamma_x$  and  $\alpha_x$  represent APs of affecting factor  $x$  for log- $F_0$ , syllable duration and syllable energy level, respectively;  $PW_n=(i,j)$ ,  $PPh_n=(i,j)$ , and  $BG/PG_n=(i,j)$  denote that *syllable*  $n$  is located at the  $j$ th place of an  $i$ -syllable PW, PPh, and BG/PG, respectively. A sequential optimization procedure based on the MMSE criterion is adopted to train these three models. The error functions for each utterance are defined by

$$E_p = \sum_{n=1}^N \left| \mathbf{pm}_n - \beta_{PW_n} - \beta_{PPh_n} - \beta_{BG/PG_n} \right|^2 \quad (3.21)$$

$$E_d = \sum_{n=1}^N \left( dm_n - \gamma_{PW_n} - \gamma_{PPh_n} - \gamma_{BG/PG_n} \right)^2 \quad (3.22)$$

$$E_e = \sum_{n=1}^N \left( em_n - \alpha_{PW_n} - \alpha_{PPh_n} - \alpha_{BG/PG_n} \right)^2 \quad (3.23)$$

Then, with proper initializations, it sequentially updates the patterns of PW, PPh and BG/PG to minimize  $E_p/E_d/E_e$  until a convergence is reached.

Figures 3.13, 3.14, and 3.15 display, respectively, the general patterns of pitch level, duration and energy level for PW, PPh and PG/BG with different lengths. It is noted that the patterns with more instances are displayed in darker lines and dots. As shown in Figure 3.13, these log- $F_0$  patterns matched the results of the previous study shown in Figure 2.11. It can be clearly observed from Figure 3.14 that the last syllables of all duration patterns of PPh and PW were lengthened significantly, while those of most BG/PG duration patterns were shortened. Interestingly, the shortening of the antepenultimate syllable in PPh, which is an important feature of tempo structure in Mandarin Chinese, is also found. These phenomena completely matched with the findings of Tseng [8]. From Figure 3.15, we find that both short  $\alpha_{BG/PG}$  and  $\alpha_{PPh}$  had falling patterns, while long  $\alpha_{BG/PG}$  and  $\alpha_{PPh}$  had, respectively, falling-sustaining-falling and falling-sustaining patterns. Compared with  $\alpha_{BG/PG}$  and  $\alpha_{PPh}$ ,  $\alpha_{PW}$  is more flat and had smaller dynamic range. It is worth to note that the last syllables of all energy level patterns had small resets illustrating a special stress style of Mandarin speech.

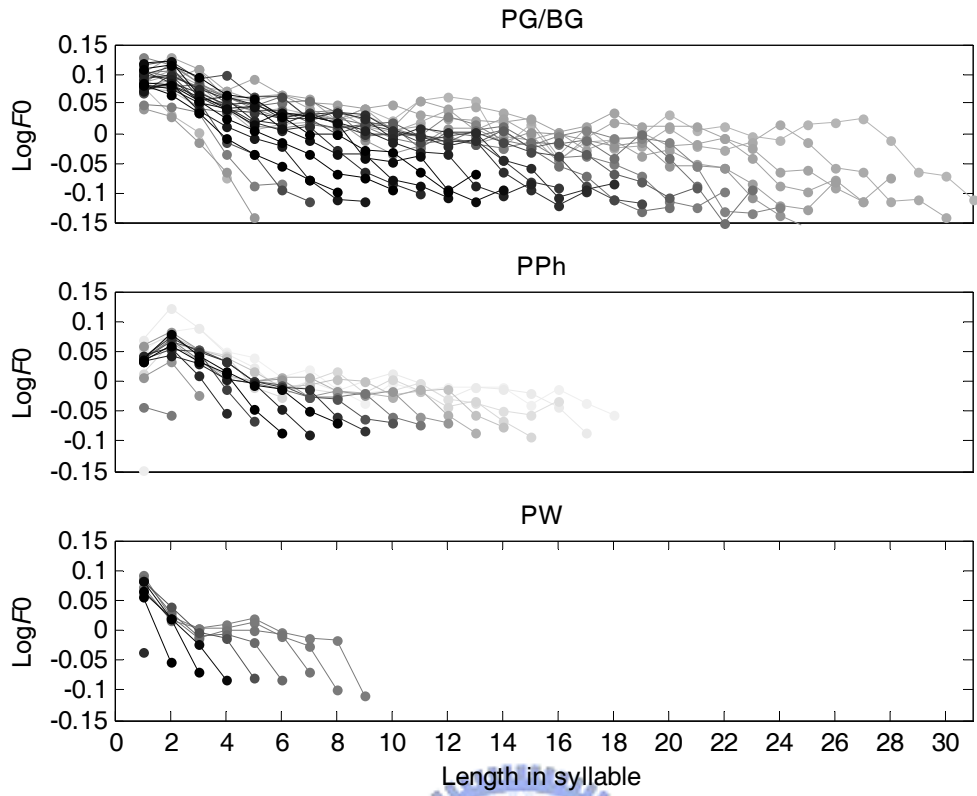


Figure 3.13: The log- $F_0$  patterns of BG/PG, PPh and PW.

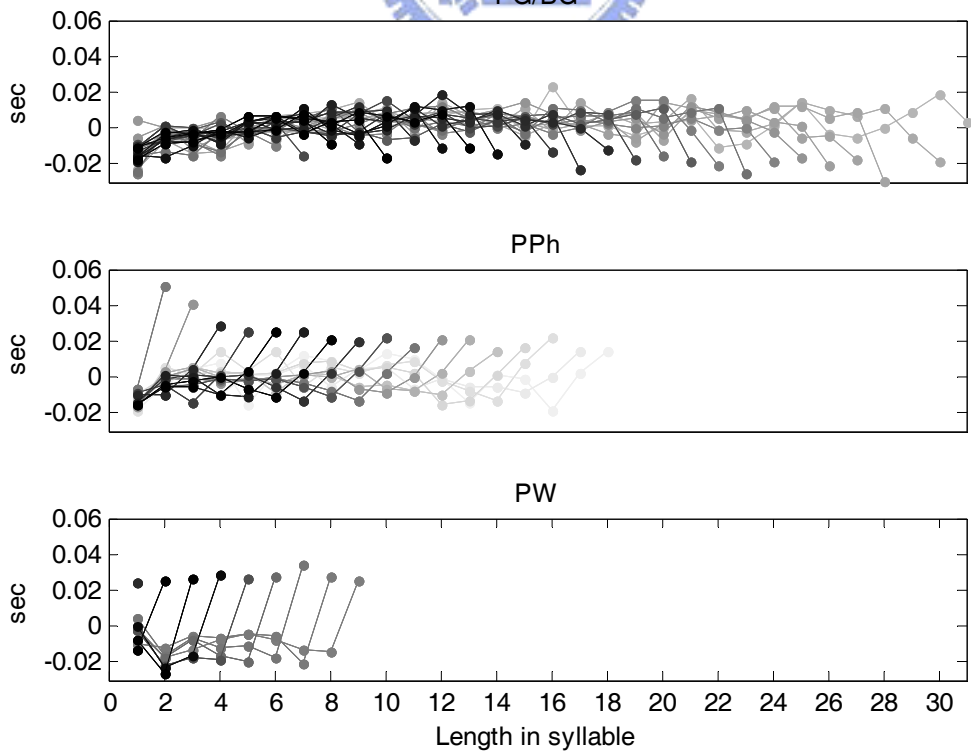


Figure 3.14: The duration patterns of BG/PG, PPh and PW.

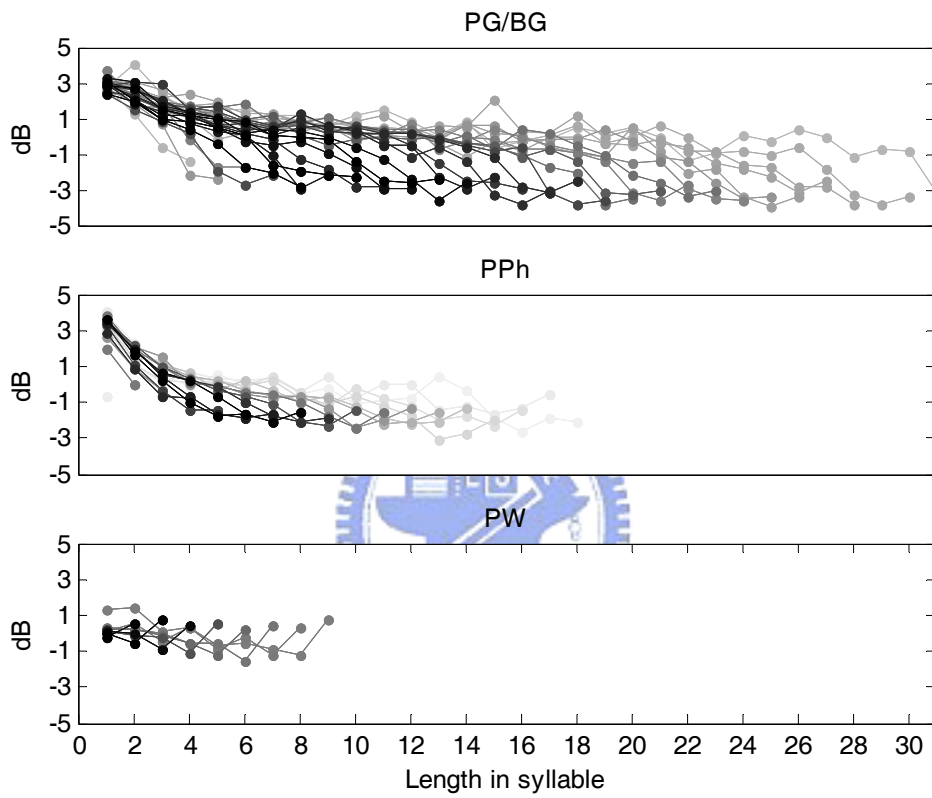


Figure 3.15: The energy level patterns of BG/PG, PPh and PW.

Table 3.8 displays the TREs of the prosodic modeling results for syllable pitch contour, duration, and energy level with respect to different combinations of affecting factors. It can be found from the table that TREs reduced as more affecting factors were used. The low-level affecting factors/linguistic features (i.e., utterance, tone, coarticulation, base-syllable, and final type) accounted for 39.7%, 37.1%, and 50.0% of prosodic variation in pitch, duration, and energy level, respectively; while the three high-level prosodic constituents (i.e. PW+ PPh + BG/PG) contributed another 22.9% (60.3% - 37.4%), 20.8%, and 22.0 % of prosodic variation in pitch, duration, and energy, respectively. Among the three high-level prosodic constituents, we find that the most significant one is PW for both pitch and duration, and PPh for energy level. However, the TREs are still high. More sophisticated representations of PW, PPh, and BG/PG are worthwhile investigating in the future.

