TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 12/22 pp500-509
Volume 13, Number 4, August 2008

Pause or No Pause?—Prosodic Phrase Boundaries Revisited

TSENG Chiu-Yu (郑秋豫)**, CHANG Chun-Hsiang (张俊祥)

Phonetics Lab, Institute of Linguistics, Academia Sinica, Taipei 11529

Abstract: This study presents evidence from analyses of the acoustic parameters of fluent continuous speech to show that within-paragraph prosodic phrase boundaries are related more to contrasts of neighborhood prosodic states rather than between-phrase pause durations; prosodic states receive more constraints from higher level discourse information. By revising a modular acoustic model by Tseng's hierarchical prosodic phrase grouping framework and examining the much varied prosodic phrase (PPh) boundary B3 within speech paragraph, we show that statistical accounts of layered contributions reveal distinct contrasts between boundary immediate duration and intensity patterns irrespective of pause duration. Contrasts of F0 contour patterns were also observed in these locations. Evidence was also obtained to illustrate how PPh boundary states are specified more by higher level discourse information than by lower level prosodic word construction. These combined results suggest that contrastive neighboring prosodic states are more significant cues to PPh boundaries than boundary pause duration. The results also help explain why in fluent speech between-phrase pause durations vary greatly, and can be applied to automatic speech segmentation.

Key words: fluent speech prosody; hierarchical prosody group; prosodic state; prosodic phrase; boundary break; discourse prosody; linear regression model

Introduction

In previous work we have collected various types of fluent Mandarin speech data from read narratives in COSPRO[1] and designed annotations on the basis of perceived boundary breaks in relation to prosodic units. Our hierarchical prosodic phrase grouping (HPG)[2-4] specifies multiple-phrase speech paragraphs as a significant discourse prosody unit above phrases. The COSPRO annotation[5] specifies 5 levels of within-paragraph boundary breaks, i.e., from lower levels upward the syllable (Syl) boundary B1, the prosodic words (PW) boundary B2, the prosodic phrase (PPh) boundary B3, the change-of-breath (breath group (BG)[6]) boundary B4, and the prosodic-group (PG) terminal boundary B5, where physical pauses apply from B2 to B5. We have shown from quantitative analyses of speech corpora that output prosody of multiple-speech paragraphs is not unrelated phrase strings, but rather the cumulative outcome of contributions from all prosodic layers specified by HPG[3,4]. Further, central to fluent speech prosody is the contribution from above-phrase higher level information related to discourse organization, in which phrases and sentences are all prosodic subunits of each speech paragraph, where speech paragraphs are subunits of spoken discourse. Among each and every prosodic level, prosodic boundaries in relation to discourse prosody organization are significant cues. Perceived boundary breaks are therefore significant prosodic units as well.

However, in a previous study[7] it was discovered

Received: 2007-10-24

** To whom correspondence should be addressed.
E-mail: cytling@sinica.edu.tw; Tel: 886-2-26525000-6143

that not all boundary breaks could be accounted for by pause durations. It was found from the consistently annotated speech data of 2 speakers at slightly different speaking rates (220 and 230 ms/syllable) that the higher level boundaries B4 and B5 all possessed pause duration over 330 ms (*m*=330, 520 ms for B4, SD=162, 124 ms; *m*=415, 595 ms for B5, SD=209, 109 ms, where *m* is the mean value and SD is the standard deviation), indicating that pause durations alone can be viewed as significant cues for BG and PG boundaries. However, boundary pauses of B3 varied considerably in duration (from 17-585, 21-538 ms at *m*=224, 248 ms, SD=150, 207 ms) from 0 to over 350 ms across speakers, indicating pause durations alone are not sufficient for PPh boundaries. Therefore, to develop automatic speech segmentation or recognition, pause durations are adequate cues to locate B4 and B5, and speech paragraphs as discourse units can be identified. Unfortunately, the rationale does not apply to within-paragraph prosodic phrase boundaries of type B3 since these cannot be located by pause duration. The question then is why the PPh boundary break B3 varies so much in duration across speakers and yet is still perceived consistently across transcribers.

We note nevertheless that the perception-based annotation makes examination of signal-perception discrepancies possible, especially when perceptions are consistent across transcribers. We therefore hypothesize that there must be cues in the speech signal other than pause durations that are significant to the PPh boundary, and are significant to the human ear as well. The same previous study also demonstrated, by including the boundary immediate prosodic state by one syllable in the immediate B3 neighborhood, that predictions of B3 were improved by 8.3%[7]. We therefore hypothesize now that B3 predictions can be further improved by including more neighborhood prosodic states in the prediction.

In the following sections we will show how the previous model can be revised to accommodate more boundary immediate syllable duration allocation patterns along the time domain, as well as intensity distribution patterns, and will compare newly obtained predictions from the same speech materials with those from the previous model.

1 Speech Data and Methodology

1.1 Speech data

The same Mandarin Chinese speech data used for previous analysis[5,7] were selected from Sinica COSPRO 0[1], i.e., one male and one female speaker (F051P and M051P). Both speakers were professional radio announcers under 35 years of age at the time of recording. Each speaker read text of 26 discourse pieces in sound proof chambers. The 26 discourse pieces ranged from 85 to 981 characters in length, amounting to a total of 11 602 syllables. The corpora were first automatically labeled for segmental identities using the HTK toolkit in SAMPA-T notation[5], and then manually tagged for perceived boundary breaks by trained transcribers using the Sinica COSPRO Toolkit[8]. The annotation results were spot-checked by professional transcribers both for segmental alignments as well as for inter-transcriber consistency.

