# Higher Level Organization and Discourse Prosody

*Chiu-yu Tseng*

Phonetics Lab, Institute of Linguistics,

Academia Sinica, Taipei

E-mail: * cytling@sinica.edu.tw

## Abstract

This paper addresses higher level organization in discourse prosody. Fluent speech prosody of text reading illustrated higher level speech planning above phrases and prosody segments above intonation units. Adopting a top-down perspective allowed clearer reflection of scope and unit involved. We examined large amount of speech data via a corpus approach, studied read discourse through perceived boundaries, analyzed prosodic characteristics of between-boundary units, and found evidence of higher prosodic specifications above phrase intonation. Through tailored quantitative analyses corresponding to a multi-layer prosodic hierarchy, we found how different prosodic levels contribute separately to prosody output, and how cumulative contributions added up to output prosody. The prosody hierarchy specifies that speech paragraphs are immediate constituents of discourse; phrases immediate constitute of speech paragraphs. Lower level nodes are subjacent units subject to higher level constraints; sister constituents bear association to one another. Hence central to discourse prosody is higher level specification as well as cross-phrase association in addition to discrete intonation patterns. Cross-phrase cadence templates could be derived to account for the melody, rhythm, loudness and boundary breaks of fluent speech. Further, evidence of cross-paragraph discourse association is also found. We believe in addition to advance understanding of discourse prosody; the knowledge is also directly applicable to speech technology development, especially speech synthesis.

Keywords: higher level organization, speech prosody, prosodic phrase grouping (PG), prosodic hierarchy, top-down, multi-phrase, cross-phrase association, templates, speech planning, global $F_0$ templates, temporal allocations, syllable duration patterns, intensity distribution, boundary breaks.

## 1. Introduction

There are three reasons specific to Chinese that made investigation of Mandarin Chinese speech prosody interesting. 1. Chinese is often misunderstood as a mono-syllabic language. This is due largely to syllable as a tone bearing unit and lack of morphological affixations, and hence much attention has been given to studies of lexical tones. 2. The syllable-based orthography that does not require word boundaries in writing and less rigid punctuation requirements in writing makes the notion of words and sentences flexible and fuzzy at the same time. 3. Also fuzzy is the demarcation between complex sentence and paragraph, making it all the more difficult to analyze discourse in text. One commonly adopted practice, as used by the CKIP group and its tree bank analysis [1] is to treat units between punctuations as independent sentences which often results in inadequate account of discourse information.

However, through reading of text, we found clear multi-phrase units in speaking that reflected higher level information. Adopting a corpus approach allowed us to examine large amount of varied speech samples, and at the same time required necessary deliberation over quantitative analyses used. A total of 9 sets of prosody-oriented speech corpus consisting of reading of text have been collected since 1997 [2]; 11.9GB of the corpora and toolkit developed were released in 2006 (*http://www.myet.com/COSPRO* ). A perception based annotation system was designed with the capacity to transcribe speech data by perceived boundaries, and in units above phrases/sentence [3]; intra- as well as inter-transcriber consistencies were maintained. Acoustic analyses included $F_0$ patterns, syllable duration, intensity distribution, and in addition, boundary breaks. We found systematic characteristics of higher organization manifested through the melody and tune, rhythm and tempo of narrative prosody that essentially associates phrases into paragraphs. As a result, we postulated a hierarchical multi-layer multi-phrase prosody framework that accounts for higher level discourse organization. Corresponding cross-phrase cadence templates were derived; a mathematical model was also constructed [4, 5]. In this paper, we will show from a simple hierarchical framework of Prosodic multiple-Phrase-grouping (PG) how cross-phrase prosodic relationship exists above individual intonation units to convey paragraph information; how intonation units are subordinate and subjacent prosody units to speech paragraph; and how speech paragraphs are also subordinate and subjacent discourse units/segments subject to even higher governing. Discussion will focus on duration patterns regarding speech rhythm and intensity distribution patterns.

## 2. Phrase grouping - organization and framework of speech paragraph

The concept of phrase grouping is not language specific to Chinese, since it is well accepted that utterances are phrased into constituents and are hierarchically organized into various domains at different levels of the prosodic organization [6, 7, 8]. We proposed [4] that by adding another layer over phrases, Prosodic Phrase Grouping (PG) could be viewed as a higher level organization that governs and constrains individual sentence whereby existing linguistic definition of intonation units still apply. By the same logic, Discourse is a higher level organization above PG. Therefore, phrases are immediate subjacent sister constituents under PG; PG immediate constituents of Discourse. Figure 1 is a schematic illustration of the framework.

