Duration, Intensity and Pause Predictions in Relation to Prosody Organization

Chiu-yu Tseng and Bau-Ling Fu

Phonetics Lab, Institute of Linguistics
Academia Sinica, Taipei, Taiwan
cytling@sinica.edu.tw

Abstract

Our research group has postulated a perceptually based multi-phrase prosody framework for speech paragraphs in fluent speech using corporal analyses. The framework features a prosody hierarchy that organizes phrases and sentences into prosodic groups (PG) in connected speech, and specifies cross-phrase prosodic relationships in the acoustic domains [1, 2]. A corresponding fluent speech prosody model with four independent acoustic modules was also constructed [3]. The model predicts cross-phrase F_0 contours, duration patterns, intensity distribution and pause insertions in accordance with prosody organization. Cumulative results from each and every prosody layer accounts for overall output prosody. We have since improved the model first by refining the duration and intensity modules through corpus analysis, and subsequently used the above improved results to facilitate better pause/break predictions. As a result, the enhanced model is now more robust than its initial version. Future works will focus on applying the improved model to synthesis of fluent connected speech.

1. Introduction

We analyzed speech corpora of read Mandarin Chinese discourses from a top-down perspective on perceived units and boundaries, and consistently identified speech paragraphs of multiple phrases that reflected discourse planning rather than sentence effects in fluent speech. Subsequent cross-speaker and cross-speaking-rate acoustic analyses of identified speech paragraphs revealed systematic cross-phrase prosodic patterns in every acoustic parameter, namely, F_0 contours, duration adjustment, intensity patterns, and in addition, boundary breaks. We therefore argue for a higher prosodic node Prosodic Phrase Group (PG) that governs, constrains, and organizes multiple phrases to derive speech paragraphs. A hierarchical multi-phrase framework is constructed to account for the governing effect, with complimentary production and perceptual evidences. We also showed how each prosody layer contributes to overall prosody and how cross phrase-F_0- and syllable-duration templates could be derived. These templates account for the tune and rhythm characteristic to fluent speech prosody, as well as the look-ahead and forecast in fluent speech planning and processing. Therefore, we argue for a prosody framework that specifies phrasal intonations not as unrelated prosody units but rather, as subjacent sister constituent subject to higher constraints. Output fluent speech prosody is thus cumulative results of contributions from every prosodic layer, and respective contributions from each layer accounted for. From bottom up, the layered nodes are syllables (SYL), prosodic words (PW), prosodic phrases (PPh) or utterances, breath-group (BG) and prosodic phrase groups (PG). These constituents are, respectively, associated with break indices B1 to B5. A corresponding liner modular model of F_0, syllable duration, intensity distribution and pause/break prediction was also constructed. The model was capable of predicting fluent speech prosody satisfactorily [4, 5, 6]. A most comprehensive and recent account of the framework and model is in [3].

However, we noted that we had used different normalization procedures for each module during the course of developing the model, resulting variation between speakers on the one hand and causing difficulties to perform between-module comparisons on the other hand. So we sought after ways to improve the situation and test if better prediction could also be achieved. In the following sections, we discuss how we first refined the syllable duration module and intensity distribution module through improved normalization analyses, and how we used the results obtained to further enhance pause/break predictions.

2. Methods of Analysis

2.1. Speech Corpora Annotation

The speech data consisted of read Mandarin Chinese speech from 1 female (F051P) and 1 male (M051P) radio announcers. The two speakers read identical text of 26 long paragraphs ranging from 85 to 981 syllables. A total of 11591 syllables of F051P and 11596 syllables of M051P were analyzed. Segmental identities were first automatically labeled using the HTK toolkit and SAMPA-T notation [7], then hand labeled for perceived prosodic boundaries. All labeling was spot-checked by trained transcribers. Segmental intensity was first derived using an ESPS toolkit. For each segment, the averaged intensity was calculated using 10 equally spaced frames in the target segment time span. Segment duration less than 10 frames were directly averaged. Table-1 summarizes derived speech features of the two speakers.