1.2 Methods of speech data analysis

The speech data was analyzed in three steps: (1) Three acoustic parameters were extracted from annotated speech data, corresponding to pause, syllable duration, and intensity; (2) The derived acoustic parameters were then normalized; (3) The respective layered contributions specified by the HPG framework were obtained through a step-wise linear regression model. Figure 1 gives a flowchart showing the basic HPG analysis. Table 1 summarizes the derived acoustic features of both speakers, where μ and σ represent the mean and standard deviations of each acoustic feature (pause, duration, and intensity).

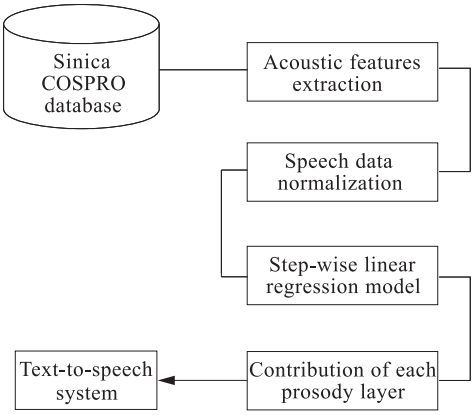

Fig. 1 Flowchart of analysis using the HPG framework

Table 1 Derived acoustic features by speaker

Speaker	μ_{Pause}	σ_{Pause}	μ_{Duration}	σ_{Duration}	$\mu_{\text{Intensity}}$	$\sigma_{\text{Intensity}}$
F051P	37	106	200	65	3.65	0.07
M051P	45	138	190	60	3.62	0.05

1.3 Speech data normalization

In order to eliminate between-speaker variations, each set of data was normalized with the mean and standard deviation of the entire class. The original method of normalization[4] is sensitive to extreme data, causing a shift in the normalized data distribution, and thereby making comparisons between speakers meaningless. To overcome the problem, the normalization was modified as follows:

$$Y_{\text{nor}}(i) = (Y(i) - \mu(Y)) / \sigma(Y),$$
$$Y_{\text{nor}} = \{Y_{\text{nor}}(1), Y_{\text{nor}}(2), ..., Y_{\text{nor}}(n)\},$$

where $Y(i)$ and $Y_{\text{nor}}(i)$ represent each datum in class Y and normalized class Y, and $\mu(Y)$ and $\sigma(Y)$ represent the mean and standard deviation in class Y. The same modification was made for each of the three acoustic features under consideration.

1.4 Revising the duration model

A syllable duration model corresponding to the HPG framework was constructed previously[7] to predict and locate boundary breaks B2 to B5 across continuous speech rather than simply predicting pauses. The predictions thus bear discourse information in relation to prosody organization specified by HPG. Higher level BG and PG boundary breaks (B4 and B5, respectively), which indicate multiple-phrase speech paragraphs across fluent continuous speech, could be easily located using pause durations alone, whereas lower level within-paragraph boundary breaks (B3 and B2 corresponding to PPh and PW, respectively) were predicted using both boundary break pause and duration information of one immediate neighboring syllable.

The goal of the present study is to revise the syllable duration model by altering both the syllable (the bottom) layer and the PW (the immediate higher) layer of the previous regression model to allow a better prediction of the PPh boundary B3. Using the same step-wise regression technique[2,3], a linear model with four layers[9,10] was modified and developed to predict speakers' timing behavior through temporal allocation of syllable duration modification. At the syllable layer, we used 6 consonant groups and 6 vowel groups to decrease the difference between groups. The revised syllable layer model could be written as Eq. (1):

Y_{nor} = Const+CCt+CVt+CTt+PCt+PVt+PTt+FCt+FVt+
FTt+2-way factors of each factor above+
3-way factors of each syllable above+
PW boundary constraint of each factor above+
Deltal (1)

In Eq. (1), we added a new condition that is constrained for each factor in the PW boundary to include co-articulation effects, such as tone sandhi, at the PW layer. The prefixes C, P, and F represent the current, preceding, and following syllable, respectively. Ct, Vt, and Tt represent consonant, vowel, and tone type, respectively. Residuals (Delta1) that cannot be predicted by the syllable layer are analyzed in the immediate higher layer.

Figure 2 shows the distribution of Delta1 (the residuals of the syllable layer) of the revised duration model from the speech data. The horizontal axis represents breaks from B1 to B5, and the vertical axis represents the residual value from –2 to 3. A significant difference ($p < 0.001$) was found with respect to the durations between the distributions of B2 and those

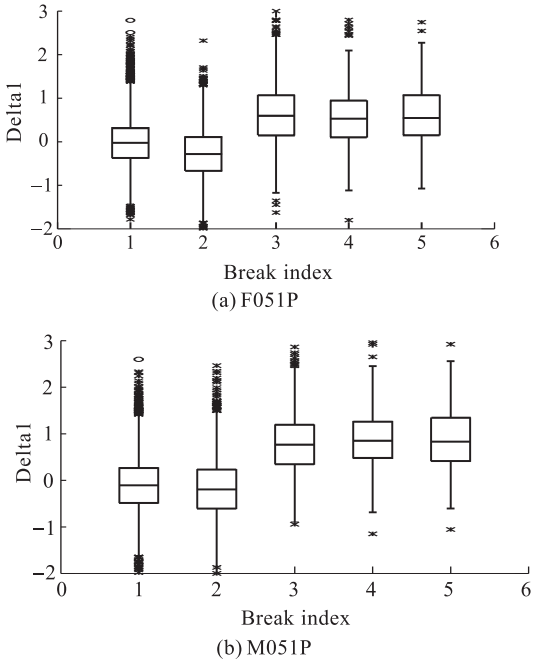

Fig. 2 Distribution of Delta1 of the revised duration model for speakers F051P and M051P. The block indicates the interval of standard deviation for one duration distribution and the beeline in the block denotes the mean value of the distribution.

greater than B2, as well as between speakers. The results enabled us to avoid overestimating the contribution of B2 from the PW layer, and thus we decided to add a constraint condition to only calculate f(PW) in the B2 level. The revised PW layer model can be written as Eq. (2):

$$\text{Delta1} = f(\text{PW length, PW sequence}) + \text{the calculation of } f(\text{PW}) \text{ is constrained in B2 level} + \text{Delta2} \quad (2)$$

The Delta2 residuals, which cannot be predicted by the PW layer, are again assumed to be contributions from the immediate higher level and are therefore subsequently analyzed at the next layer upward.

