EXPLORATION OF HIGH-LEVEL PROSODIC PATTERNS FOR CONTINUOUS MANDARIN SPEECH

Chen-Yu Chiang[1], Hsiu-Min Yu[2], Yih-Ru Wang[1] and Sin-Horng Chen[1]

[1]Dept. of Communication Engineering, National Chiao Tung University, Taiwan
[2]Dept. of Foreign Languages and Literature, Chung Hua University, Taiwan

ABSTRACT

In this paper, the high-level prosodic patterns of prosodic word (PW), prosodic phrase (PPh) and breath group/prosodic phrase group (BG/PG) for syllable pitch-level and duration are explored using an automatic joint prosody labeling and modeling method. Experimental results on a treebank speech corpus showed that the explored high-level prosodic patterns not only matched well with our a priori knowledge about Mandarin prosody, but also conformed well to other previous studies. They can therefore be integrated to form a meaningful Mandarin prosody hierarchy.

Index Terms— speech processing, prosody modeling, Mandarin prosody

1. INTRODUCTION

Prosody modeling is an important research topic for text-to-speech (TTS). Its task is to explore the hierarchical structure of the prosody of a language. For Mandarin speech, the conventional "…small ripples riding on large waves" theory [1] suggests that the tones are integrated with intonation just like small ripples riding on large waves. Recently, [2] proposed a five-layer prosody model: syllable, prosodic word, prosodic phrase, breath group, and prosodic phrase group. Based on the model, [2] explored syllable duration, energy level, and pitch patterns for prosodic units of each layer by using a well-annotated speech database with 5 break types being properly labeled manually [2]. In our previous study [3], an automatic joint prosody labeling and modeling method was proposed based on a four-layer model, which is a modification of the five-layer model, by using an unlabeled speech database. Two types of prosody tags, including 6 inter-syllable beak types and 16 prosodic states of syllable pitch level, were labeled. In this paper, we extend our previous study by taking these two types of prosody tags as primitive features to exploit the syllable pitch-level and duration patterns for prosodic units of the upper three layers, i.e., PW, PPh, and BG/PG. The construction of a quantitative prosody hierarchy for Mandarin speech is therefore completed.

The paper is organized as follows. Section 2 introduces the four-layer prosody model. Section 3 briefly reviews the findings of our previous study on the unsupervised joint prosody labeling and modeling for Mandarin speech. Section 4 presents the proposed formulations to explore the prosodic patterns of the upper three layers. Experimental results are discussed in Section 5. Some conclusions are given in the last section.

2. PROSODIC PHRASE STRUCTURE

Fig. 1 depicts a conceptual diagram of the four-layer prosody hierarchy of Mandarin speech [3] used in our previous and current studies. It is a modification of the five-layer model proposed in [2]. It consists of four layers: syllable (SYL), PW, PPh, and BG/PG. Two types of prosody tags, inter-syllable break type and prosodic state of syllable, are employed to characterize the prosodic units of these four layers. For the break type tag, we modify the 6-type break labeling scheme proposed by Tseng [2] 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 for dividing B2 into B2-1 and B2-2 is due to the difference in their acoustic characteristics, while the combination of B4 and B5 is owing to the similarity of their acoustic characteristics. A set of 6 break types, $\mathbf{\Lambda}$= {B0, B1, B2-1, B2-2, B3, B4}, is therefore adopted and used to delimit these four prosody layers. For the prosodic state tag, we regard the relative log-F0 level of a syllable in the log-F0 contour of an utterance as prosodic state for F0 modeling. Similarly, relative syllable duration is taken as prosodic state in the syllable duration model. A sequence of prosodic states can be regarded as an aggregation of log-F0/duration variation patterns of the four prosody layers. This study tries to separate the affections of these four layers and exploit the representative log-F0/duration patterns of PW, PPh, and BG/PG.