Table 3.8: Total residual errors (TREs) w.r.t. the use of different combinations of affecting factors for pitch/duration/energy level modeling

Pitch		Duration		Energy level	
APs	TRE	APs	TRE	APs	TRE
		+ Utterance	98.8%	+ Utterance	77.8%
+ Tone	71.6%	+ Tone	88.1%	+ Tone	74.5%
+ Coarticulation	60.3%	+ Base-syllable	62.9%	+ Final	48.0%
+ PW	51.7%	+ PW	48.6%	+ PW	46.9%
+ PPh	44.6%	+ PPh	45.0%	+ PPh	32.7%
+ BG/PG	37.4%	+ BG/PG	42.1%	+ BG/PG	26.0%
+ Prosodic state	1.1%	+ Prosodic state	1.1%	+ Prosodic state	1.0%

### 3.4.3 A Labeling Example

A typical example of the labeling results by A-UJPLM is given in Figure 3.16. Compared with the labeling result by UJPLM shown in Figure 2.14, most breaks labeled were the same except for an inserted *B2-3* and a substitution of *B3* with *B4* at the end of the first PP. The insertion of *B2-3* seemed to be reasonable because there existed an apparent syllable duration lengthening on the syllable “院”. For each prosodic feature of syllable log-*F0* mean, duration and energy level, the curve formed by integrating the prosodic-state APs and global mean showed smoother PW patterns as compared with those of the observed zigzag curve. The last syllables of all PWs had longer syllable duration illustrating the pre-boundary duration lengthening effect.

It is also found that apparent resets existed on the energy prosodic state of the last syllables of most PWs manifesting clear stress patterns.

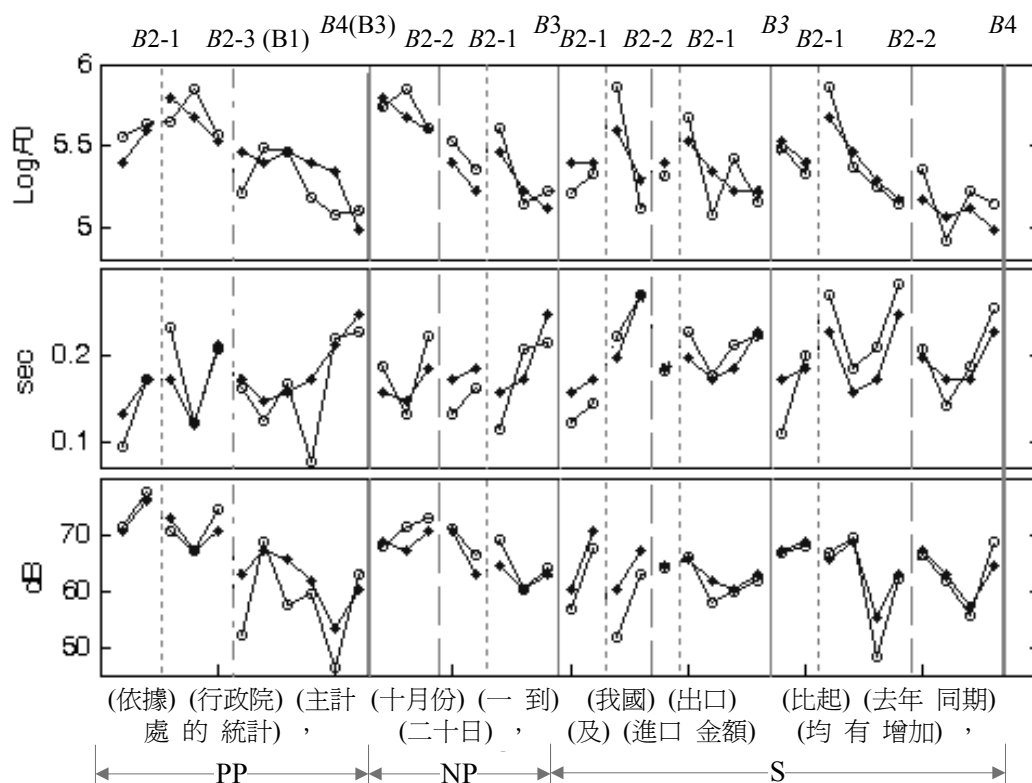
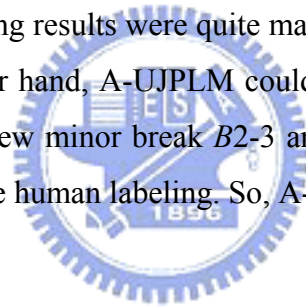


Figure 3.16: An example of the automatic prosody labeling by A-UJPLM. Upper, middle and lower panels represent observed (open circle) and prosodic state+global mean (solid diamond) of syllable log- $F_0$  means, syllable duration and syllable energy level, respectively. The utterance is “*yi-ju*(according to) *xing-zheng-yuan*(the Executive Yuan) *zhu-ji-chu*(Directorate-General of Budget, Accounting and Statistics) *de*(DE) *tong-ji*(statistics), *shi-yue-fen*(October) *yi*(1st) *dao*(to) *er-shi-ri*(20th), *wo-guo*(our country) *chu-kou*(export) *ji*(and) *jin-kou*(import) *jin-e*(the amount of money) *bi-qi*(in comparison with) *qu-nian*(last year) *tong-qi*(the same period) *jun*(both) *you*(to have some) *zeng-jia*(increase).”

### 3.5 Conclusions

In this chapter, the A-UJPLM method designed based on the UJPLM method discussed in Chapter 2 is proposed. It employs a new prosodic model to incorporate more acoustic features, more prosodic tags, and more affecting factors. Basically, A-UJPLM functions like UJPLM to perform the works of prosodic labeling and modeling jointly. It extends the UJPLM method to additionally model syllable duration and energy level. Besides, it adds a new break type *B2-3* to take care of minor break with pre-break syllable lengthening. Besides, some additional inter-syllable acoustic features, including normalized pitch jump, and normalized duration lengthening factors, are also incorporated to help the labeling and modeling task. Experimental results on the same Sinica Treebank corpus showed that A-UJPLM performed very well. The parameters of the eight prosodic sub-models are all linguistically/prosodically meaningful. A comparison with the results of UJPLM showed that their break labeling results were quite matched for the cases of non-break and major break. On the other hand, A-UJPLM could insert more minor breaks than UJPLM via introducing the new minor break *B2-3* and resulted in a more consistent labeling of minor breaks to the human labeling. So, A-UJPLM is a promising method.





# Chapter 4 An Application to Prosody Generation for TTS

## 4.1 Introduction

Prosody generation plays a very importance role on the naturalness of the synthesized speech in a TTS system. The main concern is to explore an appropriate mapping from the linguistic features of various levels extracted from the input text to the prosodic features representing the prosody hierarchy of the synthesized speech. Many methods have been proposed in the past, including the conventional rule-based approach [68,80,81], the linear regressive method [82], the decision tree-based method [83], the recurrent neural network-based method [12], the template tree-based approach [84], etc. Prosodic features can be divided into two types: numerical and categorical (or symbolic). For Mandarin speech, numerical prosody features can be some explicit values such as duration/pitch contour/energy level of each syllable and inter-syllable pause duration, while categorical features can be parameters representing the prosody hierarchy such as break indices on syllable junctures. Many existing TTS systems [8,10,11,37,82,85,86,87] generate prosody in two steps. First, symbolic prosodic features such as inter-syllable break types are predicted from the input linguistic features. A prosody hierarchical structure of the input text is then derived from the labeled symbolic prosodic features. Last, numerical prosodic features are obtained by superimposing prosodic patterns of various levels pre-stored or generated from a model, or by selecting prosodic templates from a speech inventory.

In this chapter, an A-UJPLM-based approach is proposed for prosody generation. Figure 4.1 displays the schematic diagram of the approach. It is composed of two steps: break prediction and prosodic feature prediction. In the break prediction step, a break type sequence is predicted for each input text by the break-syntax model  $p(\mathbf{B}|\mathbf{L})$  using some linguistic features extracted from the input text. The break type sequence implicitly forms a representation of the prosody hierarchy with PW as the basic synthesis unit of prosody generation. It has been suggested that a synthesized

utterance concatenated by PWs sounds more natural and pleasant than one by LWs [45]. Hence break prediction plays an importance role to properly parse the input text into strings of PWs, PPhs, and BG/PG. In the prosodic feature prediction step, four types of prosodic features, including syllable pitch contour, syllable duration, syllable energy level, and inter-syllable pause duration, are generated from input linguistic features and the break-type sequence generated in the first step by using the syllable prosodic model  $p(\mathbf{X}|\mathbf{B},\mathbf{PS},\mathbf{L})$ , the prosodic state model  $p(\mathbf{PS}|\mathbf{B},\mathbf{L})$ , and the break-acoustic model  $P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L})$ .

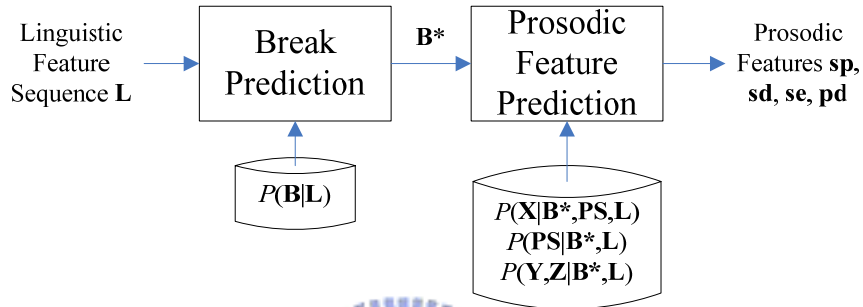


Figure 4.1: The proposed prosody generation method.

To evaluate the performance of the proposed break prediction and prosodic feature prediction methods, the test set of the Sinica Treebank corpus is adopted. The dataset consists of 46 utterances with 4801 syllables. It is labeled in advance with seven types of break  $\mathbf{B} = \{B0, B1, B2-1, B2-2, B2-3, B3, B4\}$  and three types of prosodic states  $\mathbf{PS} = \{\mathbf{p}, \mathbf{q}, \mathbf{r}\}$  by a Viterbi decoding algorithm which maximizes the objective function  $Q$  defined in Eq. (3.14). The prosodic models used in Eq. (3.14) are learned from the training set of the Sinica Treebank corpus discussed in Chapter 3. In the labeling process, all model parameters except the utterance APs are fixed. Steps of the labeling process are listed below:

- Step 1:* Initialize the utterance APs,  $\gamma_u$  and  $\alpha_u$ , by simply averaging syllable durations and syllable energy levels of each utterance.
- Step 2:* Re-label the prosodic state and break-type sequences of each utterance by using the Viterbi algorithm that maximizes  $Q$  defined in Eq. (3.14).
- Step 3:* Update the utterance APs with all other APs being fixed.
- Step 4:* Repeat Steps 2 to 3 until a convergence is reached.

In this study, the labeling process took 15 iterations to reach a convergence.

This Chapter is further organized as follows. Section 4.2 presents the proposed break prediction method. Then the proposed prosodic feature prediction method is discussed in Section 4.3. Some conclusions are given in the last section.

## 4.2 The Proposed Break Prediction Method

### 4.2.1 Linguistic Features

The linguistic features used for break prediction span a wide range from syllable level, such as initial type, and syllable juncture type (intra-word or inter-word); word level, such as, word length, POS, and type of punctuation mark (PM); to sentence level, such as length of sentence. They are discussed in more detail as follows.

#### (1) Syllable level

As illustrated in Subsection 2.4.4, we find that some syllable-level linguistic features are very useful for predicting break types. These features include the initial type of the following syllable and the syllable juncture type. Seven initial types are used in this study, including null initial,  $\{m, n, l, r\}$ ,  $\{b, d, g\}$ ,  $\{f, s, sh, shi, h\}$ ,  $\{ts, ch, chi\}$ ,  $\{p, t, k\}$ , and  $\{tz, j, ji\}$ ; and three types of syllable juncture are used, including inter-word, Type-1 intra-word and Type-2 intra-word. Here, Type-1 intra-words represent normal intra-word locations, while Type-2 intra-words are special intra-word locations of some specific long words, which have high potential to be pronounced with pauses such as “百分之\*三十二 *bai-fen-zhi san-shi*.”