The PPh and BG layer models are the same as our previous models, given by Eqs. (3) and (4).

$$\text{Delta2} = f(\text{PPh length, PPh sequence}) + \text{Delta3} \quad (3)$$

$$\text{Delta3} = f(\text{BGIMF, PPh length, PPh sequence}) + \text{Delta4} \quad (4)$$

1.5 Revising the intensity model

Based on the revised duration model, we used the same method to analyze the characteristics of the intensity parameter. Figure 3 shows the distribution of Delta1 of the revised intensity model for both speakers. The horizontal axis again represents breaks from B1 to B5,

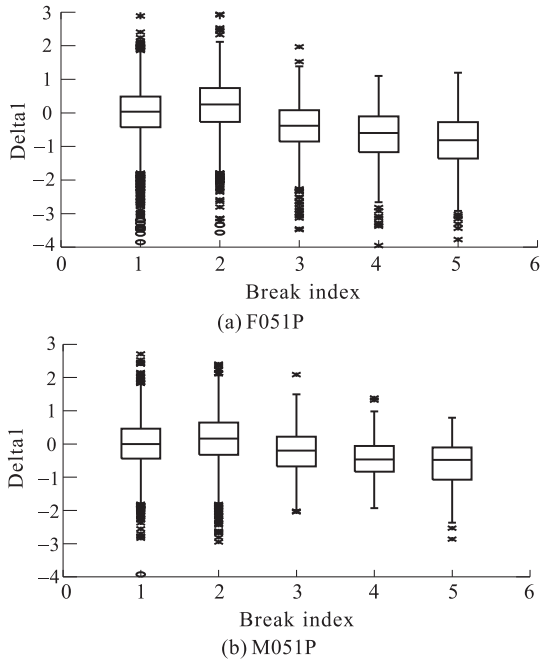

Fig. 3 Distribution of Delta1 of the revised intensity model for speakers F051P and M051P. The block indicates the interval of standard deviation for one duration distribution and the beeline in the block denotes the mean value of the distribution.

and the vertical axis represents the residual value from −4 to 3. A significant difference (p<0.001) was also found with respect to the intensity patterns between the distributions of B2 and those greater than B2, as well as between speakers. Therefore, the same rationale of modification can be applied to the revised intensity model as well.

1.6 Revising the pause model

In our previous pause model[7], we calculated the contribution of pauses from B1 to B4. However, we observed that the distribution of real pauses for B1 is very narrow and therefore decided that the contribution of B1 could be ignored in the revised model. Figure 4 shows the distribution of pauses from B2 to B4 for speakers F051P and M051P.

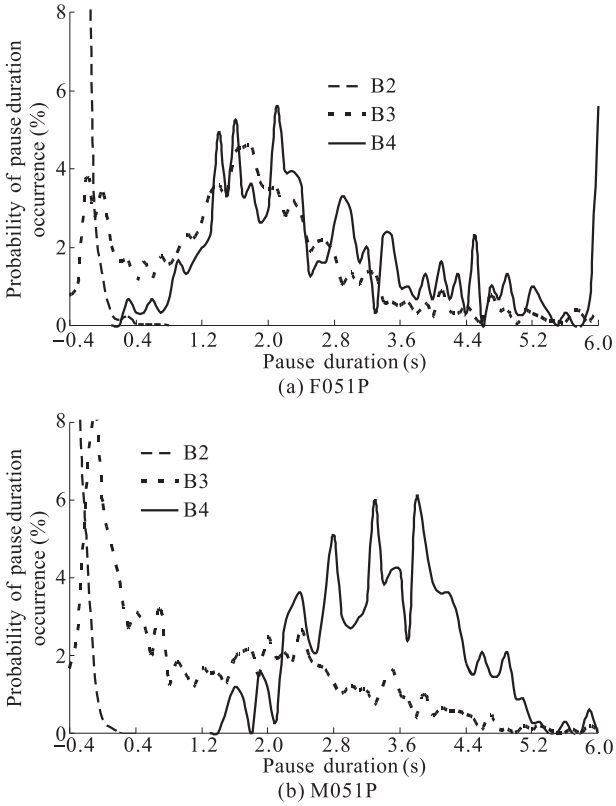

Fig. 4 Distribution of pauses as boundary breaks for speakers F051P and M051P

For the new pause model, the revised PW layer model can be written as

$$Y_{\text{nor}} = f(\text{PW length, PW sequence}) + \text{the calculation of } f(\text{PW}) \text{ is constrained in B2 level} + \text{Delta1} \quad (5)$$

The Delta1 residuals, which cannot be predicted by the PW layer, are analyzed in the immediate higher layer

subsequently.

We also analyzed the distribution of B3 pauses in relation to punctuation marks in the text used. B3 occurrences in the speech data in relation to punctuations of comma, period, and zero punctuation in text were analyzed; and their distributions were calculated (Fig. 5). The results indicate that the value of the B3 pause is indeed affected by the presence of punctuation marks, and that the mean values of B3 follow the sequence period > comma > no punctuation mark. In other words, although both speakers paused where no punctuation marks appeared in the text, the presence of punctuation marks resulted in more B3 occurrences in the speech data.