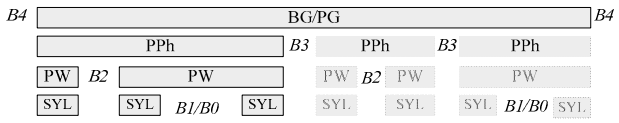

Fig. 1: A conceptual prosody hierarchy of Mandarin speech.

3. REVIEW OF THE UNSUPERVISED JOINT PROSODY LABELING AND MODELING

An unsupervised joint prosody labeling and modeling method was proposed in our previous study [3] to simultaneously label a large Mandarin speech database with the two types of prosody tags and build the relationship between the labeled prosody tags and some relevant linguistic features of the associated texts. Since its results are the foundation of the current study, we briefly review the main findings of the previous study as follows.

3.1. The Syllable Pitch Contour Modeling

The task of joint syllable pitch contour modeling and prosody tag labeling is formulated as a parametric optimization problem to find the best prosody tags ($\mathbf{B}^*,\mathbf{P}^*$) given with the input acoustic prosodic features ($\mathbf{SP},\mathbf{PD},\mathbf{PE}$) and linguistic features ($\mathbf{L},\mathbf{T}$):

$$\begin{aligned}\mathbf{B}^*,\mathbf{P}^* &= \arg\max_{\mathbf{B},\mathbf{P}} P(\mathbf{B},\mathbf{P}|\mathbf{SP},\mathbf{PD},\mathbf{PE},\mathbf{L},\mathbf{T}) \\ &= \arg\max_{\mathbf{B},\mathbf{P}} P(\mathbf{B},\mathbf{P},\mathbf{SP},\mathbf{PD},\mathbf{PE}|\mathbf{L},\mathbf{T}) \\ &= \arg\max_{\mathbf{B},\mathbf{P}} P(\mathbf{SP},\mathbf{PD},\mathbf{PE}|\mathbf{B},\mathbf{P},\mathbf{L},\mathbf{T}) P(\mathbf{B},\mathbf{P}|\mathbf{L},\mathbf{T})\end{aligned} \quad (1)$$

where $P(\mathbf{SP},\mathbf{PD},\mathbf{PE}|\mathbf{B},\mathbf{P},\mathbf{L},\mathbf{T})$ is a general prosody model

describing the variations of prosodic features (**SP**,**PD**,**PE**) controlled by (**B**,**P**) and (**L**,**T**); P(**B**,**P**|**L**,**T**) is a prosody-syntax model describing the relationship between (**B**,**P**) and (**L**,**T**); $\mathbf{B}=\{B_{k,n}\}$ is the set of break type tags of the whole speech corpus with $B_{k,n}\in\mathbf{\Lambda}$ representing the break type of the inter-syllable location following syllable n in utterance k (referred to as boundary (k,n) thereafter); $\mathbf{P}=\{p_{k,n}\}$ is the set of prosodic state tags of the speech corpus and $p_{k,n}$ represents the pitch prosodic state of syllable n in utterance k (referred to as syllable (k,n) thereafter); $\mathbf{SP}=\{\mathbf{sp}_{k,n}\}$ and $\mathbf{sp}_{k,n}$ is the vector of first four orthogonal expansion coefficients [4] representing the log-F0 contour of syllable (k,n); $\mathbf{PD}=\{pd_{k,n}\}$ and $pd_{k,n}$ is the pause duration of boundary (k,n); $\mathbf{PE}=\{pe_{k,n}\}$ and $pe_{k,n}$ is the energy dip of boundary (k,n); $\mathbf{L}=\{\mathbf{l}_{k,n}\}$ and $\mathbf{l}_{k,n}$ is the vector of contextual linguistic features around boundary (k,n); and $\mathbf{T}=\{t_{k,n}\}$ and $t_{k,n}$ is the tone of syllable (k,n).

