International Symposium on Chinese Spoken
Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

INTENSITY IN RELATION TO PROSODY ORGANIZATION

Chiu-yu Tseng and Yelling Lee

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

ABSTRACT

Mandarin fluent speech prosody is most significantly characterized by phrase grouping. A hierarchical prosody framework of phrase grouping was proposed, where corresponding evidences of governing effects from the prosody organization were found in two acoustic correlates already, namely, overall F_0 contours and temporal allocations [1]. In this paper, we present results of investigating through corpus analyses the third acoustic correlates, i.e., intensity, to look for corresponding evidence in relation to prosody organization as well. The specific questions raised are (1) how intensity pattern could be explained by the governing effect of prosody organization, (2) whether the governing effect could be used in predicting intensity distribution. We argue that the acoustic roles of speech rhythm and intensity are as much integrated part of speech prosody as F_0 contour patterns. Therefore, we conclude that in order to construct a working prosody model, all three acoustic correlates should be considered in relation to prosody organization. The conclusion is also directly applicable to TTS to improve output naturalness. [2]

1. INTRODUCTION

In investigating and modeling speech prosody, much of the collective effort in the speech community has devoted to simple sentences or short phrases as units of study. The acoustic correlate receiving most attention has been F_0 contour patterns. However, we have shown that phrase grouping is essential to characterize the prosody for Mandarin fluent speech and that a larger prosody unit is necessary [1]. The constitution of phrases under grouping represents a possible cognitive unit of speech planning, while interaction with the articulatory apparatus and breathing necessary. This rationale implies that in a framework of speech prosody, cognitive, articulatory and physiological constraints are to be integrated. Our hierarchical framework was based on the unit located inside the perceived different levels of boundary breaks across speech flow, and their perceptual effect to human listeners. Using corpora of read speech, the boundaries are marked with a Tobi-based self-designed labeling system [3] that tagged small to large boundaries with a set of break indices (BI); i.e., B1 to B5. From top down, the layered nodes are phrase groups (PG), breath groups (BG), prosodic phrases (PPh), prosodic words (PW), and syllables (SYL). These constituents are, respectively, associated with break indices B5 to B1.

Evidence of prosodic phrase grouping was found both in adjustments of F_0 contours and temporal allocations within and across phrases [1, 4]. Under this framework, phrases are sister nodes instead of unrelated intonation units, and phrasal intonations no longer independent prosody units. Temporal allocations of syllable durations are also constrained by phase grouping. Phrasal intonations are governed by their respective positions within a PG where PG-initial, PG-medial and PG-final PPhs undergo modification to signal the onset, non-terminal effect and offset of a PG. Corresponding contribution of duration patterns of each prosodic layer are also found in relation to prosody organization across speakers and speaking rate [1, 4], demonstrating the governing effect of PG on speech rhythm as well. We note in particular that speech rhythm is a necessary feature in prosody; its specification also goes beyond the syllable level and its final output an integrated outcome by prosody organization.

For the present study, we used speech corpora of two Taiwan Mandarin radio announcers (1 male TMS and 1 female TFS), reading 15 large paragraphs from 81 to 981 syllables at the average speaking rate of 182 and 202 ms/syllable, respectively. Intensity analyses in relation to duration patterns and speech prosody organization were conducted to show how each prosodic layer contributed to intensity distribution within and across prosodic units. Following our approach [1, 2, 4] modeled after Zellner Keller and Keller [5] for Mandarin speech prosody, we adopted a step-wise regression technique, in which a linear model with four layers was developed to predict patterns of intensity distribution in fluent connected Mandarin Speech.

2. METHODS OF ANALYSIS

2.1. Speech corpora annotation and RMS value calculation

Segmental identities were first automatically labeled using the HTK toolkit and SAMPA-T notation, then hand labeled for perceived prosodic boundaries. All labeling was spot-checked by trained transcribers. Segmental RMS values were first derived using an ESPS toolkit. For each segment, the averaged RMS value was calculated using 10 equally spaced frames in the target segment time span. Segment duration less than 10 frames are directly averaged. In addition, to eliminate the level difference between paragraphs possibly caused by slight changes during recording, the RMS values within each paragraph were normalized, hence NRMS. The rationale was to focus on investigating the pattern of intensity ratio difference within prosodic units rather than deriving absolute intensity of each syllable.