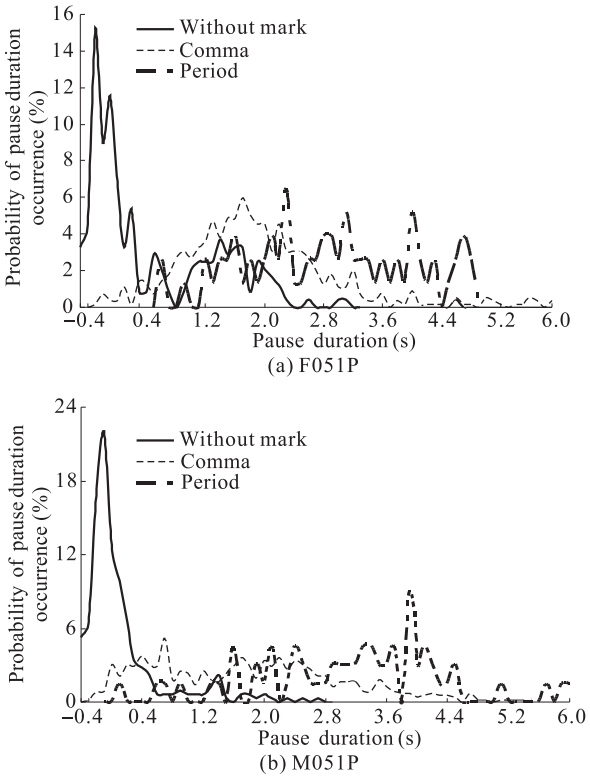

Fig. 5 Distribution of pauses as punctuation mark in B3 for F051P and M051P

According to these results, we categorized three groups of punctuation marks according to the mean values. In this way we can use punctuation marks as a feature of the PPh layer model, as shown in Eq. (6).

$$\text{Delta1} = f(\text{mark group, PPh length, PPh sequence}) + \text{Delta2} \quad (6)$$

The BG layer model is the same as our previous model,

$$\text{Delta2} = f(\text{BGIMF, PPh length, PPh sequence}) + \text{Delta3} \quad (7)$$

2 Results

2.1 Comparison of duration predictions

Figure 6 shows the duration patterns of PW (1-4 syllables in length) along the temporal course by syllable number and by speakers from the previous model[7], while Fig. 7 shows patterns from the current revised model. Each line represents the corresponding regression coefficient of one syllable at the specific position in a prosodic word. From Figs. 6 and 7 we can see that the PW patterns from the previous model show an opposite behavior to the revised models. In particular, the previous model showed final syllable lengthening of PW by syllable number and across speakers, whereas the revised model shows the reverse, namely, final syllable shortening PW by syllable number and across speakers. In general the results from the revised model attribute less contribution from the PW layer to the total output prediction.

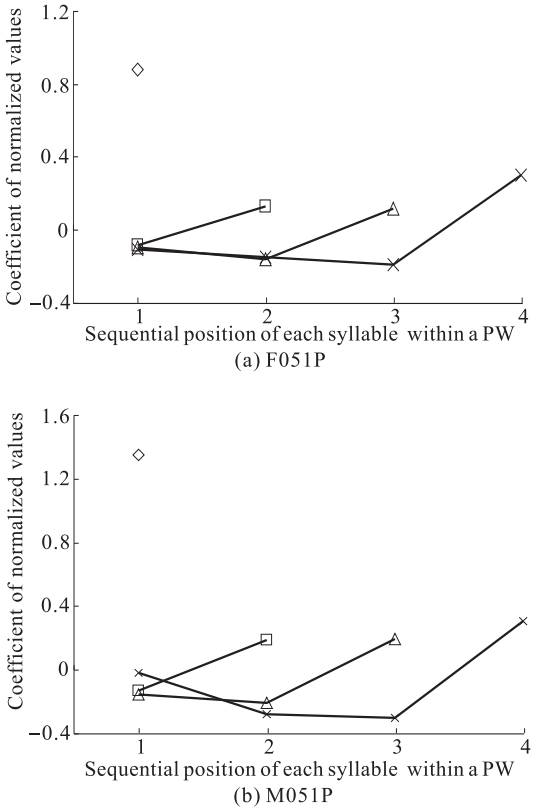

Fig. 6 PW patterns of the previous duration model for speakers F051P and M051P. PW's are from 1 to 4 syllables. 0 on vertical axis is defined as the mean of syllable duration. Each line represents the corresponding regression coefficient of one syllable at the specific position in a prosodic word.

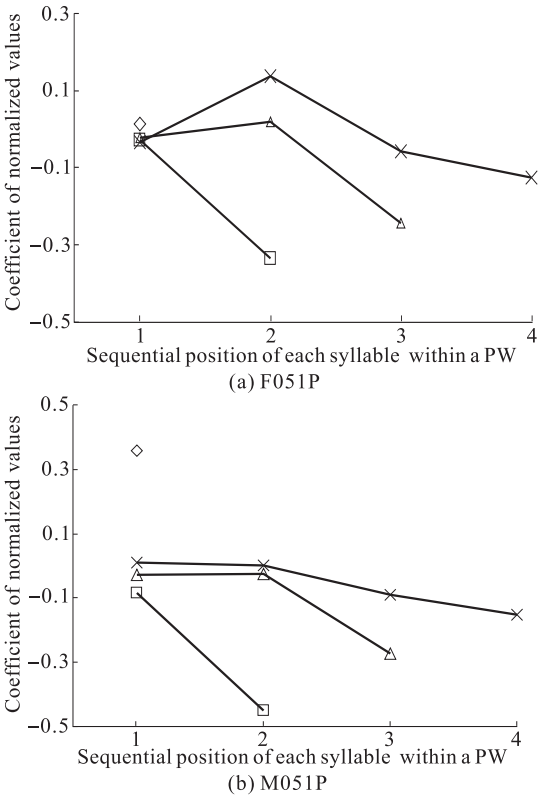

Fig. 7 PW patterns of the revised duration model for speakers F051P and M051P. PW's are from 1 to 4 syllables. 0 on vertical axis is defined as the mean of syllable duration. Each line represents the corresponding regression coefficient of one syllable at the specific position in a prosodic word.