The general prosody model P(**SP**,**PD**,**PE**|**B**,**P**,**L**,**T**) is then simplified by

$$P(\mathbf{SP},\mathbf{PD},\mathbf{PE}|\mathbf{B},\mathbf{P},\mathbf{L},\mathbf{T}) \approx \prod_{k=1}^{K}\prod_{n=1}^{N_k} P(\mathbf{sp}_{k,n}|p_{k,n},B_{k,n-1},B_{k,n},t_{k,n-1},t_{k,n},t_{k,n+1})P(pd_{k,n},pe_{k,n}|B_{k,n},\mathbf{l}_{k,n}) \quad (2)$$

where $P(\mathbf{sp}_{k,n}|p_{k,n},B_{k,n-1},B_{k,n},t_{k,n-1},t_{k,n},t_{k,n+1})$ is a syllable pitch contour model which describes the dependence of $\mathbf{sp}_{k,n}$ on the nearby prosody tags and tones; $P(pd_{k,n},pe_{k,n}|B_{k,n},\mathbf{l}_{k,n})$ is a pause acoustic model which describes the dependence of $pd_{k,n}$ and $pe_{k,n}$ on $B_{k,n}$ and some nearby linguistic features $\mathbf{l}_{k,n}$. Similarly, we simplify P(**B**,**P**|**L**,**T**) by

$$P(\mathbf{B},\mathbf{P}|\mathbf{L},\mathbf{T}) \approx \prod_{k=1}^{K}\left\{\left[P(p_{k,1})\prod_{n=2}^{N_k}P(p_{k,n}|p_{k,n-1},B_{k,n-1})\right]\left[\prod_{n=1}^{N_k}P(B_{k,n}|\mathbf{l}_{k,n})\right]\right\} \quad (3)$$

where $P(p_{k,1})$ is the initial prosodic state probability; $P(p_{k,n}|p_{k,n-1},B_{k,n-1})$ is the prosodic state transition probability; and $P(B_{k,n}|\mathbf{l}_{k,n})$ is a break-syntax model which describes the dependence of $B_{k,n}$ on some nearby linguistic features $\mathbf{l}_{k,n}$. In this study, the break-syntax model is trained by the decision tree method.

$P(\mathbf{sp}_{k,n}|p_{k,n},B_{k,n-1},B_{k,n},t_{k,n-1},t_{k,n},t_{k,n+1})$ is then elaborated to consider four major affecting factors. With an assumption that all affecting factors are combined additively, we have

$$\mathbf{sp}_{k,n}=\mathbf{sp}^{r}_{k,n}+\boldsymbol{\beta}_{t_{k,n}}+\boldsymbol{\beta}_{p_{k,n}}+\boldsymbol{\beta}^{f}_{B_{k,n-1},tp_{k,n-1}}+\boldsymbol{\beta}^{b}_{B_{k,n},tp_{k,n}}+\boldsymbol{\mu} \quad (4)$$

where $\mathbf{sp}^{r}_{k,n}$ is the normalized (i.e., residual) pitch contour; $\boldsymbol{\beta}_{t_{k,n}}$ and $\boldsymbol{\beta}_{p_{k,n}}$ are the affecting patterns of $t_{k,n}$ and $p_{k,n}$, respectively; $\boldsymbol{\beta}^{f}_{B_{k,n-1},tp_{k,n-1}}$ and $\boldsymbol{\beta}^{b}_{B_{k,n},tp_{k,n}}$ are the forward and backward coarticulation affecting patterns; $tp_{k,n}$ is the tone pair $(t_{k,n},t_{k,n+1})$; $\boldsymbol{\mu}$ is the global mean pattern. The model is further simplified by assuming that $\mathbf{sp}^{r}_{k,n}$ is normally distributed. $P(pd_{k,n},pe_{k,n}|B_{k,n},\mathbf{l}_{k,n})$ is also simplified and expressed by the product of a Gamma distribution for pause duration and a normal distribution for energy dip.

Lastly, the model is trained by a sequential optimization procedure using a large Mandarin speech database with all texts being manually parsed with syntactic trees. After well training, all parameters of the model are obtained and the whole database is properly labeled with the two types of prosody tags, i.e., inter-syllable break type and prosodic state of syllable pitch level.