#### (2) Word level

Word-level linguistic features used in this study are POS, word length, and PM. Four POS sets are used in this study, i.e. broad class word, level-1, level-2, and level-3 POS sets. The details of the four POS sets are listed in Appendix B. Word length is also an important feature for break prediction. It is observed in Subsection 2.5.1 that some short words are easy to be combined with its previous or following word. We categorize word length into five classes: 1 syllable, 2 syllables, 3 syllables,

4 syllables, and >4 syllables. A window up to six words is adopted in this study to extract the POS and word length features for the break prediction of the current word juncture: three words before and three words after the juncture. PM is the most significant feature to predict major break. Five types of PM are used in this study, including comma, period, question mark, dun hao “、,” and others.

### (3) Sentence level

The sentence-level features used are length of sentence, distance to the beginning of the current sentence, and distance to the end of the current sentence. By investigating the correlation between the length and the number of major/minor breaks in a sentence, we find that the number of minor/major breaks increases as the sentence becomes longer. It is also found that a major break is more likely to be inserted in a syllable juncture if it is far away from the beginning or end of a sentence. Table 4.1 summarizes the linguistic features used in this study.

Table 4.1: Summary of linguistic features used and their abbreviations

FI	Type of following syllable's initial: null initial, $\{m, n, l, r\}$ , $\{b, d, g\}$ , $\{f, s, sh, shi, h\}$ , $\{ts, ch, chi\}$ , $\{p, t, k\}$ , $\{z, j, ji\}$
SB	Type of syllable boundary: inter-word, Type-1 intra-word, Type-2 intra-word.
POS0	Broad class of preceding/following word: substantive word, function word
POS1	11-type POS: A, C, D, N, I, P, T, V, DE, SHI, DM
POS2	19-type POS : A, C, Dfa, Dfb, D, N, Nd, Ne, Ng, Nh, P, T, VA, VC, VH, V_2, DE, SHI, DM
POS3	47-type POS : A, Caa, Cab, Cba, Cbb, Da, Dfa, Dfb, Di, Dk, D, Na, Nb, Nc, Ncd, Nd, Neu, Nes, Nep, Neqa, Neqb, Nf, Ng, Nv, Nh, I, P, T, VA, VAC, VB, VC, VCL, VD, VE, VF, VG, VH, VHC, VI, VJ, VK, VL, V_2, DE, SHI, DM
WL	Length of word in syllable: 1, 2, 3, 4, >4
PM	Type of PM: comma, period, question mark, dun hao and others
LS	Length of sentence in syllable
LPS	Length of previous sentence
LFS	Length of following sentence
DPP	Distance to previous PM (the beginning of the sentence)
DFP	Distance to following PM (the end of the sentence)

## 4.2.2 Prediction Methods

In this study, three break prediction methods are discussed, namely (1) the baseline all-in-one CART-based method, (2) the two-stage method, and (3) the Markov model-based method.

### (1) The baseline all-in-one CART-based method

As shown in Figure 4.2, this method predicts seven break types according to the question set formed from the linguistic features discussed in Subsection 4.2.1 by a single decision tree trained by the CART algorithm. The split criterion used in the CART algorithm is the maximum information gain.



Figure 4.2: All-in-one CART for break prediction.

Table 4.2 displays the prediction result by the baseline method. It can be found from the table that the prediction rates were high for  $B1$  and  $B4$ , medium for  $B3$  and  $B0$ , and low for  $B2-1$ ,  $B2-2$  and  $B2-3$ . The overall prediction accuracies were 76.2% and 73.7% for the inside and outside tests, respectively. The low prediction accuracies of the three types of minor break mainly resulted from their confusions with  $B1$  caused in part by the relatively large counts of  $B1$  in many leaf nodes of the trained decision tree.  $B0$ s were also easily confused with  $B1$ s. Since both  $B0$  and  $B1$  are defined as intra-PW boundary, this type of confusion may not harmful to the following prosodic feature prediction.  $B3$ s were mainly confused with  $B2-2$  and  $B4$ . This result seemed reasonable because the acoustic characteristics of  $B2-2$ ,  $B3$  and  $B4$  are overlapped with apparent inter-syllable pause duration.

Table 4.3 displays the confusion matrix of the target and predicted break types which are reduced to three broad classes of break, i.e. non-break  $\{B0, B1\}$ , minor break  $\{B2-1, B2-2, B2-3\}$  and major break  $\{B3, B4\}$ . It can be seen from the table that the overall prediction accuracies were 87.5% and 85.8% for the inside and outside tests, respectively. Although the overall prediction accuracies were high, the

accuracies for minor break were still too low. This may result in parsing the input text into too long PWs so as to make an over-smoothed intonation to lose pleasant rhythm. The prediction ability for minor break certainly needs to be improved.

Table 4.2: The confusion matrix of the target and predicted break types (%) using the baseline all-in-one CART-based method for (a) the inside and (b) outside tests.

(a)								
Tar\Pre	B0	B1	B2-1	B2-2	B2-3	B3	B4	Total
B0	<b>74.1</b>	19.9	3.5	1.2	0.9	0.3	0.0	10247
B1	4.8	<b>89.8</b>	2.8	1.3	1.0	0.3	0.0	23744
B2-1	5.4	26.0	<b>56.4</b>	7.5	3.4	1.2	0.0	4782
B2-2	2.9	15.7	14.1	<b>52.9</b>	3.7	10.5	0.1	3443
B2-3	4.9	34.5	17.5	10.6	<b>29.9</b>	2.5	0.0	3046
B3	0.8	4.6	3.3	8.6	0.9	<b>72.8</b>	9.0	3527
B4	0.0	0.2	0.1	0.3	0.0	13.9	<b>85.4</b>	2703
								<b>Avg = 76.2</b>
(b)								
Tar\Pre	B0	B1	B2-1	B2-2	B2-3	B3	B4	Total
B0	<b>73.0</b>	20.3	4.5	1.1	0.5	0.6	0.0	626
B1	5.8	<b>87.3</b>	3.8	1.5	1.2	0.4	0.0	2107
B2-1	2.4	26.8	<b>51.8</b>	12.0	4.5	2.4	0.0	332
B2-2	0.9	20.3	15.7	<b>51.1</b>	2.8	9.2	0.0	325
B2-3	7.0	39.3	20.4	12.9	<b>16.9</b>	3.5	0.0	201
B3	0.4	6.0	4.3	16.7	1.1	<b>58.9</b>	12.8	282
B4	0.0	0.0	0.0	1.2	0.0	12.3	<b>86.4</b>	162
								<b>Avg = 73.7</b>

Table 4.3: The confusion matrix of the break prediction for the baseline method evaluated using 3 broad classes of break: (a) The inside and (b) outside tests. (NB: non-break, MiB: minor break, MB: major break)

(a)				
Tar\Pre	NB	MiB	MB	Total
NB	<b>94.4</b>	5.3	0.3	33991
MiB	29.7	<b>65.9</b>	4.4	11271
MB	3.1	7.4	<b>89.4</b>	6230
				<b>Avg = 87.5</b>
(b)				
Tar\Pre	NB	MiB	MB	Total
NB	<b>93.1</b>	6.4	0.5	2733
MiB	30.2	<b>64.6</b>	5.2	858
MB	4.1	14.4	<b>81.5</b>	444
				<b>Avg = 85.8</b>

## (2) The two-stage method

By detailed analysis of the break prediction results of the baseline method, we find that minor breaks are easily confused with non-breaks. The deletion of a minor break may make the synthesized speech too hasted so as to degrade its naturalness. To improve the break prediction accuracy of minor break, a two-stage method is proposed. Figure 4.3 shows its block diagram. In the first stage, a three-class CART is trained to classify the three broad classes of major break, minor break, and non-break. In the second stage, each syllable juncture is further classified by one of other three decision trees into seven-class break type. For example, if a syllable juncture is determined as a major break in the first stage, then it is fed into  $B3/B4$  classification to be classified as  $B3$  or  $B4$ .

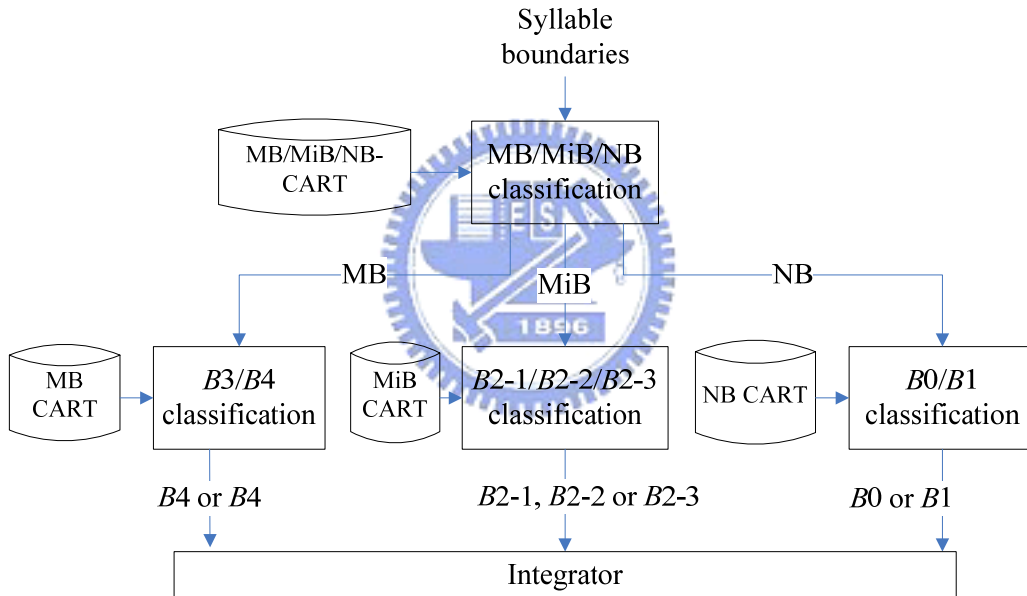


Figure 4.3: A block diagram of the two-stage break prediction method.

The performance of the method is listed in Tables 4.4 and 4.5. It can be seen from Table 4.4 that the first stage achieved 12.7% and 13.3% improvements on the minor break prediction in the inside and outside tests, respectively, as compared with the baseline method. Although the detection accuracies of both non-break and major break degraded slightly, the overall accuracies were improved. For the second-stage prediction, as shown in Table 4.5 the overall predictions of seven break types were improved in both inside and outside tests, especially in the predictions of  $B2-1$ ,  $B2-2$ , and  $B2-3$ .

Table 4.4: The confusion matrix of target and predicted reduced three classes break types using the two-stage approach: (a) inside test (b) outside test.

(a)				
Tar\Pre	NB	MiB	MB	Total
NB	<b>94.0</b>	5.7	0.3	33991
MiB	18.4	<b>78.6</b>	2.9	11271
MB	1.8	11.0	<b>87.2</b>	6230
				<b>Avg = 89.8</b>

(b)				
Tar\Pre	NB	MiB	MB	Total
NB	<b>92.4</b>	7.4	0.2	2733
MiB	18.4	<b>77.9</b>	3.7	858
MB	1.6	16.4	<b>82.0</b>	444
				<b>Avg = 88.2</b>

Table 4.5: The confusion matrix of target and predicted seven break types using the two-stage approach: (a) inside test and (b) outside test.