Figure 8 shows the duration patterns of PPh (6-11 syllables in length) along the temporal course by syllable number and by speaker from the previous model[7]. Figure 9 shows patterns for the same data from the current revised model. Instead of considering only one immediate neighboring syllable of each annotated B3, i.e., only one pre- and post-B3 syllable, in the new model we define the immediate between-PPh neighborhood as the last 4 syllables of a preceding PPh and the first 3 syllables of the following PPh. With this definition, the PPh neighborhood is defined by units that encompass the boundary immediate PW rather than single syllable, a definition that better reflects the rationale of the HPG framework. Note that the cross-boundary contrast is more distinct in the revised model than in the previous model.

In addition, Figs. 7 and 9 combined also show how patterns derived from the revised model show more contrast in general than patterns derived from the previous model (Figs. 6 and 8).

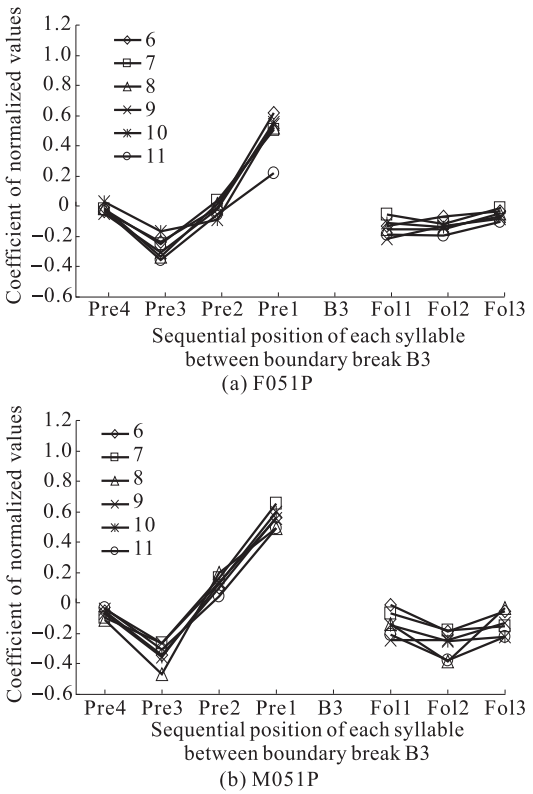

Fig. 8 PPh patterns of the previous duration model for speakers F051P and M051P. 0 on the vertical axis is defined as the mean of syllable duration.

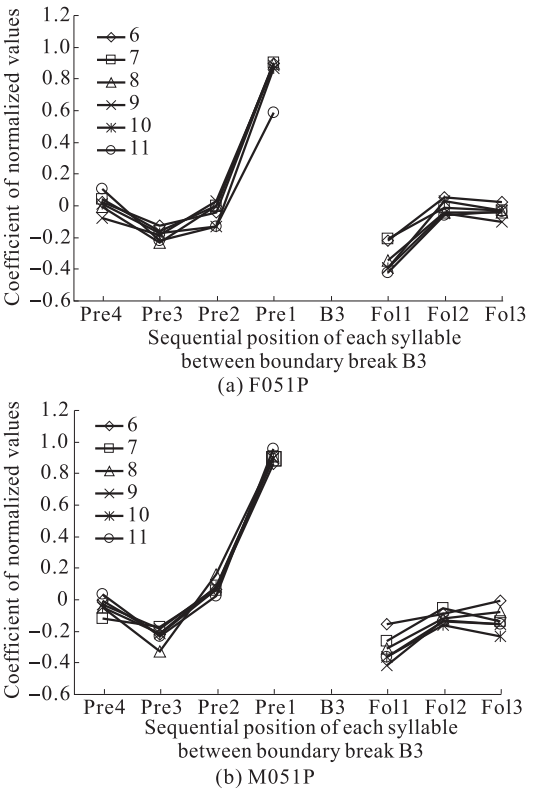

Fig. 9 PPh patterns of the revised duration model for speakers F051P and M051P. 0 on the vertical axis is defined as the mean of syllable duration.

2.2 Comparison of intensity predictions

Figure 10 shows the intensity patterns of PW (1-4 syllables in length) along the temporal course by syllable number and by speaker from the previous model, and Fig. 11 shows patterns from the current revised model. Similar to the results from the revised duration model, the revised intensity prediction patterns at the PW layer are also opposite to the previous predictions. Figures 12 and 13 show the intensity distribution of PPh patterns from both the previous and revised models. In each case the PPh's range from 6 to 11 syllables. Note that the PPh patterns from the revised model decay more sharply towards each boundary, thus matching the tendency of the intensity attenuation for PPh final weakening, especially for speaker M051P. Once again the cross-boundary contrast is more pronounced in the intensity predictions. Coupled with the increased phrase-final syllable lengthening found in Section 2.1, the prediction is closer to physical speech data. Therefore, we believe that the cross-boundary contrasts in both duration and intensity patterns are significant cues to boundary perception regardless of the boundary pause duration.