	$\mu_{Duration}$	$\sigma_{Duration}$	$\mu_{Intensity}$	$\sigma_{Intensity}$	μ_{Pause}	σ_{Pause}
F051P	200	65	1298	680	37	106
M051P	190	60	897	350	45	138

Table-1 Speech features in F051P and M051P

2.2. Speech Data Normalization

To eliminate the variation between the speakers, each set of data was normalized with the mean and standard deviation of the entire class, instead of maximum and minimum used previously. The original method of normalization would easily be affected by extreme data, causing the distribution of normalized data to shift, and thereby making comparisons between speakers meaningless. To rectify the situation, we modified the normalization as follows:

$$Y_{nor(i)} = (Y_{(i)} - \mu_Y) / \sigma_Y$$

$$Y_{nor} = \{ Y_{nor(1)}, Y_{nor(2)}, \dots Y_{nor(n)} \}$$

$Y_{(i)}$ and $Y_{nor(i)}$ represent each datum in Class Y and Normalized Class Y respectively. μ_Y and σ_Y represent the mean and standard deviation in Class Y. The same modification was made for the three modules under consideration hence Y would be duration, intensity and pause in the following sections.

2.3. Duration Module

A layered, hierarchical regression model corresponding to our prosody framework was built from bottom up, namely, the SY layer, the PW layer, the PPh layer, and the BG layer where the PG layer is collapsed for the present study. The procedures are aimed to investigate relationships between dependent and independent variables.

Using a step-wise regression technique, a linear model with four layers [8, 9] was modified and developed to predict speakers' timing behavior. In the syllable layer, we used six consonant groups and six vowel groups. In order to reduce the difference between groups, the groupings were decided according to the concept of weight, instead of manual grouping. In other words, the number and mean of each segment was considered in relation to grouping. The Syllable Layer Model could be written as:

Y_{nor} *= Const + CCt + CVt + Ton*
Y_{nor} *= + PCt + PVt + PTt + FCt + FVt + FTt*
Y_{nor} *= + 2-way factors of each factor above*
Y_{nor} *= + 3-way factors of each syllable*
Y_{nor} *= + Delta 1*

Ct, Vt and Tt represent consonant type, vowel type and tone respectively. Prefix P, C and F represent preceding, current and following syllable. After regression, the less influential factors, prob. > 0.1, would be excluded. Residuals, Delta1, which could not be predicted by the syllable layer, would be analyzed in the immediate higher layer subsequently. The derived coefficients represent the effect unit on the specific syllable position of one prosodic unit.

In the PW layer, the PW Layer Model could be written as:

Delta 1 = f(PW length, PW sequence) + Delta 2

Residuals, Delta 2, which could not be predicted by the PW layer, would be analyzed in the immediate higher layer subsequently.

In the PPh layer, the PPh Layer Model could be written as:

Delta 2 = f(PPh length, PPh sequence) + Delta 3

In order to apply the concept of temporal allocation robustly to different corpora with different PPh length distribution, an adaptive threshold, which means the percentage of PPh length distribution over this threshold would decline to 5% minus, is necessary, instead of fixed threshold, namely 8 syllables as used before. Figure 1 shows the PPh length distribution of F051P and M051P; and the adaptive threshold would be 10. Therefore, we labeled the syllables in a PPh less than 10 syllables as [PPh length, PPh sequence]. For PPh with 11 syllables and above, we labeled the first (initial and hence I) and the last (final and hence F) 5 syllables individually, while the syllables in between were labeled as [M] for the medial positions, for example, {[I1], [I2], [I3], [I4], [I5], [M]... [M], [F1], [F2], [F3], [F4], [F5]}. By using such an adaptive threshold in different corpora with different PPh length distribution, we could avoid losing representative data or getting unrepresentative patterns. The Residual Delta 3 which could not be predicted by the PPh layer would be analyzed in the immediate higher layer subsequently.