2.2. Analysis procedures

A layered, hierarchical regression model corresponding to our prosody framework was built from bottom up, namely, the SYL layer, the PW layer, the PPh layer, and the BG layer where the

PG layer is collapsed for the present study. The procedures were aimed to investigate relationships between dependent variable(DV)s (syllable durations, NRMS) and independent variable(IV)s (segment identities, prosodic organization), and could be summarized as follows: (1.) linearize DVs, (2.) decide segmental identities groupings, (3.) build up linear regression model, (4.) prune less affecting IVs, (5.) explain residuals that can not be predicted by the immediate higher layers, and finally repeat steps (3.) and (5.). The original duration value and the square root of NRMS value for the 2 corpora were used in the following regression procedures. Linear models for discrete data were built using DataDesk with partial sum of squares (type3).

In the syllable layer, to reduce the regression complexity we used 6 consonant groups and 6 vowel groups. The groupings were decided using the mean DV value of each IV identity. In other words, segment identities with similar mean NRMS values were grouped together. Each syllable's identity in consonant type, vowel type, and tone types were thus used as IVs in the syllable layer. In addition, the preceding and following syllable of each syllable were considered as factors. Furthermore, we also took into account their 2-way interactions. Thus the syllable layer regression could be formulated as:

Square root NRMS = constant + CTy +VTy + Ton
+ PCT+ PVT + PrT + FCt + FVt + FlT
+ 2-way factors of each factor above
+ Delta 1

After regression, the less influential factors (p-value less than 0.1) were excluded. Residuals (Delta 1) that could not be predicted by segment identities were analyzed in the immediate higher layers subsequently. From the PW layer to BG layer, we used the prosodic structure as IVs. The derived coefficients represent the effect (increase/decrease of intensity)) unit on the specific syllable position of the prosodic unit.

In the PW layer, our aim was to see whether DV is affected by its position within a PW. Therefore, the PW Layer Model can thus be written as:

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

Residuals of the PW layer, i.e., Delta 2, which could not be predicted by the PW structure, were analyzed in the immediate higher level the Phi layer. From our previous results on durations patterns and temporal allocations [1, 4], we found that the most significant effect of duration pattern was on the first and the last 4 syllables within a Phi. Therefore, we labeled the syllables in a Phi less than 8 syllables as [Phi length, Phi sequence]. For PHs with more 9 syllables above, we labeled the first 4 and the last 4 syllables individually, while the syllables in between were labeled as [M] for the medial positions, for example, {[I1], [I2], [I3], [I4], [M]…. [M], [F1], [F2], [F3], [F4]}. By doing so, we bypassed the scarcity problem of longer PHs in the corpora. The Phi layer could be formulated as:

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

Residual from the PPh layer, Delta 3, which could not be predicted by the PW and PPh layers, were then analyzed in the immediate higher BG layer. We labeled the first PPh and the final PPh within a BG as Initial- and Final-PPhs, while all other PPhs were deemed the same and labeled Medial PPhs. Within each PPh, PPh less than 7 syllables were labeled as [PPh length, PPh sequence]. For PPhs from 7 syllables up, the first and final 3 labels were labeled separately, while the others were deemed the same. As a result, the initial PPh within a BG can be labeled {[II1], [II2], [II3], [IM]…[IM], [IF1], [IF2], [IF3]}. And the BG layer could be formulated as:

Delta 3 = f(PPh I|M|F, PPh length, PPh sequence) + Delta 4

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

3. RESULTS OF ANALYSIS

3.1. Syllable Layer

In this layer, different segmental groupings were used in the regression of each corpus. The grouping was decided according to the mean of square root NRMS value of each segment's identity. The consonant groupings were much more similar among corpora than the vowel groupings. Regression using both 1-way and 2-way factors were conducted and factors with p-value larger than 0.1 were neglected.