(a)								
Tar\Pre	B0	B1	B2-1	B2-2	B2-3	B3	B4	Total
B0	<b>74.7</b>	20.1	2.8	1.0	1.2	0.3	0.0	10247
B1	4.8	<b>88.8</b>	2.8	1.5	1.7	0.3	0.0	23744
B2-1	4.5	16.1	<b>62.4</b>	8.5	7.5	0.9	0.0	4782
B2-2	1.6	7.8	17.1	<b>59.2</b>	7.1	6.8	0.3	3443
B2-3	3.7	21.4	17.1	13.5	<b>43.0</b>	1.3	0.0	3046
B3	0.8	2.3	4.5	12.2	2.5	<b>68.4</b>	9.3	3527
B4	0.0	0.1	0.1	0.1	0.0	12.8	<b>86.8</b>	2703
								<b>Avg = 77.4</b>

(b)								
Tar\Pre	B0	B1	B2-1	B2-2	B2-3	B3	B4	Total
B0	<b>70.6</b>	22.8	3.7	1.0	1.9	0.0	0.0	626
B1	5.8	<b>86.3</b>	4.0	1.5	2.1	0.3	0.0	2107
B2-1	2.1	18.1	<b>58.4</b>	10.8	8.7	1.8	0.0	332
B2-2	0.6	10.5	22.8	<b>52.3</b>	7.4	5.8	0.6	325
B2-3	1.0	26.4	22.9	15.4	<b>31.8</b>	2.0	0.5	201
B3	0.0	2.5	3.2	18.1	4.3	<b>61.0</b>	11.0	282
B4	0.0	0.0	0.0	0.6	0.0	9.9	<b>89.5</b>	162
								<b>Avg = 74.5</b>



### (3) The Markov model-based method

The third method we tested is the Markov model-based method. Compared with the two methods illustrated above, the Markov model-based method not only takes linguistic features but also incorporates contextual information to predict break type. The general form of break prediction in this method can be expressed by

$$\mathbf{B}^* = \arg \max_{\mathbf{B}} p(\mathbf{B} | \mathbf{L}) = \arg \max_{\mathbf{B}} \prod_{n=1}^N p(B_n | B_1^{n-1}, \mathbf{L}_n) \quad (4.1)$$

Many methods can be used to generate the probability  $p(B_n | B_1^{n-1}, \mathbf{L}_n)$ . They include the HMM method [46,49] and the CART algorithm, etc. In this study, we adopt the CART algorithm to train a decision tree for generating the probability  $p(B_n | B_1^{n-1}, \mathbf{L}_n)$ . Here we only consider the classification of the three broad classes of break. The decision tree is a refinement of the decision tree trained in the first stage of the two-stage method. We further split the leaf nodes of the previous-trained decision tree using a question set containing the information of previous breaks. The optimal break sequence can be obtained by the Viterbi decoding algorithm solving Eq. (4.1). The experimental results are shown in Table 4.6. Compared with the results of the two-stage method shown in Table 4.4, the overall accuracies of break prediction were slightly higher for both inside and outside tests. However, the improvement was not significant for the outside test. This may suggest that the break prediction mainly relies on the linguistic features rather than the contextual information of break.

Table 4.6: The confusion matrix of target and predicted reduced three classes break types using the Markov model: (a) inside test (b) outside test.

(a)				
Tar\Pre	NB	MiB	MB	Total
NB	<b>94.2</b>	5.5	0.3	33991
MiB	18.4	<b>78.6</b>	2.9	11271
MB	1.9	10.6	<b>87.6</b>	6230
				Avg = <b>90.1</b>
(b)				
Tar\Pre	NB	MiB	MB	Total
NB	<b>92.5</b>	7.3	0.2	2733
MiB	18.3	<b>78.0</b>	3.7	858
MB	1.6	16.4	<b>82.0</b>	444
				Avg = <b>88.4</b>

### 4.3 Prosodic Feature Prediction

The prosodic features to be predicted include syllable prosodic features (**sp**, **sd**, **se**) and inter-syllable pause duration (**pd**). Among them, the inter-syllable pause duration of each syllable juncture can be simply predicted by the break-acoustic model trained in Chapter 3, i.e.

$$pd_n^* = \arg \max_{pd_n} p(pd_n | B_n^*, \mathbf{I}_n) \quad (4.2)$$

where  $B_n^*$  represents the optimal break type of syllable  $n$  predicted by the break-syntax model discussed in Section 4.2. The syllable prosodic features, including syllable pitch contour **sp**, syllable duration **sd**, and syllable energy level **se**, are predicted by the models formulated basing on the minimum mean squared error (MMSE) criterion. Given with the predicted break sequence  $\mathbf{B}^*$  and linguistic features  $\mathbf{I}$ , the MMSE predictors for **sp**, **sd**, and **se** are  $\mathbf{sp}_n^* = E[\mathbf{sp}_n | \mathbf{B}^*, \mathbf{I}]$ ,  $sd_n^* = E[sd_n | \mathbf{B}^*, \mathbf{I}]$ , and  $se_n^* = E[se_n | \mathbf{B}^*, \mathbf{I}]$ , respectively. Since **sp**, **sd**, and **se** are predicted in the same way, we only present the prediction model of **sp** here for simplicity. The MMSE predictor for **sp** can be elaborated by

$$\begin{aligned} \mathbf{sp}_n^* &= E[\mathbf{sp}_n | \mathbf{B}^*, \mathbf{I}] \\ &= \int \mathbf{sp}_n P(\mathbf{sp}_n | \mathbf{B}^*, \mathbf{I}) d\mathbf{sp}_n \\ &= \int \mathbf{sp}_n \sum_{p_n} P(\mathbf{sp}_n | p_n, B_{n-1}^*, t_{n-1}^{n+1}) P(p_n | \mathbf{B}^*, \mathbf{I}) d\mathbf{sp}_n \cdot \\ &= \sum_{p_n} (\boldsymbol{\beta}_{t_n}^{p_n} + \boldsymbol{\beta}_{p_n}^{p_n} + \boldsymbol{\beta}_{B_{n-1}^*, t_{n-1}^{n+1}}^f + \boldsymbol{\beta}_{B_n^*, t_{p_n}}^b + \boldsymbol{\mu}) P(p_n | \mathbf{B}^*, \mathbf{I}) \end{aligned} \quad (4.3)$$

It can be seen from Eq. (4.3) that the predicted syllable pitch contour is a weighted sum of the reconstructed patterns formed by superimposing various APs with weights being the *a posterior* probabilities of prosodic state  $p_n$ . The *a posterior* probability  $P(p_n | \mathbf{B}^*, \mathbf{I})$  can be formulated as

$$P(p_n = i | \mathbf{B}^*, \mathbf{I}) = \frac{P(p_n = i, \mathbf{B}^*, \mathbf{I})}{\sum_j P(p_n = j, \mathbf{B}^*, \mathbf{I})} = \frac{a_n(i)b_n(i)}{\sum_j a_n(j)b_n(j)} \quad (4.4)$$

where  $a_n(i) = P(B_1^* \cdots B_n^*, p_n = i | \mathbf{I})$  and  $b_n(i) = P(B_{n+1}^* \cdots B_N^* | p_n = i, \mathbf{I})$  are the forward and backward probabilities, respectively.  $a_n(i)$  and  $b_n(i)$  can be calculated by the

forward and backward algorithm with the probability  $P(p_n | p_{n-1}, B_{n-1}^n, \mathbf{I}_n)$  which is similar to the prosodic state model  $P(p_n | p_{n-1}, B_{n-1})$ . The probability  $P(p_n | p_{n-1}, B_{n-1}^n, \mathbf{I}_n)$  strengthens the influences of linguistic features and break  $B_n$  on the current prosodic state  $p_n$ . In practical realization, since the space of the histories  $\{p_{n-1}, B_{n-1}^n\}$  and linguistic features  $\{\mathbf{I}_n\}$  is too large, we partition the space into several classes  $C(p_{n-1}, B_{n-1}^n, \mathbf{I}_n)$  to calculate the conditional probabilities  $P(p_n | C(p_{n-1}, B_{n-1}^n, \mathbf{I}_n))$  by the decision tree method. The detail of the question set for constructing the decision tree is listed bellow:

- (1) Current word length in syllable:  $\{1, 2, 3, 4, >4\}$ .
- (2) Current syllable position in word:  $\{1^{\text{st}}, \text{intermediate}, \text{last}, \text{mono-syllable word}\}$ .
- (3) Sentence length in syllable:  $\{1, [2,5], [6,10], [11,15], [16,20], >20\}$ .
- (4) Current syllable position in sentence:  $\{1^{\text{st}}, 2^{\text{nd}}, 3^{\text{rd}}, [4^{\text{th}}, 5^{\text{th}}], [6^{\text{th}}, 7^{\text{th}}], [8^{\text{th}}, 11^{\text{th}}], \text{last}, 2^{\text{nd}} \text{ last}, 3^{\text{rd}} \text{ last}, [5^{\text{th}} \text{ last}, 4^{\text{th}} \text{ last}], [7^{\text{th}} \text{ last}, 6^{\text{th}} \text{ last}], [11^{\text{th}} \text{ last}, 8^{\text{th}} \text{ last}], \text{others}\}$ ; Smaller count number from the beginning or end wins.
- (5) PM after the current syllable (five types).
- (6) POS3: 47-types POS.
- (7) Break type of *juncture*  $n, n-1, n-2$ .
- (8) Prosodic state of  $(n-1)$ -th syllable.

The proposed prediction method is conducted with the break sequence given by the two-stage method. We choose the two-stage method because it performs better. Table 4.7 displays the TREs of the prosody prediction results for syllable pitch contour, duration and energy level. Since the performance of the proposed method should not consider the influence of utterance, the TREs of syllable duration and energy level are respectively the ratios of the sum-squared prediction errors of syllable duration and energy level over the sum-squared normalized ones with the influences from utterance being removed. The performances were acceptable. To separate the effect of break prediction on the prosodic feature prediction, we do the same experiment using the correct break labels. Table 4.8 displays the experimental results. By comparing the results shown in the two tables, we find that the latter performed better. This shows that the break prediction plays an important role in the prediction of prosodic features. Erroneous breaks predicted will make gross shifts of

PW patterns and may result in large prediction errors of prosodic features. Hence, to improve the prosodic feature generation, the break prediction task is essential and worthwhile further investigating in the future.

Table 4.7: TREs of the prosodic feature prediction results.

	<b>sp</b>	<b>sd</b>	<b>se</b>	<b>pd</b>
Inside	42.39%	45.60%	37.61%	18.50%
Outside	42.73%	46.24%	35.98%	18.92%

Table 4.8: TREs of the prosodic feature prediction using correct break labels.