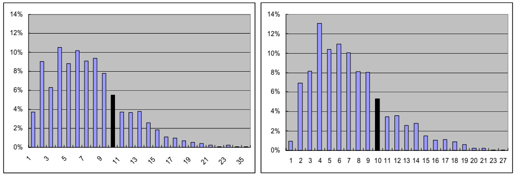

Figure 1 *Length distribution in F051P and M051P in syllable numbers*

In the BG layer, the BG Layer Model could be written as:

Delta 3 = f(PPh IMF, PPh length, PPh sequence)
Delta 3 = + Delta 4

We labeled the first PPh and the final PPh within one BG unit as "Initial" and "Final" PPh, while all other PPh were deemed the same and labeled "Medial" PPh. Within each PPh, the same rationale was used. That is, instead of using fixed threshold of 8 syllables as before, the first and last 5 syllables were labeled individually whereas the medial ones were not assigned individual identities. In other words, the structure of BG Layer Model was completely based on the PPh Layer Model so the BG Layer Model was also adaptive with the PPh Layer Model instead of using fixed threshold of 7 syllables as before. According to Figure 1, the initial PPh within one BG unit would be labeled as {[II1], [II2], [II3], [II4], [II5], [IM]... [IM], [IF1], [IF2], [IF3], [IF4], [IF5]}.

Two evaluations on the prediction outcome are used: 1. Correlation Coefficient, r, which represents how much the prediction outcome correlates with the original data. 2. The Total Residual Error, T.R.E., is the percentage of sum-squared residue over the sum-squared original data. T.R.E. indicates the residual error ratio that could not be accounted for from the bottom syllable layer is moved to the immediate layer.

2.4. Intensity Module

In the Intensity Module we used the same modified method of analysis as with the Duration Module by changing the dependent variables from duration to intensity. The process of square root is prohibited because there would be minuses in the new way of normalization.

2.5. Pause Module

The same modification was further used in the Pause Module by changing the dependent variables from duration to pause and perform the same kind of analyses.

3. Results and Discussion

3.1. Duration Module

Figure 2 and Figure 3 show respective duration patterns of PW and PPh for both the two speakers F051P and M051P. Each line represents the corresponding regression coefficient of one syllable at the specific position in a prosodic word and prosodic phrase. Y-axis represents the prediction of normalized values. Positive coefficients indicate that the syllable at this specific position possesses longer duration than the average value over the mean residue, while the negative ones shorter duration. The general pattern of PW layer is clear. Because the adaptive threshold at the PPh Layer in speakers F051P and M051P is 10, PPh over 10 syllables are shown in dark green, where the medial part of PPh is represented by the

6th syllable, while the first and the last 5 syllables are clearly shown.

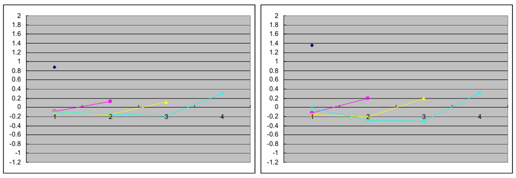

Figure 2 *Coefficients at the PW Layer in F051P and M051P*

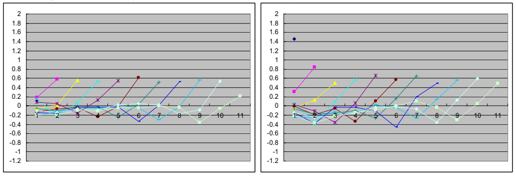

Figure 3 *Coefficients at the PPh Layer in F051P and M051P*

Figure 4 to Figure 6 show the duration patterns of BG unit in speakers F051P and M051P. Because the structure of BG Layer Model is completely based on the PPh Layer Model, the length (in syllable numbers) of the longest duration pattern at the BG Layer is equal to that at the PPh Layer.