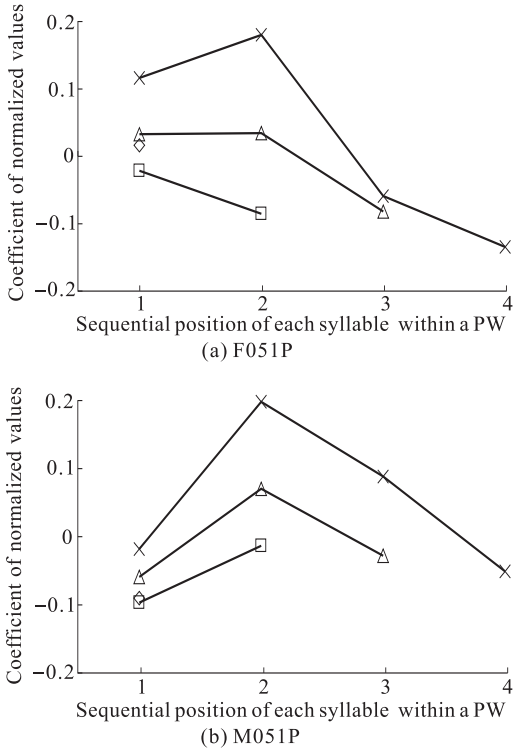

Fig. 10 PW patterns of the previous intensity model for speakers F051P and M051P. 0 on vertical axis is defined as the mean of syllable intensity. Each line represents the corresponding regression coefficient of one syllable at the specific position in a prosodic word.

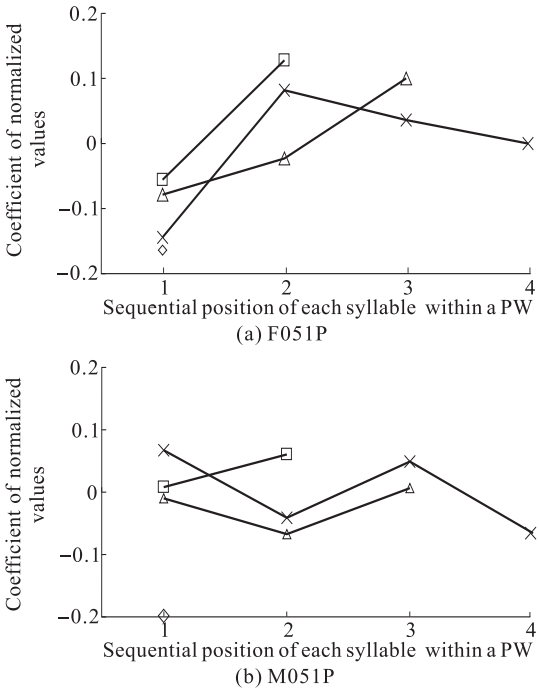

Fig. 11 PW patterns of the revised intensity model for speakers F051P and M051P. 0 on vertical axis is defined as the mean of syllable intensity. Each line represents the corresponding regression coefficient of one syllable at the specific position in a prosodic word.

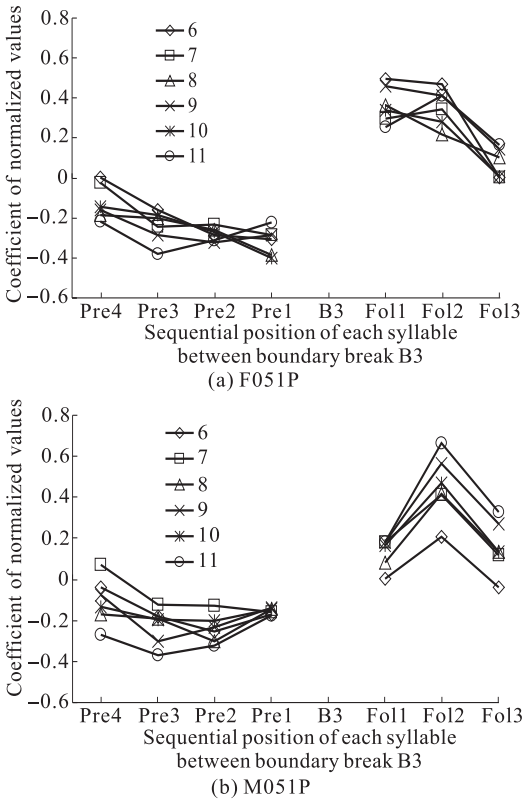

Fig. 12 PPh patterns of the previous intensity model for speakers F051P and M051P. 0 on vertical axis is defined as the mean of syllable intensity.

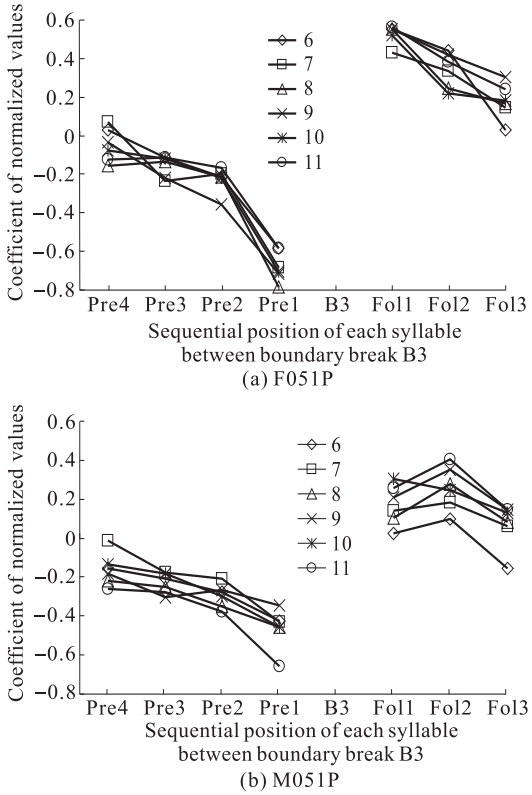

Fig. 13 PPh patterns of the revised intensity model for speakers F051P and M051P. 0 on vertical axis is defined as the mean of syllable intensity.