3.2. PW Layer

Figure 1 shows the intensity pattern of PW unit in the corpora. Each line represents the corresponding regression coefficient of a syllable at the specific position in a prosodic word. Y-axis represents the prediction of square root of NRMS value. Positive coefficients indicate that the syllable at this specific position possesses more intensity than the average value over the mean residue, while the negative ones less intensity. The general pattern of PW the layer was clear. The longer the prosodic word is, the more intensity it needs to be produced.

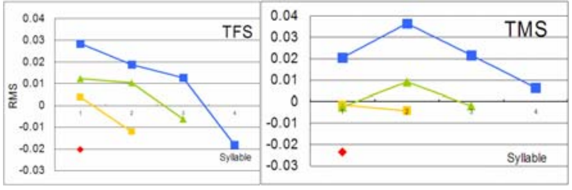

Figure 1. *Regression Coefficients of Intensity at the PW layer*

3.3. PPh Layer

Figure 2 demonstrates the intensity ratio pattern of PPh in the corpora. Each line represents the intensity pattern of a PPh of specific syllable number. PPhs over 9 syllables were shown in purple, where the medial part of PPh was represented by the 4th syllable, while the first and the last 4 syllables were clearly shown.

We found that more intensity was needed to produce longer PPhs as well. For PPhs over 9 syllables, the range of coefficients was 0.088 for TMS and 0.101 for TFS, where both speakers exhibited a change of prominence on the last 2 syllables. However, for PPhs of The coefficient range of PPhs over 7 syllables was 0.073 in TMS and 0.1 for TFS. The coefficient range of medial PPhs over 7 syllables was 0.056 for TMS and 0.081 for TFS.

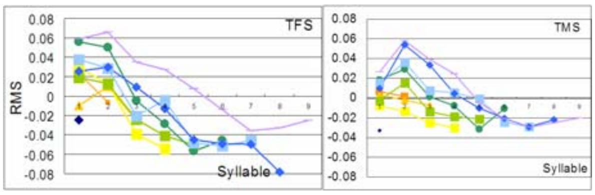

Figure 2. *Regression Coefficients of Intensity at the PPh layer,*

3.4. BG Layer

In the BG layer, our PG framework states that the contribution of different PPh in a BG is not the same. Or, the governing effect from the higher node is not the same on each PPh. The first and last PPhs need to signal the beginning and ending of a PG where the PPhs in between need to maintain a non-terminal effect. Analyses were performed accordingly. The first and the last PPh of a BG were considered as the initial and final PPh; while the others the medial PPh(s). Figure 3 shows the intensity pattern of Initial PPh of BG. Each line represents the pattern of PPh of different length. Initial PPh over 7 syllables were shown in purple.

Contrary to the pattern seen at the PPh layer as shown in Figure 2, the initial PPhs did not possess larger intensity as they became longer. Generally speaking, syllables in the initial PPh of a BG had larger intensity compared to the derived prediction value, except for the first and the last syllable. Figure 3 shows the results.

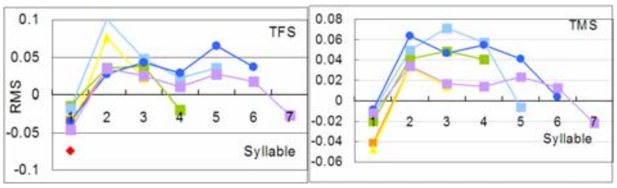

Figure 3. *Regression Coefficients of Intensity of Initial PPhs at the BG layer*

The Medial PPh of BG in Figure 4 shows the intensity pattern of the medial PPhs of a BG. Similar to the PPh layer, the medial PPhs had a similar declination pattern of intensity.