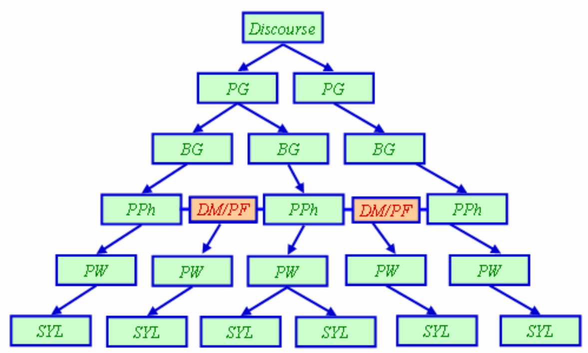


Figure 1: A schematic representation of how PGs form spoken discourse and where DM (Discourse Marker) and PF (Prosodic Filler) are located.

The framework is based on the perceived units located inside different levels of boundary breaks across speech flow. Units used were perceived prosodic entities. The boundaries are annotated using a ToBI-based self-designed labeling system [3] that annotated small to large boundaries with a set of 5 break indices (BI); i.e., B1 to B5, purposely making no reference to either lexical or syntactic properties in order to be able to study possible gaps between these different linguistic levels and units. Phrase-grouping related evidences were found both in adjustments of perceived pitch contours, and boundary breaks within and across phrases, with subsequent analyses of temporal allocations and intensity distribution [9, 10, 11].

From bottom up, the layered nodes are syllables (SYL), prosodic words (PW), prosodic phrases (PPh) or utterances, breath group (BG), prosodic phrase groups (PG) and Discourse. Optional discourse markers (DM) and prosodic fillers (PF) exist between phrases, but are linkers and transitions within and across PGs. These constituents are, respectively, associated with break indices B1 to B5. These boundary breaks are not shown in Figure 1 to keep the illustration less complicated. B1 denotes syllable boundary at the SYL layer where usually no perceived pauses exist; B2 a perceived minor break at the PW layer; B3 a perceived major break at the PPhs layer; B4 when the speaker is out of breath and takes a full breath and breaks at the BG layer; and B5 when a perceived trailing-to-a-final-end occurs and the longest break follows. In the framework, intonation unit is usually a PPh. When a speech paragraph is relatively shorter and does not exceed the speaker's breathing cycle, the top two layers BG and PG collapse into the PG layer. BGs and/or PGs are therefore units of discourse prosody.

We reported that the most significant features of PGs are where and how they begin and end. In other words, the most significant contrast regarding PG-related positions and boundaries exists between PG-initial and PG-final positions [9]. In addition, PG specifications also include how subjacent individual intonation unit adjusts, and how cumulatively layered contributions add up output prosody [4, 5]. The multi-layer framework presented assumes an independent higher level scope that reflects discourse planning and on-line processing. Hence it is feasible to assume corresponding canonical and default global templates that contribute to the planning within and across units before and during speech production, much the same as cadence templates in music. They also imply cross-phrase anticipation goes beyond physiologically conditioned articulatory maneuvers at the segmental, tonal, and intonation levels. The interacting and trading relationships between and tones and sentence intonation were described as "…small ripples riding on large waves"[12] and has been well-known to the Chinese linguistic community. By analogy our framework assumes that larger and higher layer(s) may be superimposed over intonation and tones as tides over both waves and ripples. The question then is what the tides are like, how ripples ride over waves, and how waves ride over tides.

## 2. 1. Speech Melody--Global $F_0$ patterns of PG

A canonical overall $F_0$ contour cadence template for PG representing the global melody of PG was proposed to specify how a PG begins, holds and ends [4]. Unit of the template is PPh which can be either a phrase or a short sentence. Figure 2 is a trajectory of a 5-phrase PG, preceded and followed by B4 or B5; phrases within are separated by B3.

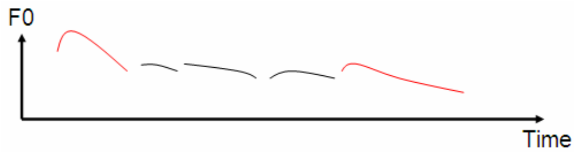


Figure 2. Schematic illustration of the global trajectory of perceived $F_0$ contours of a 5-PPh PG preceded and followed by B4 or B5. Within-PG units are PPhs and separated by boundary breaks B3s.