3.2. The Syllable Duration Modeling

Given all inter-syllable break types being properly labeled by the syllable pitch contour modeling, the syllable duration model is constructed to consider some major affecting factors. Under the assumption that all affecting factors are combined additively and we have

$$sd_{k,n}=sd^{r}_{k,n}+\gamma_{t_{k,n}}+\gamma_{s_{k,n}}+\gamma_{q_{k,n}}+\mu_{d} \quad (5)$$

where $sd_{k,n}$ and $sd^{r}_{k,n}$ represent observed syllable duration and residual duration respectively; $\gamma_{t_{k,n}}$, $\gamma_{s_{k,n}}$, $\gamma_{q_{k,n}}$ and μ^{d} are respectively, affecting factors of tone, base syllable type, duration prosodic state (treated as a latent variable) and global mean. The residual duration $sd^{r}_{k,n}$ is further modeled using a Gaussian distribution $N(sd^{r}_{k,n};0,R_d)$. The duration model is trained by the expectation-maximization (EM) training algorithm.

3.3. Joint Prosody Modeling and Labeling: some findings

An unlabeled Mandarin speech database containing read speech of a single female professional announcer was used to train both syllable log-F0 contour and duration models. Its texts were all short paragraphs composed of several sentences selected from the Sinica Treebank Corpus [5]. The database consisted of 380 utterances with 52192 syllables in total. The numbers of pitch/duration prosodic states were empirically set to be 16.

Fig. 2 displays the distributions of pause duration and energy dip for 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. These conformed to our knowledge about break types.

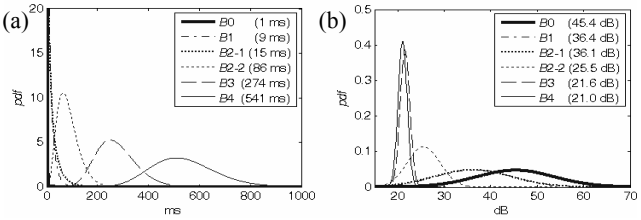

Fig. 2: The pdfs of (a) pause duration and (b) energy dip for these 6 break types. Numbers in () denote the mean values.

Fig. 3 displays the affecting factors of 16 prosodic states and their distributions for syllable log-F0 level and duration, respectively. As shown in the figure, those affecting factors spanned widely to cover the whole dynamic ranges of syllable F0 level and duration variations with lower indices of prosodic state representing lower log-F0 levels and shorter syllable durations, respectively.

Based on the break type labeling, we can divide the syllable sequences of all utterances into sequences of three prosodic units of PW, PPh and BG/PG to form a four-layer prosody hierarchical structure. According to the histograms displayed in Fig. 4, the

length of each of the three prosodic units mainly spans respectively from 1 to 12 syllables for PWs, from 1 to 20 syllables for PPhs, and from 1 to 60 syllables for BG/PGs. Statistics in Table 1 shows that the average lengths for these three types of prosodic units 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. These data approximately matched the results reported in [6-9].

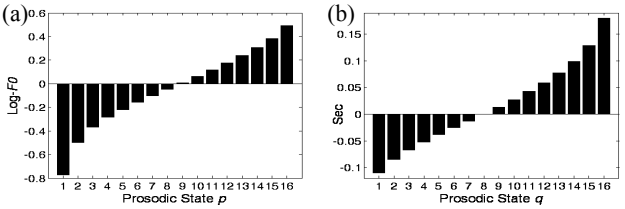

Fig. 3: The affecting patterns of 16 prosodic states of syllable (a) F0-level and (b) duration.

Table 1: Statistics of three types of prosodic units. Value in parentheses denotes standard deviation.

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

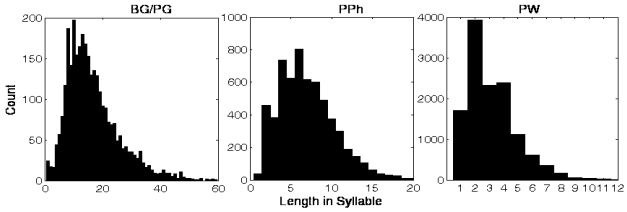

Fig. 4: Histograms of lengths for BG/PG, PPh and PW.