	<b>sp</b>	<b>sd</b>	<b>se</b>	<b>pd</b>
Inside	32.72%	34.80%	32.13%	8.64%
Outside	39.10%	41.74%	33.33%	7.00%

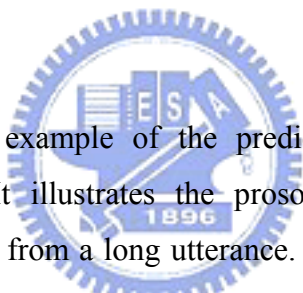
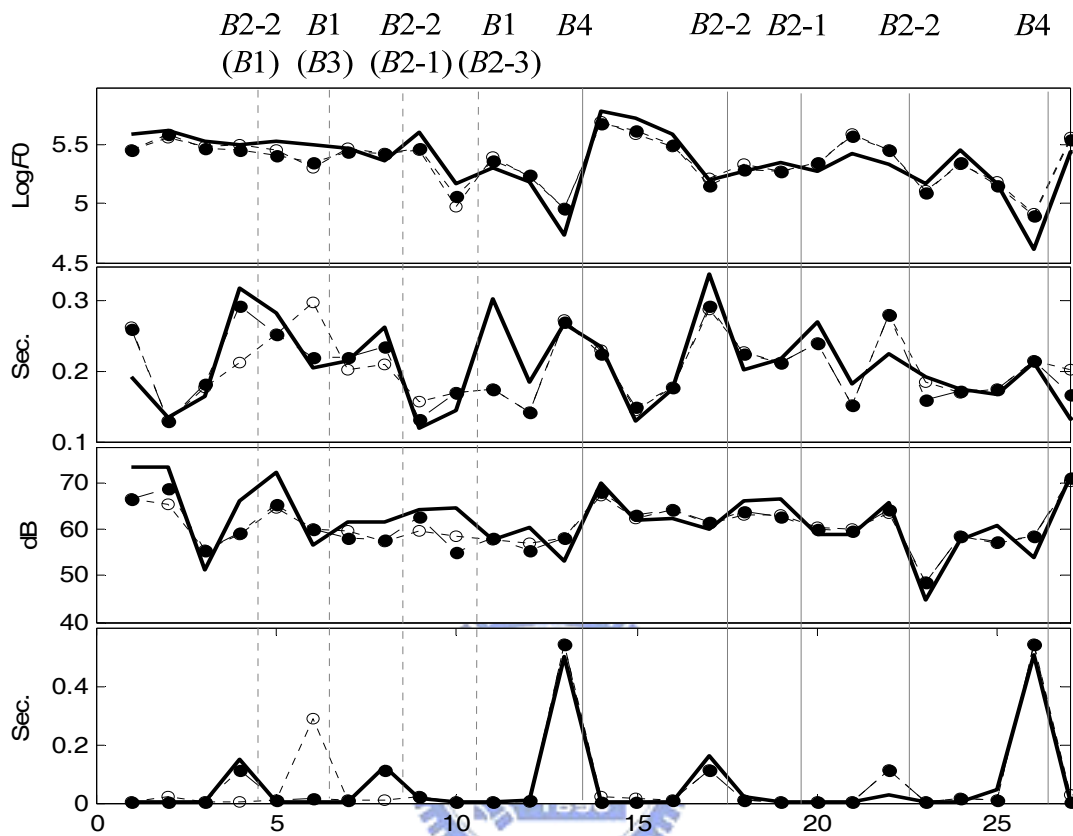


Figure 4.4 displays an example of the predicted prosodic features by the A-UJPLM-based approach. It illustrates the prosodic feature variations of two sentential utterances extracted from a long utterance. It can be found from the figure that the predicted prosodic features matched well with their original counterparts for most syllables. Some large errors can be found to occur on the syllable durations and inter-syllable pause durations of the first sentence. They were mainly resulted from a series of break prediction errors. For example, the two contiguous break prediction errors (predict  $(B2-2, B1)$  as  $(B1, B3)$ ) in the first sentence caused a gross shift of the first PW showing a move of the phrase-ending lengthening from the 4<sup>th</sup> syllable to the 6<sup>th</sup> syllable. The after-phrase long pause also shifted two syllables to the right synchronously. By using correct break labels, these large prosodic feature prediction errors disappeared. This confirmed that break prediction errors are responsible for prosodic feature prediction errors. So break prediction plays an important role on prosodic feature prediction.



勤益公司現金增資一點六億元，申購日期截至十一月五日為止，

Figure 4.4: An example of the prosodic feature prediction by the A-UJPLM-based approach. The panels from up to bottom represent, respectively, syllable log- $F_0$  means, syllable duration, syllable energy level and inter-syllable pause duration. Solid lines, open circles, and closed circles denote, correspondingly, the original features, the predicted features using predicted breaks, and the predicted features using correct break labels. Vertical dash lines represent erroneous major/minor break prediction boundaries while vertical solid lines represent correct ones. Notice that break labels in ( ) represent erroneous breaks predicted.

## 4.4 Conclusions

In this chapter, a model-based prosody generation method for TTS is discussed. The method contains two steps: break prediction and prosodic feature prediction. In the break prediction, three methods are investigated. They include the baseline all-in-one CART-based method, the two-stage method and the Markov model-based method. Among them, the Markov model-based method achieves the highest accuracy in predicting three-class break types of non-break, minor break and major break, while the baseline all-in-one CART-based method has the worst performance. However, compared with the two-stage method, the Markov model-based method can only bring negligible improvement on the outside test. We therefore conclude that the linguistic features rather than the contextual break information are primary features in break prediction.

Based on the break prediction result by the two-stage method, four prosodic features, including syllable pitch contour, syllable duration, syllable energy level, and inter-syllable pause duration, are predicted by the proposed A-UJPLM-based prosody generator. Experimental results showed that the performance of the proposed method is acceptable. An upper bound of performance obtained in the oracle experiment using correct break labels confirms that the break prediction task is essential in prosodic feature generation. Further elaboration of the break prediction model is worthwhile studying in the future.

## Chapter 5 Conclusions and Future Works

### 5.1 Conclusions

In this dissertation, an unsupervised joint prosody labeling and modeling method (UJPLM) for Mandarin speech has been proposed. Unlike the conventional prosody labeling task that is fulfilled by trained human labelers with audio-visual aids, the proposed method not only intended to objectively/consistently label prosodic tags but also to concurrently construct interpretive prosodic models. Two types of prosodic tags are determined by four prosodic models designed to illustrate the hierarchy of Mandarin prosody: the break of a syllable juncture to demarcate prosodic constituents and the prosodic states to represent any prosodic domain's pitch level variation resulting from its upper-layered prosodic constituents' influences. The four prosodic models are (1) the syllable pitch contour model which describes the variations in syllable pitch contour controlled by the prosodic tags and syllable-level linguistic features, (2) the break-acoustics model which describes the relationship between the break type of a syllable juncture and nearby acoustic features, (3) the break-syntax model which constructs the relationship between the break type of a syllable juncture and contextual linguistic features and (4) the prosodic state model which describes the relationship between the prosodic states of syllables and the break types of neighboring syllable junctures. An experiment on the Sinica Treebank corpus uttered by an experienced female announcer showed that the four prosodic models learned were all linguistically and/or prosodically meaningful. The corresponding relationship between the break indices labeled by UJPLM and their associated words were investigated to confirm the performance of UJPLM. The prosodic state labeled could be used to extract the general  $\log-F_0$  patterns of PW, PPh and BG/PG. Besides, a quantitative comparison between the break labeling results by UJPLM and human labelers showed that the breaks labeled by UJPLM were more consistent and discriminative than those by human in prosodic feature distributions, a further verification of the proposed method.

Motivated by the success of UJPLM, the A-UJPLM method was designed basing on the same idea to incorporate more acoustic features and more prosodic tags. The

new prosodic features added included syllable duration, syllable energy level, normalized pitch jump, and normalized duration lengthening factors, while the new prosodic tags are the break type *B2-3* introduced to take care of the PW boundary with pre-break syllable lengthening, the duration prosodic state and the energy prosodic state. Experimental results on the same Sinica Treebank corpus showed that A-UJPLM performed very well. The parameters of the eight prosodic sub-models were all linguistically/prosodically meaningful. A comparison with the results of UJPLM showed that their break labeling results were quite match for non-breaks and major breaks. However, A-UJPLM could insert more minor breaks than UJPLM via introducing the new minor break *B2-3* and resulted in a more consistent labeling of minor breaks to the human labeling. Besides, the general duration and energy level patterns of PW, PPh, and BG/PG could also be explored using the labeled duration and energy prosodic states. So, a more substantial prosody labeling and modeling for Mandarin speech was achieved by the A-UJPLM method.

Lastly, an A-UJPLM-based prosody generation method for TTS was proposed. It is composed of two steps: break prediction and prosodic feature prediction. In the break prediction, three prediction methods were discussed: the baseline all-in-one CART-based method, the two-stage method and the Markov model-based method. Based on the break prediction result by the two-stage method, four types of prosodic features, including syllable pitch contour, syllable duration, syllable energy level, and inter-syllable pause, were generated from input linguistic features and the break-type sequence generated in the first step by using the syllable prosodic model, the prosodic state mode and the break-acoustic model. Experimental results showed that the performance of the proposed method was acceptable.

In conclusion, the proposed unsupervised joint prosody labeling and modeling method was able to construct interpretive prosodic models and generate proper prosodic tags automatically. Therefore, it is a promising prosodic labeling and modeling approach for Mandarin speech.



## 5.2 Future Works

Some future works are worth doing. First, the four prosodic models can be directly used or further elaborated to provide useful prosodic information to assist in some applications of spoken language processing, including ASR, punctuation generation from unprescribed speech utterance, prosody generation for TTS, prosody pattern conversion for different speakers, and prosodic error detection in computer-assisted Mandarin Chinese learning. Second, the speech database with prosodic tags being properly labeled can be used to exploit the hierarchical structure of Mandarin prosody in more detail, especially for high-layer prosodic constituents. Third, via prosody labeling and modeling of large multi-speaker, emotional, and multi-speaking rate speech databases, the influences of different speaking styles on prosody can be explored.



## Bibliography

- [1] E. Selkirk, “On prosodic structure and its relation to syntactic structure,” *Nordic Prosody* (Tapir, Trondheim, Norway), Vol. 2, pp. 111–140.
- [2] E. Selkirk, *Phonology and Syntax: The Relation Between Sound and Structure* (MIT Press, Cambridge, MA, 1984).
- [3] M. Beckman and J. Pierrehumbert, “Intonational structure in Japanese and English,” *Phonology Yearbook 3* (Cambridge University Press, UK, 1986), pp. 255–309.
- [4] J.-F. Cao, “Rhythm of spoken Chinese—Linguistic and paralinguistic evidences,” *Proceedings of the ICSLP 2000*, Vol. 2, pp. 357–360.
- [5] Q. Shi, X.-J. Ma, W.-B. Zhu, W. Zhang and L.-Q. Shen, “Statistic prosody structure prediction,” *Proceedings of IEEE Workshop on Speech Synthesis 2002*, pp. 155-158.
- [6] Z. Sheng, J.-H. Tao, and D.-L. Jiang, “Chinese prosodic phrasing with extended features,” *Proceedings of the IEEE ICASSP 2003*, Vol. 1, pp. 492–495.
- [7] G.-H. Fu and K.K Luke, “Integrated approaches to prosodic word prediction for Chinese TTS,” *Proceeding of the IEEE NLP-KE 2003*, pp. 413-418.
- [8] C.-Y. Tseng, S.-H. Pin, Y.-L. Lee, H.-M. Wang, and Y.-C. Chen, “Fluent speech prosody: Framework and modeling,” *Speech Commun. special issue on quantitative prosody modeling for natural speech description and generation*, **46**, 284–309 (2005).
- [9] C.-Y. Tseng, “Higher level organization and discourse prosody,” *Proceedings of the TAL 2006*, pp. 23–34.
- [10] S.-H. Pin, Y.-L. Lee, Y.-C. Chen, H.-M. Wang, and C.-Y. Tseng, “A Mandarin TTS system with an integrated prosodic model,” *Proceedings of the ISCSLP 2004*, pp. 169–172.
- [11] N.-H. Pan, W.-T. Jen, S.-S. Yu, M.-S. Yu, S.-Y. Huang, and M.-J. Wu, “Prosody model in a Mandarin text-to-speech system based on a hierarchical approach,” *Proceedings of the ICME 2000*, Vol. 1, pp. 448–4511.
- [12] S.-H. Chen, S.-H. Hwang, and Y.-R. Wang, “An RNN-based prosodic information synthesizer for Mandarin text-to-speech,” *IEEE Trans. Speech Audio Process.* **6**, 226–239 (1998).
- [13] Y. Liu, E. Shriberg, S. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, “Enriching speech recognition with automatic detection of sentence boundaries and disfluencies,” *IEEE Trans. Audio, Speech, Lang. Process.* **14**, 526–1540 (2006).