In our corpora a speech paragraphs corresponding to a PG range from 3 to 12 phrases; B5 usually occurs in PGs over 5 phrases. Note that only the PG-initial and -final PPhs (shown in red in Figure2) possess clear declarative intonations whereas the PG-medial phrases may exhibit flatter contour patterns. The first PPh features a $F_0$ reset before declining rapidly indicating new information while final PPh features a relatively lower $F_0$ reset and final lengthening. Together, the three positions and respective phrases form speech paragraph. With these three PG-specified positions, respective intonation patterns also signify global as well as cross-phrase relationship, as does the global default multi-phrase speech melody. This is essentially why top-down higher level information makes it possible to produce and perceive a speech paragraph; the PG-initial and -final intonations are never confused. Taken one intonation unit at a time, the cross-phrase higher level relationship disappears and intonation variations unaccountable. To the listener, the effect is very similar to hearing a piece of music where cadences are heard and templates used. In short, this is why discourse prosody is NOT concatenation or string of discrete phrasal intonations. Hence, a multi-phrase $F_0$ template, coupled with PG-specified boundary breaks, can be viewed as a global melody over phrases that are intonation units, much the same way as tide over waves and ripples, and even larger tides over tides. All of the units involved are but discourse units at different layers of the prosody hierarchy. Output discourse is cumulative integrated results from higher level organization.

### 2.2. Speech Rhythm--Duration patterns within and across phrases

In this section we discuss duration patterns from corpora analyses that form speech rhythm in fluent speech. Results of how each prosodic layer in the framework accounts for the duration pattern across syllables and contributes to the final tempo of phrases under grouping is presented. Syllable-cadence templates from each prosodic layer are derived to account for the rhythmic structure associated with higher level prosody organization. In the discussion below, duration and syllable duration are used interchangeably.

Three sets of Mandarin Chinese speech corpora were analyzed. Text reading from two Taiwanese Mandarin radio announcers (1 male TMS and 1 female TFS COSPRO 05) and another Beijing female radio announcer (BFS) were used. Text scrip contains 15 large paragraphs ranging from 85 to 981 syllables per paragraph; or 5655 syllables at an average of 377 syllables per paragraph. The three speakers also differed in speaking rate. Average speaking rates for speakers TFS, TMS and BFS are 202ms, 182ms and 267ms, or 4.3, 4.2 and 3 syllables/sec respectively. Table 1 summarizes the corpora.

| Feature \ Corpora | TFS | TMS | BFS |
|---|---|---|---|
| Total Syllable Number | 5655 | 5655 | 5483 |
| Syllable Mean Duration (ms) (Only segment duration considered) | 202.45 | 181.84 | 266.91 |
| Speech Rate (Syllable #/sec) (Pause considered) | 4.23 | 4.33 | 2.816 |

Table1. Syllable numbers and speaking rates of speech data.

The three sets of speech data were first labeled automatically for segments using the HTK toolkit and SAMPA-T notations [3], then hand labeled independently by 3 trained transcribers for perceived prosodic boundaries. The HTK labeling was manually spot-checked; the manual perceptual labeling cross-checked for inter-transcriber consistency. Using a step-wise regression technique, a linear model with four layers [13] was developed and modified for Mandarin Chinese to predict speakers' temporal allocation patterns. Moving from the SYL layer upward to each of the higher prosodic units and levels, we examined each higher layer independently to see if it could account for residuals from one of the lower layers, and if so, how much was contributed by each level. All of the data were analyzed using DataDesk™ from Data Description, INC. Two benchmark values were used in this study to evaluate the closeness of the predicted value to that of the original speech data, residual error (R.E.) and correlation coefficient (r). Residual error was defined as the percentage of the sum-squared residue (the difference between prediction and original value) over the sum-squared original value. Duration analyses presented were only limited to the BG layer for lack of sufficient samples of PGs.