4. CONSTRUCTION OF PROSODIC PATTERNS

We now explore the F0-level patterns for the three prosodic units of PW, PPh, and BG/PG. First, we extract the affecting patterns of prosodic state from the observed syllable F0 levels $\mathbf{sp}_{k,n}(1)$ by eliminating the SYL-layer influence which is realized by the affections of the current tone, the coarticulations from the two nearest neighboring tones, and the global mean, i.e.,

$$pm_{k,n} = \mathbf{sp}_{k,n}(1) - \boldsymbol{\beta}_{t_{k,n}}(1) - \boldsymbol{\beta}^{f}_{B_{k,n-1},tp_{k,n-1}}(1) - \boldsymbol{\beta}^{b}_{B_{k,n},tp_{k,n}}(1) - \boldsymbol{\mu}(1) \quad (6)$$

where $\mathbf{x}(1)$ denotes the first component of vector $\mathbf{x}$ (i.e. log-F0 level).

A sequence of $pm_{k,n}$ delimited by B2-1/B2-2/B3/B4 at both sides is then regarded as a prosodic state pattern formed by integrating the F0-level patterns of the three prosodic units we considered. A model of pitch prosodic state pattern is therefore defined by

$$pm_{k,n} = pm^{r}_{k,n} + \beta_{PW_{k,n}} + \beta_{PPh_{k,n}} + \beta_{BG/PG_{k,n}} \quad (7)$$

where $pm^{r}_{k,n}$ is the residual of F0-level at syllable (k,n); $\beta_{PW_{k,n}}$ is the F0-level pattern of PW with $PW_{k,n}=(i,j)$ denoting that syllable (k,n) is located at the j-th place of an i-syllable PW; $\beta_{PPh_{k,n}}$ is the F0-level pattern of PPh with $PPh_{k,n}=(i,j)$ denoting that syllable (k,n) is located at the j-th place of an i-syllable PPh; and $\beta_{BG/PG_{k,n}}$ is the F0-level pattern of BG/PG with $BG/PG_{k,n}=(i,j)$ denoting that syllable (k,n) is located at the j-th place of an i-syllable BG/PG.

Similarly, we extracted the duration prosodic state patterns from $sd_{k,n}$ by eliminating the SYL-layer influence realized by the affections of current tone, base syllable type and the global mean:

$$dm_{k,n} = sd_{k,n} - \gamma_{t_{k,n}} - \gamma_{s_{k,n}} - \mu_d \quad (8)$$

A model of duration prosodic state pattern is then defined by

$$dm_{k,n} = dm^{r}_{k,n} + \gamma_{PW_{k,n}} + \gamma_{PPh_{k,n}} + \gamma_{BG/PG_{k,n}} \quad (9)$$

where $dm^{r}_{k,n}$ is the residual syllable duration, and $\gamma_{PW_{k,n}}$, $\gamma_{PPh_{k,n}}$ and $\gamma_{BG/PG_{k,n}}$ are duration patterns of PW, PPh and BG/PG, respectively.

A sequential optimization procedure based on the MMSE criterion is adopted to train these two models. It first defines two error functions, respectively, for pitch and duration modeling by

$$E_p = \sum_{k=1}^{K}\sum_{n=1}^{N_k}\left(pm_{k,n} - \beta_{PW_{k,n}} - \beta_{PPh_{k,n}} - \beta_{BG/PG_{k,n}}\right)^2 \quad (10)$$

$$E_d = \sum_{k=1}^{K}\sum_{n=1}^{N_k}\left(dm_{k,n} - \gamma_{PW_{k,n}} - \gamma_{PPh_{k,n}} - \gamma_{BG/PG_{k,n}}\right)^2 \quad (11)$$

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

5. EXPERIMENTAL RESULTS

Fig. 5 displays the learning curves of the sequential optimization process of exploring prosodic patterns of syllable F0-level and duration. It can be seen from the figure that the optimization process converged around 6 iterations.