- [14] Y. Gotoh and S. Renals, "Sentence boundary detection in broadcast speech transcripts," Proceedings of the ISCA Workshop: Automatic Speech Recognition: Challenges for the New Millennium ASR 2000, pp. 228–235.
- [15] E. Shriberg, A. Stolcke, D. Hakkani-Tur, and G. Tur, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Commun.* **32**, 127–154 (2000).
- [16] J.-H. Kim and P. C. Woodland, "A combined punctuation generation and speech recognition system and its performance enhancement using prosody," *Speech Commun.* **41**, 563–577 (2003).
- [17] J.-H. Kim and P. C. Woodland, "The use of prosody in a combined system for punctuation generation and speech recognition," Proceedings of the Eurospeech 2001, pp. 2757–2760.
- [18] H. Christensen, Y. Gotoh, and S. Renals, "Punctuation annotation using statistical prosody models," Proceedings of the ISCA Workshop on Prosody in Speech Recognition and Understanding 2001, pp. 35–40.
- [19] J.-F. Yeh and C.-H. Wu, "Edit disfluency detection and correction using a cleanup language model and an alignment model," *IEEE Trans. Audio, Speech, Lang. Process.* **14**, 1574–1583 (2006).
- [20] M. Lease, M. Johnson, and E. Charniak, "Recognizing disfluencies in conversational speech," *IEEE Trans. Audio, Speech, Lang. Process.* **14**, 1566–1573 (2006).
- [21] C.-K. Lin and L.-S. Lee, "Improved spontaneous Mandarin speech recognition by disfluency interruption point (IP) detection using prosodic features," Proceedings of the Eurospeech 2005, pp. 1621–1624.
- [22] K. Chen and M. Hasegawa-Johnson, "How prosody improves word recognition," Proceedings of the ISCA International Conference on Speech Prosody 2004, pp. 583–586.
- [23] K. Chen, M. Hasegawa-Johnson, A. Cohen, S. Borys, S.-S. Kim, J. Cole, and J.-Y. Choi, "Prosody dependent speech recognition on radio news corpus of American English," *IEEE Trans. Audio, Speech, Lang. Process.* **14**, 232–245 (2006).
- [24] K. Chen and M. Hasegawa-Johnson, "Improving the robustness of prosody dependent language modeling based on prosody syntax dependence," Proceedings of the IEEE ASRU 2003, pp. 435–440.
- [25] E. Shriberg and A. Stolcke, "Direct modeling of prosody: An overview of applications in automatic speech processing," Proceedings of the ISCA International Conference on Speech Prosody 2004, pp. 575–582.
- [26] J.-H. Yang, Y.-F. Liao, Y.-R. Wang, and S.-H. Chan, "A new approach of using temporal information in Mandarin speech recognition," Proceedings of the ISCA

International Conference on Speech Prosody 2006, Vol. **SPS4-3**.

- [27] X. Lei and M. Ostendorf, “Word-level tone modeling for Mandarin speech recognition,” Proceedings of the IEEE ICASSP 2007, Vol. **4**, pp. 665–668.
- [28] C.-Y. Tseng, “Recognizing Mandarin Chinese fluent speech using prosody information—An initial investigation,” Proceedings of the ISCA International Conference on Speech Prosody 2006.
- [29] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, “ToBI: A standard for labeling English prosody,” Proceedings of the ICSLP 1992, Vol. **2**, pp. 867–870.
- [30] A. Batliner, J. Buckow, H. Niemann, E. Noth, and V. Warnke, “The prosody module,” in *VerbMobil: Foundations of Speech-to-Speech Translation*, edited by W. Wahlster (Springer, New York, 2000).
- [31] M. Selting, *Prosody in Conversation* (Max Niemeyer, Tuebingen, Germany, 1995), in German.
- [32] D. J. Hirst, “The symbolic coding of fundamental frequency curves: From acoustics to phonology,” Proceedings of the International Symposium on Prosody 1994.
- [33] P. A. Taylor, “The tilt intonation model,” Proceedings of the ICSLP 1998, Vol. **4**, pp. 1383–1386.
- [34] S.-H. Peng, M. K. M. Chan, C.-Y. Tseng, T. Huang, O.-J. Lee, and M. Beckman, “Towards a Pan-Mandarin system for prosodic transcription,” in *Prosodic Typology: The Phonology of Intonation and Phrasing*, edited by S.-A. Jun (Oxford University Press, Oxford, 2005), pp. 230–270.
- [35] A.-J. Li, “Chinese prosody and prosodic labeling of spontaneous speech,” Proceedings of the ISCA International Conference on Speech Prosody 2002, pp. 39–46.
- [36] G.-P. Chen, G. Bailly, Q.-F. Liu, and R.-H. Wang, “A superposed prosodic model for Chinese text-to-speech synthesis,” Proceedings of the ISCSLP 2004, pp. 177–180.
- [37] M.-S. Yu, N.-H. Pan, and M.-J. Wu, “A statistical model with hierarchical structure for predicting prosody in a Mandarin text-to-speech system,” Proceedings of the ISCSLP 2002, pp. 21–24.
- [38] G. Bailly and B. Holm, “SFC: A trainable prosodic model,” *Speech Commun.* **46**, 348–364 (2005).
- [39] M. Ostendorf and N. Veilleux, “A hierarchical stochastic model for automatic prediction of prosodic boundary location,” *Comput. Linguist.* **20**, 27–52 (1994).

- [40] X. Shen and B. Xu, “A CART-based hierarchical stochastic model for prosodic phrasing in Chinese,” Proceedings of the ISCSLP 2000, pp. 105–109.
- [41] H.-J. Peng, C.-C. Chen, C.-Y. Tseng, and K.-J. Chen, “Predicting prosodic words from lexical words—A first step towards predicting prosody from text,” Proceedings of the ISCSLP 2004, pp. 173–176.
- [42] J. Hirschberg and P. Prieto, “Training intonational phrasing rules automatically for English and Spanish text-to-speech,” *Speech Commun.* **18**, 281–290 (1996).
- [43] D.-W. Xu, H.-F. Wang, G.-H. Li, and T. Kagoshima, “Parsing hierarchical prosodic structure for Mandarin speech synthesis,” Proceedings of the IEEE ICASSP 2006, Vol. **1**, pp. 14–19.
- [44] X. Sun and T. H. Applebaum, “Intonational phrase break prediction using decision tree and n-gram model,” Proceedings of the Eurospeech 2001, pp. 537–540.
- [45] M. Chu, Y. Qian, “Locating Boundaries for Prosodic Constituents in Unrestricted Mandarin Texts,” *Computat. Linguist. and Chinese Language Processing*, **6**, 61-82 (2001).
- [46] A. W. Black and P. Taylor, “Assigning phrase breaks from part-of-speech sequences,” Proceedings of the Eurospeech 1997, pp. 995–998.
- [47] Z. Sheng, J.-H. Tao, and D.-L. Jiang, “Chinese prosodic phrasing with extended features,” Proceedings of the IEEE ICASSP 2003, Vol. **1**, pp. 492–495.
- [48] J.-F. Li, G.-P. Hu, and R.-H. Wang, “Chinese prosody phrase break prediction based on maximum entropy model,” Proceedings of the Interspeech 2004, pp. 729–732.
- [49] Y.-Q. Shao, Y.-Z. Zhao, J.-Q. Han, and T. Liu, “Using different models to label the break indices for mandarin speech synthesis,” Proceedings of the ICMLC 2005, Vol. **6**, pp. 3802–3807.
- [50] J.-F. Li, G.-P. Hu, R.-H. Wang, and L.-R. Dai, “Sliding window smoothing for maximum entropy based intonational phrase prediction in Chinese,” Proceedings of the IEEE ICASSP 2005, Vol. **1**, pp. 285–288.
- [51] Z.-P. Zhao, T.-J. Zhao, and Y.-T. Zhu, “A maximum entropy Markov model for prediction of prosodic phrase boundaries in Chinese TTS,” Proceedings of the IEEE GrC 2007, pp. 498–498.
- [52] C. W. Wightman and M. Ostendorf, “Automatic labeling of prosodic patterns,” *IEEE Trans. Speech Audio Process.* **2**, 469–481(1994).
- [53] K. Chen, M. Hasegawa-Johnson, and A. Cohen, “An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model,” Proceedings of the IEEE ICASSP 2004, Vol. **1**, pp. 509–512.

- [54] V. Rangarajan, S. Narayanan, and S. Bangalore, “Acoustic-syntactic maximum entropy model for automatic prosody labeling,” Proceedings of the IEEE Spoken Language Technology Workshop 2006, pp. 74–77.
- [55] X.-J. Ma, W. Zhang, Q. Shi, W.-B. Zhu, and L.-Q. Shen, “Automatic prosody labeling using both text and acoustic information,” Proceedings of the IEEE ICASSP 2003, Vol. 1, pp. 516–519.
- [56] A. F. Muller, H. G. Zimmermann, and R. Neuneier, “Robust generation of symbolic prosody by a neural classifier based on autoassociators,” Proceedings of the IEEE ICASSP 2000, Vol. 3, pp. 1285–1288.
- [57] J.-H. Tao, “Acoustic and linguistic information based Chinese prosodic boundary labeling,” Proceedings of the TAL 2004, pp. 181–184.
- [58] S.-H. Chen, W.-H. Lai, and Y.-R. Wang, “A statistics-based pitch contour model for Mandarin speech,” J. Acoust. Soc. Am. **117**, pp. 908–925(2005).
- [59] S.-H. Chen, W.-H. Lai, and Y.-R. Wang, “A new duration modeling approach for Mandarin speech,” IEEE Trans. Speech Audio Process. **11**, 308–320 (2003).
- [60] C.-R. Huang, K.-J. Chen, F.-Y. Chen, Z.-M. Gao, and K.-Y. Chen, “Sinica Treebank: Design criteria, annotation guidelines, and pn-line interface,” Proceedings of the Second Chinese Language Processing Workshop 2000, pp. 29–37.
- [61] Young, S. J., Evermann, G., Gales, M. J. F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P. C., 2006. The HTK Book, version 3.4. Cambridge University Engineering Department, Cambridge, UK.
- [62] K. Sjlinder and J. Beskow, “Wavesurfer - an open source speech tool,” In Proceeding of the ICSLP 2000, Vol. 4, pp. 464-467.
- [63] S.-H. Chen and Y.-R. Wang, “Vector quantization of pitch information in Mandarin speech,” IEEE Trans. Commun. **38**, 1317–1320 (1990).
- [64] Y. Qian and W.-Y. Pan, “Prosodic word: The lowest constituent in the Mandarin prosody processing,” Proceedings of the ISCA International Conference on Speech Prosody 2002, pp. 591–594.
- [65] J.-H. Tao, H.-G. Dong, and S. Zhao, “Rule learning based Chinese prosodic phrase prediction,” Proceedings of the IEEE NLP-KE 2003, pp. 425–432.
- [66] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees* (Wadsworth, Belmont, CA, 1984).
- [67] Y.-R. Chao, *A Grammar of Spoken Chinese* (Berkeley Press, Berkeley, CA, 1968).
- [68] L.-S. Lee, C.-Y. Tseng, and M. Ouh-Young, “The synthesis rules in a Chinese