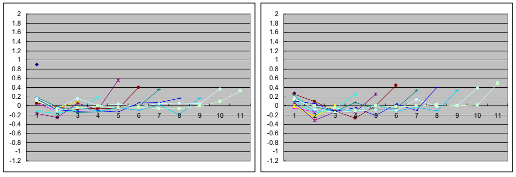

Figure 4 *Coefficients of Initial PPh at the BG Layer in F051P and M051P*

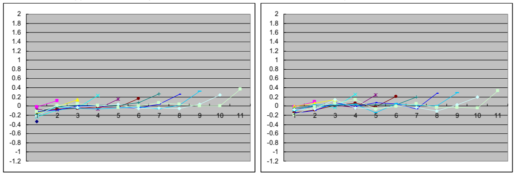

Figure 5 *Coefficients of Middle PPh at the BG Layer in F051P and M051P*

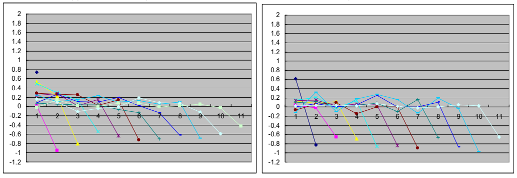

Figure 6 *Coefficients of Final PPh at the BG Layer in F051P and M051P*

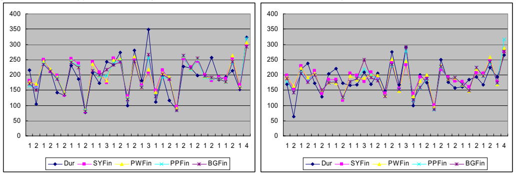

Figure 7 *Duration Comparison in F051P and M051P*

Although the length of duration patterns at the PPh and BG Layer is longer than before, the duration patterns are clearly shown. The results indicate that our previous analyses using the fixed threshold have resulted in loss of meaningful and representative patterns from the data, and may be detrimental to the prediction. Therefore, setting up a proper threshold, in this case the adaptive threshold instead of the fixed threshold, has facilitated in getting the most representative patterns from data with minimum distortion.

The contribution of each prosody layer is quite clear from Figure 2 to Figure 6 using the same scale. Figure 7 shows the comparison between the original speech data and predictions of each prosody layer in one BG unit. The curve "Dur" represents actual duration from the speech data, "SYFin" the prediction from syllable layer, "PWFin" the prediction form syllable to PW layer, "PPFin" the prediction form syllable to PPh layer and finally, "BGFin" the prediction form syllable to BG layer. Each number in X-axis represents one syllable labeled with break. The comparisons demonstrate the contribution of each prosody layer practically.

3.2. Intensity Module

The prediction patterns of each prosody layer in the intensity module are also very clear. Figure 8 and Figure 9 show results of refined intensity analyses at the PW and PPh layers, respectively; while Figure 10 to Figure 12 show results at the PPh layer in three BG positions, i.e., BG-Initial, BG-Medial and BG-Final. Overall cumulative predictions are shown in Figure 13. The error rate of intensity prediction is higher than that of duration prediction. Compared with duration predictions in relation to the original speech data as shown in Figure 7, it is expected that the predicted curves are farther from the original curve as Figure13 so shown.