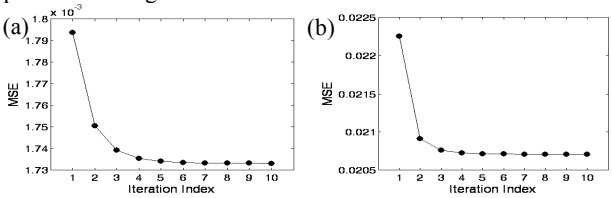

Fig. 5: The learning curve of the sequential optimization procedure for (a) syllable F0-level and (b) syllable duration modeling.

Fig. 6 displays the syllable F0-level patterns of $\beta_{PW_{k,n}}$, $\beta_{PPh_{k,n}}$, and $\beta_{BG/PG_{k,n}}$ 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 F0-level patterns. As shown in Fig. 6(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 Fig. 6(b), we find that short β_{PPh} had rising-falling patterns, while long β_{PPh} had rising-falling-sustaining-falling patterns. Moreover, they had smaller dynamic range spanning approximately in [-0.07, 007]. Lastly, we find from Fig. 6(c) that short β_{PW} showed high-falling patterns, while long β_{PW} showed falling-sustaining-falling patterns. Their dynamic range spanned approximately from -0.1 to 0.1. All these three types of F0-level patterns generally agree with our knowledge about Mandarin prosody.

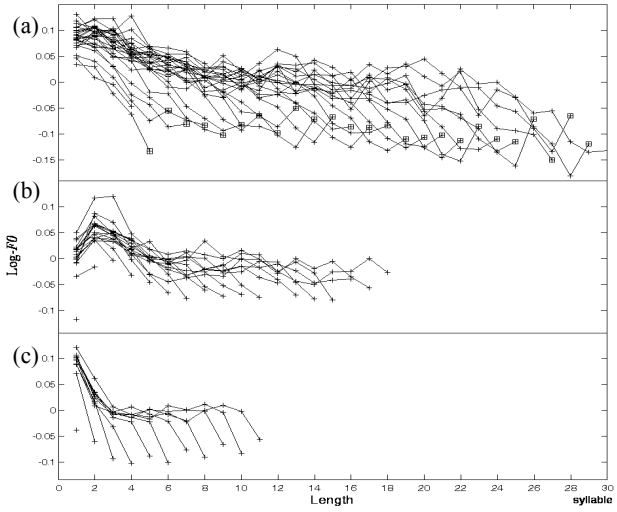

Fig. 6: The syllable F0-level patterns of (a) BG/PG, (b) PPh, and (c) PW. "□" in (a) indicates the ending syllable of a F0-level pattern.

Fig. 7 displays the syllable duration patterns of $\gamma_{PW_{k,n}}$, $\gamma_{PPh_{k,n}}$ and $\gamma_{BG/PG_{k,n}}$ with different lengths. It can be clearly observed from Fig. 7 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 [2].

Table 2 displays the total residual error (TRE), which is the percentage of sum-squared residue over the observed sum-squared log-F0/syllable duration, with respect to the use of different combination of affecting factors. It can be found from the table that TRE reduced as more affecting factors were used. The low level affecting factors/linguistic features (i.e., tone, coarticulation and base-syllable) accounted for 59.4% and 36.6% of prosodic variation in pitch and duration, respectively. However, high-level prosodic units (i.e. PW+ PPh + BG/PG) only contributed 17.2% (40.6% - 23.4%) and 18.1% for pitch and duration, respectively. By further investigating the contributions of high-level prosodic patterns, we find that the most significant one is PW. These results also matched well with the findings of [2].