2.3 Comparison of pause predictions

Due to space limitations, we will present a comparison of pause prediction for one speaker only. Figure 14 shows a comparison of the predicted boundary pauses from the previous and revised models for speaker M051P. We can see that the differences of pauses between the previous and revised models for B1 and B2 are greater because the previous pause model could be mistaken for contributions from lower break levels. In the revised boundary pause model, we ignored the contribution from B1 to refine the prediction of lower breaks since the contribution of B1 is about 0.4 ms which cannot be perceived by the human ear.

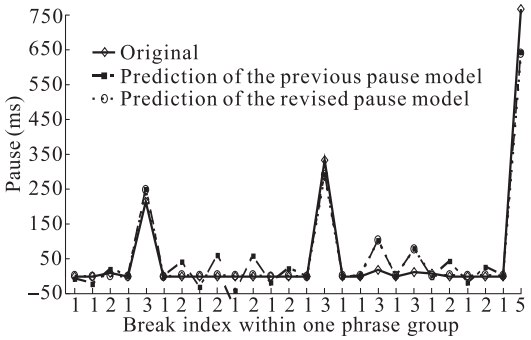

Fig. 14 An example of comparing the pause predictions between the previous and revised models for speaker M051P for one prosodic group

2.4 Prediction error improvement

Our analyses showed a reduction of overall total residual error (TRE) by about 20% from the previous model as compared to the revised model. Table 2 shows the TRE of the previous and revised models for both speakers. Therefore, revising the previous model by including more boundary neighborhood state resulted in improved predictions compared to the previous model, indicating that the current predictions deviate less from actual speech data.

Table 2 TRE values for speakers F051P and M051P

	F051P TRE (%)				M051P TRE (%)			
	Duration	Intensity	Pause	Average	Duration	Intensity	Pause	Average
Previous	36	54	32	41	33	48	27	36
Revised	32	47	22	34	31	41	13	28
TRE reduction	11	13	31	17	6	15	52	22

We also examined why the TRE value of the intensity prediction is always higher than that of the duration prediction. By comparing the distribution of Delta1 of the intensity model with that from the patterns shown in Figs. 2 and 3, one can see that the previous case has a broader distribution. This means that the variation of intensity is greater than that of duration, most notably for F051P. A broader distribution of Delta1 will result in greater deviation of the acoustic parameter. The pause prediction was increased effectively by ignoring the contribution of B1 and by adding punctuation mark as a feature. The order of prediction performance can therefore be stated as pause > duration > intensity.

2.5 Analysis of B3 pauses shorter than B2 pauses

As mentioned in Section 1, the range of pauses for breaks is very wide for B3, as plotted in Fig. 4. Therefore, in addition to revising the prediction models

above, we also studied B3 in more detail. We further analyzed the performance of duration and intensity predictions for cases where between-PPh pause B3 values were smaller than the B2 values. These are cases that contradict the annotation definition but are still consistently perceived by transcribers. Accordingly, we defined two conditions to analyze short B3 pauses: (1) where the lengths of the preceding and following PPhs are equal to or over 6 syllables, and (2) where the maximum pause of B2 is used as a B3 threshold.

The analysis of short B3 pauses for duration and intensity distributions is depicted in Figs. 15 and 16. In Figs. 15 and 16, we chose to include the last 4 syllables of the preceding PPh and the first 3 syllables of the following PPh at B3 for analysis. Except for the first 3 syllables of the following PPh of intensity for M051P, there are significant differences between the preceding and following syllables of B3. These results also indicate that our previous model attributed more contribution from the lower PW layer to output prosody, whereas the revised model entails instead more contribution from the higher PPh layer. Since the revised model yielded better overall predictions, it is clear that

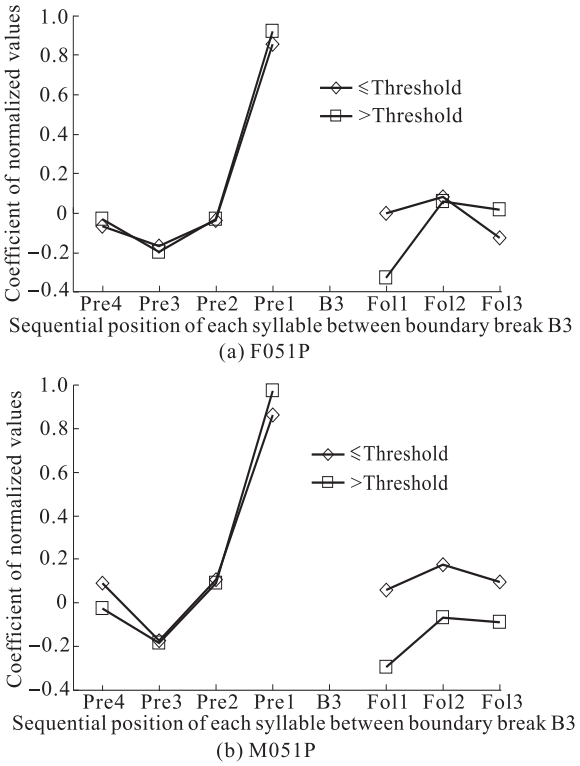

Fig. 15 PPh patterns of the duration model for F051P and M051P. PW's are from 1 to 4 syllables. 0 on vertical axis is defined as the mean of syllable duration.

increased contribution from higher level information accounts for the speech data better, hence proving further the significance of higher level contribution to the prosody output and showing how such information is perceived.