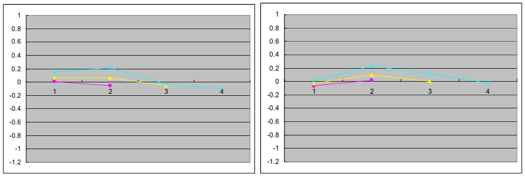

Figure 8 *Coefficients at the PW Layer in F051P and M051P*

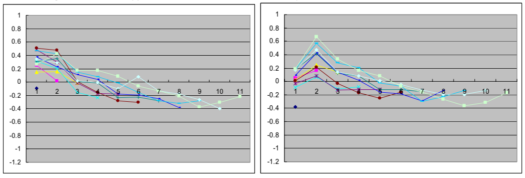

Figure 9 *Coefficients at the PPh Layer in F051P and M051P*

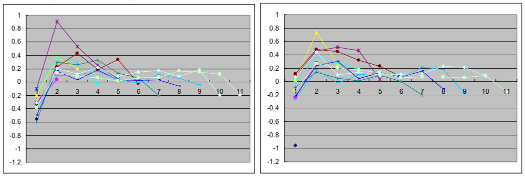

Figure 10 *Coefficients of Initial PPh at the BG Layer in F051P and M051P*

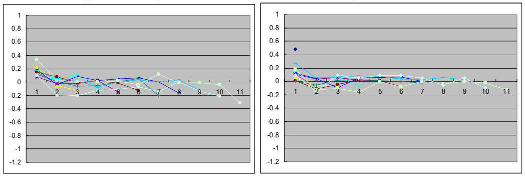

Figure 11 *Coefficients of Medial PPh at the BG Layer in F051P and M051P*

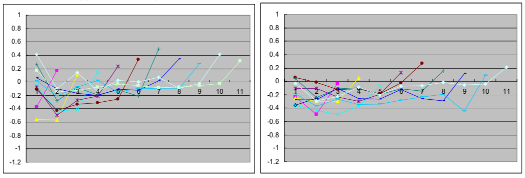

Figure 12 *Coefficients of Final PPh at the BG Layer in F051P and M051P*

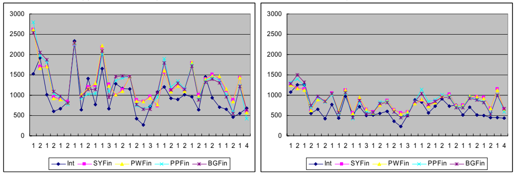

Figure 13 *Intensity Comparison in F051P and M051P*

3.3. Pause/Break Module

The prediction patterns of each prosody layer in the pause/break module are also clear. Figure 14 to Figure 18 show the modified and refined analyses at each prosody layer. The error rate of pause prediction is lower than that of duration prediction. Compared with duration predictions as shown in Figure 7, it is expected that the predicted curves are closer to the original curve as shown in Figure 19.

3.4. Prediction Evaluation

Evaluations on predictions of each prosody layer in duration, intensity and pause are depicted in Table-2. The lower T.R.E. means the higher performance. Therefore, the order of prediction performance is: pause > duration > intensity

4. Conclusions

We have shown in the present study that under our prosody framework, we were able to further enhance our prosody model by analyzing speech corpora in a more refined manner. The improvement was targeted to capture one of the major features of fluent speech prosody, namely, the organization of phrase groups corresponding to speech paragraphs, most notably signaled by how they begins and end in speech flow. By using the adaptive threshold and modified normalization, our model now is more robust than before. In particular, the better prediction achieved in pauses/breaks makes it possible to develop software towards locating and labeling prosody breaks not independently but in relation to prosody organization. These improvements should be more than constructive to speech synthesis for better prosody output.

5. References

[1] Tseng, C., 2003. Towards the organization of Mandarin speech prosody: units, boundaries and their characteristics. In Proceedings of ICPhS2003.

[2] Tseng, C., Pin, S., Lee, Y., 2004a. Speech prosody: issues, approaches and implications. in Fant, G., H. Fujisaki, J. Cao and Y. Xu Eds. From Traditional Phonology to Mandarin Speech Processing, Foreign Language Teaching and Research Process, pp. 417-438.

[3] Tseng, C., Pin, S., Lee, Y., Wang, H. and Chen, Y. (2005). Fluent Speech Prosody: Framework and Modeling. To appear in Speech Communication: Special Issue on Speech Prosody

[4] Tseng, C., and Lee, Y., 2004. Speech rate and prosody units: evidence of interaction from Mandarin Chinese. In: Proceedings of Speech Prosody 2004, pp. 251-254.

[5] Tseng, C., and Lee, Y., 2004. Intensity in relation to prosody organization. In Proceedings of the 4th International Symposium on Chinese Spoken Language Processing, 2004.