Table 2: Total residual errors (TRE) w.r.t. the use of different combinations of affecting factors for pitch/duration modeling

Pitch Modeling		Duration Modeling	
Affecting factors	TRE	Affecting factors	TRE
+ Tone	46.3%	+ Base Syllable	69.1%
+ Coarticulation	40.6%	+ Tone	63.4%
+ PW	32.2%	+ PW	53.7%
+ PPh	28.3%	+ PPh	48.4%
+ BG/PG	23.4%	+ BG/PG	45.3%

5. CONCLUSIONS

In this paper, we exploited high-level prosodic patterns of PW, PPh and BG/PG for syllable pitch-level and duration using an automatic joint prosody labeling and modeling method. Experimental results on a treebank speech corpus confirmed that most prosodic patterns found were linguistic meaningful. More sophisticated exploration of prosodic patterns of pitch and duration as well as the extension to energy modeling are worth further studying.

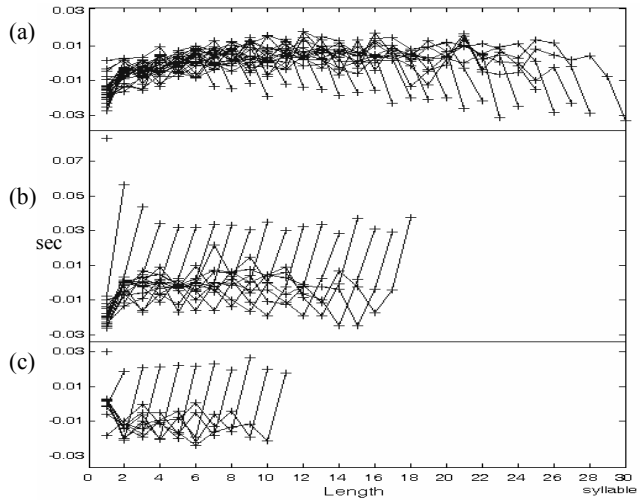

Fig. 7: Syllable duration patterns of (a)BG/PG, (b)PPh, and (c) PW.

ACKNOWLEDGEMENTS

This work was supported in part by NSC under contract NSC95-2218-E-002-027 and NSC95-2752-E009-014-PAE. The authors also want to thank Academia Sinica, Taiwan for providing the Treebank Corpus.

REFERENCES

[1] Y.R. Chao, 1968, "A Grammar of Spoken Chinese," University of California Press, Berkeley, Los Angeles, CA.

[2] C. Y. Tseng, S. H. Pin, Y. L. Lee, H. M. Wang and Y. C Chen, "Fluent speech prosody: framework and modeling," Speech Communication, Vol.46, Issues 3-4 (July 2005), Special Issue on Quantitative Prosody Modelling for Natural Speech Description and Generation, 284-309.

[3] C. Y. Chiang, H. M. Yu, Y. R. Wang, S. H. Chen, "An Automatic Prosody Labeling Method for Mandarin Speech," Proc. of Interspeech 2007, Antwerp, Belgium, pp. 494-497.

[4] S. H. Chen and Y. R. Wang, "Vector Quantization of Pitch Information in Mandarin Speech," IEEE Trans. Communications, vol. 38, no.9, pp.1317-1320, Sept. 1990.

[5] C. R. Huang, K. J. Chen, F. Y. Chen, Z. M. Gao and K. Y. Chen, "Sinica Treebank: Design Criteria, Annotation Guidelines, and On-line Interface," Proc. of 2nd Chinese Language Processing Workshop, pp.29-37, 2000, Hong Kong

[6] C. Y. Tseng, "Recognizing Mandarin Chinese Fluent Speech Using Prosody Information—An Initial Investigation," The 3rd Int. Conf. on Speech Prosody 2006, Dresden, Germany.

[7] J. F. Cao(2000): "Rhythm of spoken Chinese - linguistic and paralinguistic evidences," in ICSLP-2000, vol.2, 357-360.

[8] Y. Qian and W. Y. Pan, "Prosodic Word: the Lowest Constituent in the Mandarin Prosody Processing," Proceeding of International Conference on speech prosody, P591-594, Aix-en-Provence, April 2002.

[9] J. H. Tao, H. H. Dong, S. Zhao, "Rule Learning Based Chinese Prosodic Phrase Prediction," Proceeding of International Conference on Natural Language Processing and Knowledge Engineering, Beijing, China, P425-432, Oct. 2003.