- text-to-speech system,” *IEEE Trans. Acoust., Speech, Signal Process.* **37**, 1309–1320 (1989).
- [69] Y. Xu, “Contextual tonal variations in Mandarin,” *J. Phonetics* **25**, 61–83 (1997).
- [70] A.-J. Li, Y.-Q. Zu, and Z.-Q. Li, “A national database design and prosodic labeling for speech synthesis,” *Proceedings of the Oriental COCODA Workshop 1999*, pp. 13–16.
- [71] A.-J. Li and M.-C. Lin, “Speech corpus of Chinese discourse and the phonetic research,” *Proceedings of the ICSLP 2000*, Vol. **4**, pp. 13–18.
- [72] K.-J. Chen and C.-R. Huang, “Part of speech (POS) analysis on Chinese language,” CKIP Technical Report No. 93-05, Institute of Information Science, Academia Sinica, Taiwan, R.O.C., 1993 (in Chinese).
- [73] Chinese knowledge Information Processing (CKIP), Academia Sinica, “An introduction to Academia Sinica balanced corpus for modern Mandarin Chinese,” CKIP Technical Report No. 95-02, Institute of Information Science, Academia Sinica, Taiwan, R.O.C. 1995 (in Chinese).
- [74] C. Shih, “Declination in Mandarin,” *Proceedings of the ESCA Workshop on Intonation: Theory, Models and Applications 1997*, pp. 293–296.
- [75] Y. Yufang and W. Bei, “Acoustic correlates of hierarchical prosodic boundary in Mandarin,” *Proceedings of the ISCA International Conference on Speech Prosody 2002*, pp. 707–710.
- [76] C.-Y. Tseng and S.-H. Pin, “Mandarin Chinese prosodic phrase grouping and modeling: Method and implications,” *Proceedings of the TAL 2004*, pp. 193–196.
- [77] C.-Y. Tseng and S.-H. Pin, “Modeling prosody of Mandarin Chinese fluent speech via phrase grouping,” *Proceedings of the Speech and Language Systems for Human Communication (SPLASH-2004/Oriental-COCOSDA2004)*, 2004, pp. 53–57.
- [78] C.-Y. Tseng and Z.-Y. Su, “Corpus approach to phonetic investigation—Methods, quantitative evidence and findings of Mandarin speech prosody,” *Proceedings of the Oriental COCODA Workshop 2006*, pp. 123–138.
- [79] S. Theodoridis and K. Koutroumbas, *Pattern Recognition* 2nd ed.(Elsevier, London, UK, 2003).
- [80] R.-H. Wang, Q. Liu, D. Tang, “A New Chinese Text-To-Speech System with High Naturalness,” *Proceedings of the ICSLP 1996*, pp. 1441–1444.
- [81] Yu Hu, Bo Yin, Xiaoru Wu, “KD2000 Chinese Text-To-Speech System,” *Proceedings of the ICMI 2000*. pp.300~307.
- [82] A.-W. Black, A.-J. Hunt, “Generating F0 contours from ToBI labels using linear

regression,” Proceedings of the ICSLP 1996, pp. 1385-1388.

- [83] Q. Guo, N. Katae, “Duration Prediction in Mandarin TTS System,” Proceedings of the ISCA International Conference on Speech Prosody 2006.
- [84] C. H. Wu and J. H. Chen, “Automatic generation of synthesis units and prosodic information for Chinese concatenative synthesis,” *Speech Commun.* **35**, 219–237 (2001).
- [85] K.-N. Ross, M. Ostendorf, "A dynamical system model for generating fundamental frequency for speech synthesis," *IEEE Trans. Speech Audio Process.* Vol.7, no.3, pp.295-309 (1999).
- [86] Chou Fu-chiang, Tseng Chiu-yu and Lee Lin-Shan, “Automatic Generation of Prosodic Structure for High Quality Mandarin Speech Synthesis,” Proceedings of the ICSLP 1996, pp.1624-1627.
- [87] X. D. Huang, A. Acero, J. Adcock, H.-W. Hon, J Goldsmith, J. Liu and M. Plumpe, “Whistler: a trainable text-to-speech system,” Proceedings of the ICSLP 1996, pp. 2387-2390.





## Appendix A

### (1) Determinations of $Th1$ , $Th2$ , and $Th3$

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

### (2) Determination of $Th5$

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

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

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

$f_{PM}(\xi_n) > f_{intra}(\xi_n)$  which can exclude non-PM inter-word junctures with normalized log- $F0$  level jump similar to intra-word juncture. Lastly, the threshold  $Th5$  is set as the equal-probability intersection of  $f_{intra}(\xi)$  and  $f_{B2-1}(\xi)$ .

### (3) Determinations of $Th4$ and $Th6$

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

### (4) Determination of $Th7$ and $Th8$

The normalized syllable duration lengthening thresholds,  $Th7$  and  $Th8$ , are set to distinguish between  $B2-3$  and  $B0/B1$  for inter-word junctures with normalized syllable duration lengthening factors 1 and 2 (i.e.  $dl_n$  and  $df_n$ ) greater than  $Th7$  and  $Th8$ . Four empirical Gaussian distributions of normalized duration lengthening factors,  $\{f_{intra}^{dl}(\tau) / f_{intra}^{df}(\tau)\}$  and  $\{f_{PM}^{dl}(\tau) / f_{PM}^{df}(\tau)\}$ , for intra-word and PM junctures are constructed using all samples of intra-word syllable junctures and all PM junctures, respectively. We then construct an empirical Gaussian distributions of normalized syllable duration lengthening factors  $\{f_{B2-3}^{dl}(\tau) / f_{B2-3}^{df}(\tau)\}$  for  $B2-3$  by using all samples of inter-word junctures without PM but with apparent normalized duration syllable lengthening factors. The condition of apparent normalized syllable duration lengthening factors are evaluated based on the criterion of  $f_{PM}^{dl}(dl_n) > f_{intra}^{dl}(dl_n)$  and  $f_{PM}^{df}(df_n) > f_{intra}^{df}(df_n)$  which can exclude non-PM inter-word junctures with normalized syllable duration lengthening factors similar to intra-word juncture. Lastly, the threshold  $Th7$  and  $Th8$  are set as the equal-probability intersection of  $\{f_{intra}^{dl}(\tau)$  and  $f_{B2-3}^{dl}(\tau)\}$  and  $\{f_{intra}^{df}(\tau)$  and  $f_{B2-3}^{df}(\tau)\}$ , respectively.

## Appendix B

Table B.1: The contextual linguistic features considered in this study. Note that the notations of POS symbols follow Ref. [72].

<b>Type of following syllable's initial:</b> null initial, $\{m, n, l, r\}$ , $\{b, d, g\}$ , $\{f, s, sh, shi, h\}$ , $\{ts, ch, chi\}$ , $\{p, t, k\}$ , $\{tz, j, ji\}$
<b>Type of syllable boundary:</b> inter-word, Type-1 intra-word, Type-2 intra-word. Here, Type-1 intra-words represent normal ones, while Type-2 intra-words are specific intra-word locations of some special long words which have potential to be pronounced with pauses.
<b>Type of PM:</b> period, exclamation mark, semicolon, and question mark, comma, dun hao, colon
<b>Length of preceding/following word in syllable:</b> 1, 2, 3, 4, >4
<b>Broad class of preceding/following word:</b> substantive word, function word
<b>Level-1 POS of preceding/following word:</b> A (adjective), C (conjunction), D (adverbial), N (noun), I (interjection), P (preposition), T (particle), V (verb), DE ( <i>de, zhi, di</i> ), SHI (shi), DM (determiner-measure compound)
<b>Level-2 POS of preceding/following word:</b> Ca (coordinate conjunction), Cb (correlative conjunction), Da (adverb of quantity), Db (adverb of evaluation), Dc (negation), Dd (adverb of time), Df (adverb of degree), Dg (adverb of place), Dh (adverb of manner), Di (aspectual adverb), Dj (interrogative adverb), Dk (sentential adverb), Na (general noun), Nb (special noun), Nc (place noun), Nd (time noun), Ne (determiner), Nf (measure), Ng (localizer), Nh (pronoun), VA (active intransitive verb), VB (active pseudo-transitive verb), VC (active transitive verb), VD (ditransitive verb), VE (active verb with a sentential object), VF (active verb with a verbal object), VG (classificatory verb), VH (stative intransitive verb), VI (stative pseudo-transitive verb), VJ (stative transitive verb), VK (stative verb with a sentential object), VL (stative verb with a verbal object), V 2 ( <i>you</i> )
<b>Level-3 POS of preceding/following word:</b> Caa (conjunctive conjunction), Cab (listing conjunction), Cba (movable before correlative conjunction), Cbb (unmovable before correlative conjunction), Dfa (pre-verbal degree adverbs), Dfb (post-verbal degree adverbs), Ncd (localizer), Neu (numeral determiner), Nes (specific determiner), Nep (anaphoric determiner), Neq (quantitative determiner), VA2 (active intransitive verb), VC1 (active transitive verb), VH16 (stative intransitive verb), VH22 (stative intransitive verb)
<b>Position of following syllable in a syntactic phrase:</b> beginning, otherwise
<b>Position of preceding syllable in a syntactic phrase:</b> ending, otherwise
<b>Number of syntactic phrase levels that preceding/following word terminates/initiates:</b> 0, >0, >1, >2
<b>Length of the smallest syntactic phrase covering both preceding and following words in syllable :</b> $\leq 2, >2, >3, >4, \dots, >15$
<b>Length of the largest syntactic phrase covering the following word but not the preceding word in syllable:</b> $\leq 2, >2, >3, >4, \dots, >10$
<b>Length of the largest syntactic phrase covering the preceding word but not the following word in syllable:</b> $\leq 2, >2, >3, >4, \dots, >10$

# Appendix C

## C.1 The question set $\Theta_1$

The question set  $\Theta_1$  used to construct the decision trees for building the break-acoustics model  $P(pd_n, ed_n | B_n, I_n)$  is listed below:

### 1. Syllable Level

$Q_1$ 1.1: Is the initial of the following syllable a null one or in  $\{m, n, l, r\}$ ?

$Q_1$ 1.2: Is the initial of the following syllable a null one?

$Q_1$ 1.3: Is the initial of the following syllable in  $\{b, d, g\}$ ?

$Q_1$ 1.4: Is the initial of the following syllable in  $\{f, s, sh, shi, h\}$ ?

$Q_1$ 1.5: Is the initial of the following syllable in  $\{m, n, l, r\}$ ?

$Q_1$ 1.6: Is the initial of the following syllable in  $\{ts, ch, chi\}$ ?

$Q_1$ 1.7: Is the initial of the following syllable in  $\{p, t, k\}$ ?