[6] Tseng, C., Pin, S., 2004d. Modeling prosody of Mandarin Chinese fluent speech via phrase grouping. In: Proceedings of ICSLT-O-COCOSDA 2004.

[7] Tseng, C., Chou, F., 1999. A prosodic labeling system for Mandarin speech database. In: Proceedings of ICPhS'99, pp. 2379-238.

[8] Keller, E., Zellner Keller, B. "A Timing model for Fast French", *York Papers in Linguistics*, 17, University of York. 53-75. (1996)

[9] Zellner Keller B, Keller E., "Representing Speech Rhythm" Improvements in Speech Synthesis. (pp. 154-164). Chichester: John Wiley. (2001)

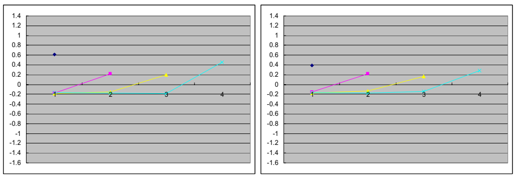

Figure 14 *Coefficients at the PW Layer in F051P and M051P*

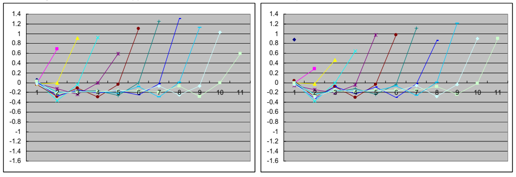

Figure 15 *Coefficients at the PPh Layer in F051P and M051P*

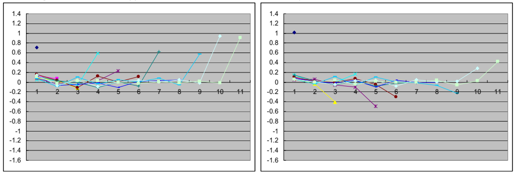

Figure 16 *Coefficients of Initial PPh at the BG Layer in F051P and M051P*

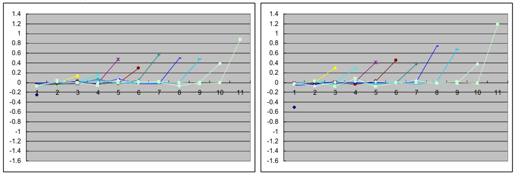

Figure 17 *Coefficients of Medial PPh at the BG Layer in F051P and M051P*

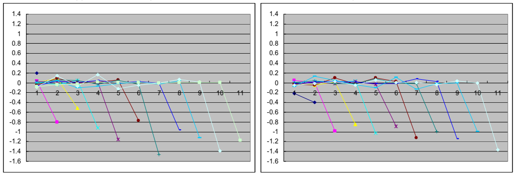

Figure 18 *Coefficients of Final PPh at the BG Layer in F051P and M051P*

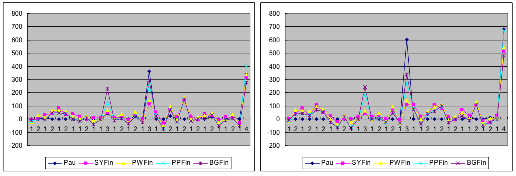

Figure 19 Pause Comparison in F051P and M051P

F051P		SY	PW	PP	BG
Duration	T.R.E.	46%	44%	39%	36%
	r	0.734	0.748	0.782	0.799
Intensity	T.R.E.	63%	62%	56%	54%
	r	0.611	0.613	0.662	0.682
Pause	T.R.E.	58%	54%	40%	32%
	r	0.649	0.681	0.799	0.827

M051P		SY	PW	PP	BG
Duration	T.R.E.	48%	44%	36%	33%
	r	0.718	0.747	0.805	0.822
Intensity	T.R.E.	56%	55%	51%	48%
	r	0.666	0.669	0.701	0.718
Pause	T.R.E.	50%	47%	34%	27%
	r	0.707	0.731	0.835	0.858

Table-2 Prediction Evaluations in F051P and M051P