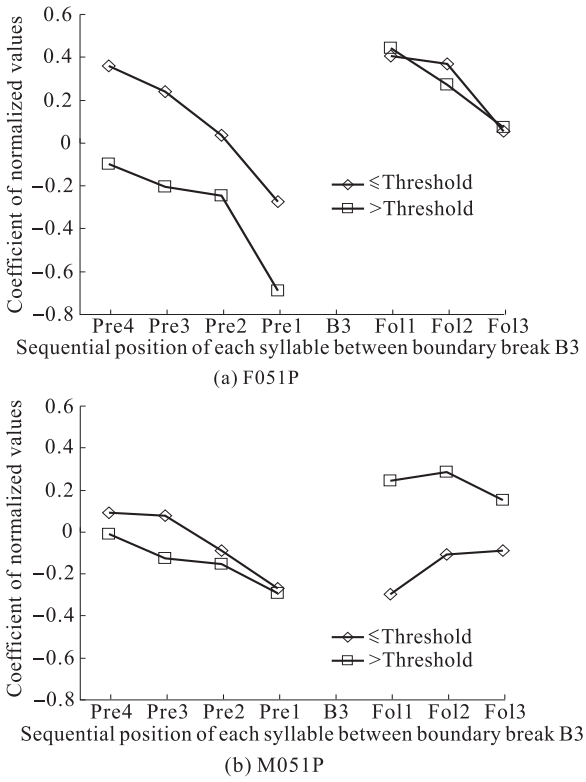

Fig. 16 PPh patterns of the intensity model for F051P and M051P. 0 on vertical axis is defined as the mean of syllable intensity.

3 Discussion

Instead of analyzing the duration and intensity patterns of one syllable before and after each annotated PW and PPh boundary break, as in a previous model[5,7], here we have analyzed the B3 boundary immediate prosodic states, in terms of the duration and intensity distribution along the time domain, using PW (4 syllables before and 3 syllables after). A comparison with the analysis from only immediate neighboring B2 state reveals different yet corresponding patterns in these two acoustic parameters. Accordingly, we have included factors of duration and intensity to revise and fine-tune the linear regression model[7], and recalculated the predicted contributions from the PW layer to the final prosody output under the HPG framework. The TRE of duration and intensity at the PW layer is

improved by 10%, the overall prediction of the output prosody is consequently improved by 5%. In addition, the layered predictions are now more consistent with the actual break distributions in the speech data.

Based on the above results, we believe that a detailed analysis of residual distributions of every prosodic layer (from syllable to PPh) can yield more stable and general patterns that lead to better prediction. In Figs. 6 and 10, the duration and intensity patterns at the PPh layer yield clearer evidence that the coefficients of the last 4 syllables are similar irrespective of PPh lengths (from 6 to 11 syllables). Thus, it is clear that to the human ear, the PW boundary break B2's and PPh break B3's can be distinguished from each other not by pause duration alone, but by contrastive neighborhood prosodic states as well. Evidence of boundary neighboring F0 contour patterns also showed similar results. Our analysis also shows how contrasts are constituted more by higher level constraints from discourse information than from lower level concatenation smoothing. To the human ear, it is clear that B2 and B3 boundaries are within- rather than between-paragraph signals as their respective pause durations are less relevant.

The results enable a better prediction of B3 and provide support to the idea that prosodic states relate more to higher level information such that prosody in fluent speech is more than just lower level co-articulation driven smoothing.

4 Conclusions

The presented results in this paper offer an alternative rationale for automatic segmentation of fluent speech and speech recognition of Mandarin Chinese in general. This is particularly true for the most commonly adopted approach which focuses on individual syllabic tone identities and F0 contour patterns, and perhaps inadvertently disregards boundaries as well as higher level information. The improved model can also be incorporated to enhance the prosody output of speech synthesis, showing where boundary breaks can be varied greatly to yield more natural prosody. Last, but not least, although the evidence was drawn from Mandarin Chinese, we believe boundary properties in relation to higher level discourse information are not language specific.

References

[1] Tseng Chiu-Yu, Cheng Yun-Ching, Chang Chun-Hsiang. Sinica COSPRO and toolkit—Corpora and platform of Mandarin Chinese fluent speech. In: Proceedings of Oriental COCOSDA 2005. Jakarata, Indonesia, 2005: 23-28. http://www.myet.com/cospro.

[2] Tseng Chiu-Yu, Lee Yeh-Lin. Speech rate and prosody units: Evidence of interaction from Mandarin Chinese. In: Proceedings of the International Conference on Speech Prosody. Nara, Japan, 2004: 251-254.

[3] Tseng Chiu-Yu, Pin Shao-Huang, Lee Yeh-Lin, Wang Hsin-Min, Chen Yong-Cheng. Fluent speech prosody: Framework and modeling. *Speech Communication*, 2005, **46**(3-4): 284-309.

[4] Tseng Chiu-Yu. Recognizing Mandarin Chinese fluent speech using prosody information—An initial investigation. In: The 3rd International Conference on Speech Prosody. Dresden, Germany, 2006.

[5] Tseng Chiu-Yu, Chou Fu-Chiang. A prosodic labeling system for Mandarin speech database. In: Proceedings of ICPhS'99. San Francisco, USA, 1999.

[6] Lieberman P. Intonation, Perception and Language. Cambridge, Massachusetts: The MIT Press, 1967.

[7] Tseng C, Fu B. Duration, intensity and pause predictions in relation to prosody organization. In: Proceedings of Interspeech. Lisbon, Portugal, 2005: 1405-1408.

[8] Sinica COSPRO and Toolkit. http://reg.myet.com/cospro.

[9] Keller E, Zellner K B. A timing model for fast french. York Papers in Linguistics, 17, University of York, 1996: 53-75.

[10] Zellner K B, Keller E. Representing Speech Rhythm Improvements in Speech Synthesis. Chichester: John Wiley, 2001.