$Q_1$ 1.8: Is the initial of the following syllable in  $\{tz, j, ji\}$ ?

$Q_1$ 1.9: Is the inter-syllable location an inter-word?

$Q_1$ 1.10: Is the inter-syllable location a Type-1 intra-word?

$Q_1$ 1.11: Is the inter-syllable location a Type-2 intra-word?

### 2. PM

In the following questions, we define major PMs = {period, exclamation mark, semicolon, question mark} and minor PMs={comma, dun hao(a mark in Chinese punctuation used to set off items in a series), colon}.

$Q_1$ 2.1: Does a PMs exist at the inter-syllable location?

$Q_1$ 2.2: Does a major PM exist at the inter-syllable location?

$Q_1$ 2.3: Does a minor PM exist at the inter-syllable location?

$Q_1$ 2.4: Does a comma exist at the inter-syllable location?

$Q_1$ 2.5: Does a dot or colon exist at the inter-syllable location?

### 3. Questions related to tree-level linguistic features

All the following questions are subject to a prerequisite condition that the current inter-syllable location is an inter-word.

#### 3.1 Phrase beginning or ending

$Q_{13.1.1}$ : Are the preceding and following words at the same level of a tree?

$Q_{13.1.2}$ : Does the following word initiate a syntactic phrase? Here syntactic phrases include noun phrase (NP), verb phrase (VP), preposition phrase (PP), geographic phrase (GP), and clause (S).

$Q_{13.1.3}$ : Does the preceding word terminate a syntactic phrase?

#### 3.2 Number of syntactic phrase levels

$Q_{13.2.1\sim 3}$ : If the following word initiates a syntactic phrase, is the number of syntactic phrase levels greater than or equal to  $n \in \{1, 2, 3\}$ ?

$Q_{13.2.4\sim 6}$ : If the preceding word terminates a syntactic phrase, is the number of syntactic phrase levels greater than or equal to  $n \in \{1, 2, 3\}$ ?

#### 3.3 Number of syllable in a syntactic phrase

$Q_{13.3.1\sim 14}$ : Is the length of the smallest syntactic phrase covering both the preceding and following words in syllable greater than  $n \in \{2, 3, 4, \dots, 15\}$ ?

$Q_{13.3.15\sim 23}$ : Is the length of the largest syntactic phrase covering the following word but not the preceding word in syllable greater than  $n \in \{2, 3, 4, \dots, 10\}$ ?

$Q_{13.3.24\sim 32}$ : Is the length of the largest syntactic phrase covering the preceding word but not the following word in syllable greater than  $n \in \{2, 3, 4, \dots, 10\}$ ?

## C.2 The question set $\Theta_2$

The question set  $\Theta_2$  used to construct the decision trees for building the break-syntax model  $P(B_n | \mathbf{I}_n)$  is listed below:

### 1. Syllable Level

$Q_21.1$ : Is the initial of the following syllable a null one or in  $\{m, n, l, r\}$ ?

$Q_21.2$ : Is the inter-syllable location an inter-word?

$Q_2$ 1.3 : Is the inter-syllable location a Type-1 intra-word?

$Q_2$ 1.4 : Is the inter-syllable location a Type-2 intra-word?

## 2. Word Level

All the following questions are subject to a prerequisite condition that the current inter-syllable location is an inter-word.

### 2.1 PM

$Q_2$ 2.1.1~5 : The same questions as  $Q_1$ 2.1~5 .

### 2.2 Word length

$Q_2$ 2.2.1~4 : Is the preceding word an  $n \in \{1, 2, 3, 4\}$ -syllable word?

$Q_2$ 2.2.5~8 : Is the following word an  $n \in \{1, 2, 3, 4\}$ -syllable word?

$Q_2$ 2.2.9 : Is the length of the preceding word in syllable greater than 4?

$Q_2$ 2.2.10 : Is the length of the following word in syllable greater than 4?

### 2.3 Substantive/function words

$Q_2$ 2.3.1~2 : Is the preceding word a substantive word/function words?

$Q_2$ 2.3.3~4 : Is the following word a substantive word/function words?

### 2.4 Level-1 POS and special tags

$Q_2$ 2.4.1~11 : Is the POS of the preceding word A/C/D/N/I/P/T/V/DE/SHI/DM?

$Q_2$ 2.4.12~22 : IS the POS of the following word A/C/D/N/I/P/T/V/DE/SHI/DM?

### 2.5 Level-2 POS

$Q_2$ 2.5.1~33 : Is the POS of the preceding word Ca/Cb/Da/Db/Dc/Dd/Df/Dg/Dh/Di/Dj/Dk/Na/Nb/Nc/Nd/Ne/Nf/Ng/Nh/VA/VB/VC/V D/VE/VF/VG/VH/VI/VJ/VK/VL/V\_2?

$Q_2$ 2.5.34~66 : Is the POS of the following word Ca/Cb/Da/Db/Dc/Dd/Df/Dg/Dh/Di/Dj/Dk/Na/Nb/Nc/Nd/Ne/Nf/Ng/Nh/VA/VB/VC/V D/VE/VF/VG/VH/VI/VJ/VK/VL/V\_2?

### 2.6 Level-3 POS

$Q_2$ 2.6.1~15 : Is the POS of the preceding word Caa/Cab/Cba/Cbb/Dfa/Dfb/Ncd/Neu/Nes/Nep/Neq/VA2/VC1/VH16/VH22?

$Q_2$ 2.6.16~30 : Is the POS of the following word Caa/Cab/Cba/Cbb/Dfa/Dfb/Ncd/Neu/Nes/Nep/Neq/VA2/VC1/VH16/VH22?

### 2.7 Combination of POS

$Q_2$ 2.7.1~7: Does the POS of the preceding word belong to {Da, Db, Dc, Dd, Dg, Dh, Di, Dj, Dk}/{Na, Nb, Nc}/{Ncd, Ng}/{I, T}/{VA, VG}/{VB, VC, VD, VE, VF, VJ, VK, VL}/{VH, VI}?

$Q_2$ 2.7.8~14: Does the POS of the following word belong to {Da, Db, Dc, Dd, Dg, Dh, Di, Dj, Dk}/{Na, Nb, Nc}/{Ncd, Ng}/{I, T}/{VA, VG}/{VB, VC, VD, VE, VF, VJ, VK, VL}/{VH, VI}?

### 3. Tree-level features

All tree-level features here are the same as the tree-level features used in the question set  $\Theta_1$ , i.e.,  $Q_2$ 3.1.1~3= $Q_1$ 3.1.1~3,  $Q_2$ 3.2.1~6= $Q_1$ 3.2.1~6 and  $Q_2$ 3.3.1~32= $Q_1$ 3.3.1~32.



# Appendix D

## D.1 The question set $\Theta_3$

The question set  $\Theta_3$  used to construct the decision trees for building the break-acoustics model  $p(pd_n, ed_n, pj_n, dl_n, df_n | B_n, \mathbf{I}_n)$  is listed below:

### 1. Syllable Level

$Q_3$ 1.1~11: The same questions as  $Q_1$ 1.1~11.

### 2. PM

$Q_3$ 2.1~5: The same questions as  $Q_1$ 2.1~5.

### 3. Questions related to sentence level features

All the following questions are subject to a prerequisite condition that the current inter-syllable location is an inter-word.

#### 3.1 Length of sentence

$Q_3$ 3.1.1~30: Is the length of the current sentence greater or equal to 1~30?

$Q_3$ 3.1.31~60: Is the length of the previous sentence greater or equal to 1~30?

$Q_3$ 3.1.61~90: Is the length of the following sentence greater or equal to 1~30?

#### 3.2 Distances to PM

$Q_3$ 3.2.1~15: Is the distance to the nearest previous PM in syllable greater or equal to 1~15?

$Q_3$ 3.2.16~30: Is the distance to the nearest following PM in syllable greater or equal to 1~15?

## D.2 The question set $\Theta_4$

The question set  $\Theta_4$  used to construct the decision trees for building the break-syntax model  $P(B_n | \mathbf{I}_n)$  is listed below:

### 1. Syllable Level

All the syllable level questions are identical to the syllable level question in Appendix C.2.



## **2. Word Level**

All the word level questions are identical to the word level question in Appendix C.2.

## **3. Questions related to sentence level features**

All the sentence level questions are identical to the sentence level question in Appendix D.1.



# Publication List

## Journal Paper

- [1] **Chen-Yu Chiang**, Sin-Horng Chen, Hsiu-Min and Yu, Yih-Ru Wang, "Unsupervised Joint Prosody Labeling and Modeling for Mandarin Speech," J. Acoust. Soc. Am. **125**, No. **2**, pp. 1164-1183(2009).

## Conference Papers

- [1] **Chen-Yu Chiang**, Yih-Ru Wang and Sin-Horng Chen, "On the Inter-syllable Coarticulation Effect of Pitch Modeling for Mandarin Speech," Proceedings of Interspeech 2005, Lisboa, Portugal, pp. 3269-3272.
- [2] **Chen-Yu Chiang**, Xiao-Dong Wang, Yuan-Fu Liao, Yih-Ru Wang, Sin-Horng Chen, Keikichi Hirose, "Latent Prosody Modeling Of Continuous Mandarin Speech," Proceedings of ICASSP 2007, Honolulu, USA, Vol. **IV**, pp. 625-628.
- [3] **Chen-Yu Chiang**, Hsiu-Min Yu, Yih-Ru Wang, Sin-Horng Chen, "An Automatic Prosody Labeling Method for Mandarin Speech," Proceedings of Interspeech 2007, pp. 494-497, Antwerp, Belgium
- [4] Chi-Feng Chen, **Chen-Yu Chiang**, Yih-Ru Wang, and Sin-Horng Chen, "A Study on Prosodic Modeling for Isolated Mandarin Words," Proceedings of ROCLING 2007, Taipei, ROC. pp. 273-286
- [5] **Chen-Yu Chiang**, Hsiu-Min Yu, Yih-Ru Wang, Sin-Horng Chen, "Exploration of High-level Prosodic Patterns for Continuous Mandarin Speech," Proceedings of ICASSP 2008, Las Vegas, USA, pp. 4381-4384
- [6] Hung-Kuang Shih, **Chen-Yu Chiang**, Yih-Ru Wang and Sin-Horng Chen, "Prosodic Modeling For Isolated Mandarin Words And Its Application," Proceedings of ISCSLP 2008, Kunming, China, pp. 1-4.

## Others

- [1] Xiao-Dong Wang, Jin-Song Zhang, Keikichi Hirose, Nobuaki Minematsu, **Chen-Yu Chiang**, Yih-Ru Wang, Yuan-Fu Liao, "Tone Recognition of Continuous Mandarin Speech Based on Tone Nucleus Model and Neural Network," IPSJ SIG Technical Reports, Vol. 2006, No.136(SLP-64), pp. 107-112.
- [2] 江振宇、蕭希群、余秀敏、廖元甫, "語音韻律簡介," 中華民國計算語言學學會通訊, 第十八卷第二期, 2007, pp. 5-19.

## 博士候選人資料

姓 名：江振宇

性 別：男

出生年月日：民國 69 年 3 月 9 日

籍 貫：台北市

學 歷：

國立交通大學電信工程學系學士班畢業(88 年 8 月~91 年 6 月)

國立交通大學電信工程研究所碩士班畢業(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~)

論文題目：

非監督式中文語音韻律標記及韻律模式

Unsupervised Joint Prosody Labeling and Modeling for Mandarin  
Speech