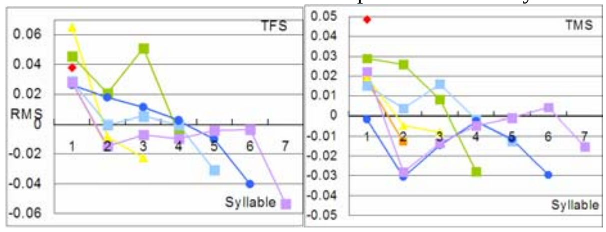

Figure 4. *Regression Coefficients of Intensity of Medial PHs at the BG layer*

The Final PPh of BG in Figure 5 shows the intensity pattern in the final PPh of a BG. Both corpora showed that instead of different degree of declination of intensity seen in the Initial and Medial PPhs, the final PPhs showed a converse tendency, ending with a rather stronger syllable. Shorter final PPhs had a wider coefficient range. For final PPh over 7 syllables, the range is 0.059 in TMS, and 0.077 in TFS.

Since the overall prediction shall take into account prediction of all the layers, while the BG final is still weakened. The results of intensity analyses at each layer are consistent with the results of F_0 and duration analyses [1, 4, and 7].

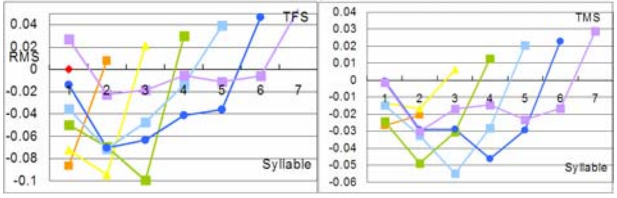

Figure 5. *Regression Coefficients of Intensity of Final PHs at the BG layer*

3.5. Evaluation

Evaluations on the overall prediction are depicted in Table 1. Note that the evaluation is made between the original DV (NRMS value) and the transformation of summed prediction (square of prediction sum); therefore the residual error may be increased due to this nonlinear transform. On the other hand, correlation r indicates that the prediction ability is improved when a higher layer is added. In the corpora, the result correlation r is 0.787 for TMS and 0.693 for TFS.

Table 1. *Overall Evaluations on Intensity Prediction of TMS Corpora*

Corpus / Layer	TFS		TMS	
	T.R.E.	r	T.R.E.	r
Syllable	63.80%	0.616	47.65%	0.732
PW	62.10%	0.621	47.02%	0.736
PP	48.19%	0.666	37.43%	0.766
BG	49.04%	0.693	35.37%	0.787

4. DISCUSSION

We have analyzed and derived intensity patterns across phrase groups of Taiwan Mandarin Chinese fluent connect speech using segmental identities and a layered hierarchical prosodic structure. Results from the above analyses are discussed in sections below.

4.1. Syllable Layer

Nearly 50% of the prediction could be derived from the segmental identity factors. For intensity prediction, nearly 52% of the original data (T.R.E = 47.65% in TMS) were predicted.

4.2. PW Layer

The gradual declination of intensity represents the prototype of intensity pattern at the prosodic word level. For intensity prediction, T.R.E reduced 0.63% in TMS and 1.7% in TFS. The contribution of the PW layer may not be pronounced at this level, and will be considered with other higher levels. However, note that a scoop-like duration pattern was also found at this layer [1, 4]. Both evidences indicate contributions from prosodic levels above the syllables. From the intrinsic pattern of speech rhythm found [1,4], intensity distribution is also governed by prosody organization, though at a smaller scale.

4.3. PPh Layer

Among the three prosodic layer (PW, PPh, BG layer), PPh layer accounted for the most part in both predictions. Nearly 10% of T.R.E. was explained by PPh layer in the corpora of TMS and 14% for TFS. This is a clear evidence that PPh information interacted with the lower prosodic levels and contributed significantly to the overall intensity distribution. The contribution of the PPh layer echoes duration contributions at the same layer towards the rhythmic structure [1,4], as well as the F_0 outcome [2, 7].

4.4. BG Layer

The grouping of phrase into a larger prosody unit is best exemplified at the BG layer from our duration studies [1,4], clear corresponding patterns in intensity distribution is also found in the present study. For the initial PPh, the first and last syllables are weaker while the medial syllables were pronounced with much more intensity. Note that if intensity distribution is viewed together with duration pattern for PG-initial PPhs, in that the final syllable is always lengthened, then a clear distinction between the first and last syllable in a PPh is obtained. For the medial PPh, the first syllable was stressed and shortened, while the last syllable is weakened and lengthened. As for the final PPh, similar to duration patterns found, opposite pattern to the derived PPh pattern were found in intensity pattern as well. The last two syllables showed a weakened-stressed pattern. Besides, the final syllable is also shortened comparing to other PPh units in the same BG.

In this hierarchical regression model, the prediction ability was increased whenever a higher layer of prosodic unit is added.

6. CONCLUSION

We have investigated the overall F_0 contour pattern as well as duration patterns in relation to prosody organization and found that a prosodic organization of phrase grouping contributed in layers to the final output of fluent speech prosody. Phrasal intonations are thus related and modifications are governed. In this light, the role of in dependent phrasal intonation is less significant [9]; lower levels of prosodic information such as PW and PPh only contribute partially to speech prosody. Corresponding duration patterns in relation to prosody organization were also found from corpus analyses, providing a prototype of speech rhythm in fluent speech [1,4]. In this paper, we investigated the third acoustic correlates, i.e., intensity, as well to look for corresponding evidence. We found that intensity pattern could be explained by prosody organization, and the governing effect from prosodic phrase grouping to lower prosodic levels was necessary to derive the final output. The derived F_0, rhythmic and intensity models could serve as basis of speech synthesis for connected fluent speech. From the evidences presented, we argue that F_0 contour patterns alone are insufficient to characterize the major part of speech prosody. Further, the acoustic roles of speech rhythm and intensity deem reconsideration. In fact, integrated adjustment in relation to prosody organization is called for in all three acoustic correlates. Initial attempt to simulate, predict and integrate our prosody model into a Mandarin TTS system was reported in separate papers [2,7]. We believe our studies have shown that more understanding of speech prosody in operation is helpful to better simulation of speech prosody in modeling.

7. REFERENCES

[1] Tseng, C, S. Pin and Y Lee, “Speech Prosody: Issues, Approaches and Implications”, in Font, G., H. Fujisaki, J. CIO and Y. Up Eds. *From Traditional Phonology to Mandarin Speech Processing,* Foreign Language Teaching and Research Process, Beijing, China, 2004, pp. 417-438.

[2] Pin, S., Y. Lee, Y. Chen, H. Wang and C. Tseng, “A Mandarin TTS System with an Integrated Prosody Model”, paper submitted to ISCSLP-2004

[3] Tseng, C. and F. Chou, “A Prosodic Labeling System for Mandarin Speech Database”, *Proc. Of With International Congress of Phonetic Science*, San Francisco, California. Pp. 2379-238

[4] Tseng, C. and Y. Lee, “Speech Rate and Prosody Units: Evidence of Interaction from Mandarin Chinese”, *Proceedings of Speech Prosody 2004*, Nara, Japan, March 23-26, 2004, pp. 251-254.

[5] Zellner Keller, B. & Keller, E. Representing Speech Rhythm. *Improvements in Speech Synthesis*, 154-164. Chichester: John Wiley. 2001

[6] Tseng, C. and S. Pin, “Mandarin Chinese Prosodic Phrase Grouping and Modeling--Method and Implications”, *Proceedings of International Symposium on Tonal Aspects of Languages—with Emphasis on Tonal Languages (TAL 2004)*, Beijing, China, March 28-30, 2004, pp. 193-19.

[7] Tseng, C. and S. Pin, “Modeling Prosody of Mandarin Chinese Fluent Speech via Phrase Grouping”, paper submitted to O-COCOSDA 2004

[8] Tseng, C., “Towards the Organization of Mandarin Speech Prosody: Units, Boundaries and Their Characteristics”, *Proceedings of ICPhS 2003*, Barcelona, Spain.

[9] Tseng, C., “On the Role of Intonation in the Organization of Mandarin Speech Prosody”, *Eurospeech 2003 / Interspeech 2003*, Geneva, Switzerland.