#### *2.2.1. Results*

**Syllable Layer**

In this layer, different segmental grouping were used in the regression of each corpora. Table2 to 4 indicate the segment grouping of each corpus in the duration regression. The grouping is determined by the mean duration value of each segment's identity. Regression using both 1-way and 2-way factors were made; factors with p-value larger than 0.1 were neglected. Table5 to 7 show the ANOVA table of the remained factors. In the listed factors, where CTy/CT represents the consonant identity factor; VTy/VT represents the vowel identity factor and Ton/Tn represents the tone identity factor. The prefix P and F indicate whether the contributed factor is of the proceeding or following syllable.

| Group | Consonant | Group | Vowel |
|---|---|---|---|
| CDUR1 | d,b,g | VDUR1 | @,o,U`,U |
| CDUR2 | dz`,l,f | VDUR2 | i,u,a,ei |
| CDUR3 | n,Z` | VDUR3 | yE,y,@n,in |
| CDUR4 | m,dz,dj | VDUR4 | uo,iE,ai,ou,uei |
| CDUR5 | t,p,k,h | VDUR5 | @N,oN,iN,an,au |
| CDUR6 | s`,ts`,sj,s | VDUR6 | yn,iau,aN |
| CDUR7 | ts,tj | VDUR7 | ia,iou,u@n,@`,iEn ,ua |
| CDUR8 | Zero | VDUR8 | uan,yEn,iaN,uaN,u ai,yoN |

Table2. Segment groupings for TFS, Duration

| Group | Consonant | Group | Vowel |
|---|---|---|---|
| CDUR1 | d,b,g | VDUR1 | @,o |
| CDUR2 | dz`,l | VDUR2 | o,U,ei |
| CDUR3 | n,f,Z` | VDUR3 | i,u |
| CDUR4 | m,t,p,dz,dj, k | VDUR4 | a,in,uo,y,@n,iE,yE ,ou |
| CDUR5 | h,ts` | VDUR5 | uei,iN,ai,@n,yn |
| CDUR6 | ts,sj,tj | VDUR6 | oN,iou,au,aN,iau, an |
| CDUR7 | s,s` | VDUR7 | ia,iEn,u@n,uaN,ua i,ua,uan |
| CDUR8 | Zero | VDUR8 | yEn,iaN,yoN |

Table3. Segment groupings for TMS, Duration

| Group | Consonant | Group | Vowel |
|---|---|---|---|
| CDUR1 | d,b | VDUR1 | @,U,U` |
| CDUR2 | g,l | VDUR2 | u,o |
| CDUR3 | n,dz`,Z`,m | VDUR3 | a,I,yE,ou,ua |
| CDUR4 | dz,dj | VDUR4 | ei,ai,uo,au |
| CDUR5 | t,p,f,k,h | VDUR5 | oN,in,@n,u@n,ua N,@N,an,aN,iE,iN |
| CDUR6 | ts` | VDUR6 | uei,y,ia,yn,uan,iau |
| CDUR7 | tj,s`,sj,s,ts | VDUR7 | iEn,@`,iaN,iou |
| CDUR8 | Zero | VDUR8 | yEn,uai,yoN |

Table4. Segment groupings for BFS, Duration

| Source | df | Sums of Squares | Mean Square | F-ratio | Prob |
|---|---|---|---|---|---|
| CTy | 7 | 615273 | 87896.1 | 42.767 | <0.0001 |
| VTy | 7 | 367437 | 52491.1 | 25.54 | <0.0001 |
| Ton | 4 | 109460 | 27365 | 13.315 | <0.0001 |
| Ton*FCT | 32 | 356105 | 11128.3 | 5.4146 | <0.0001 |
| CTy*Ton | 25 | 183320 | 7332.81 | 3.5679 | <0.0001 |
| VTy*Ton | 23 | 152705 | 6639.33 | 3.2305 | <0.0001 |
| PVT | 7 | 42277.2 | 6039.59 | 2.9386 | 0.0045 |
| PTn | 4 | 23319.3 | 5829.82 | 2.8366 | 0.023 |
| PCT | 8 | 37596.2 | 4699.52 | 2.2866 | 0.0193 |
| FCT*FVT | 56 | 262766 | 4692.25 | 2.2831 | <0.0001 |
| FCT*FTn | 29 | 134237 | 4628.86 | 2.2522 | 0.0001 |
| CTy*VTy | 49 | 199615 | 4073.77 | 1.9821 | <0.0001 |
| CTy*FTn | 35 | 140762 | 4021.78 | 1.9568 | 0.0006 |
| FCT | 8 | 31730.5 | 3966.32 | 1.9299 | 0.0514 |
| PCT*PTn | 25 | 95864.1 | 3834.56 | 1.8658 | 0.0056 |
| CTy*PTn | 35 | 126427 | 3612.19 | 1.7576 | 0.0039 |
| PCT*PVT | 49 | 173919 | 3549.36 | 1.727 | 0.0013 |
| PVT*PTn | 23 | 76195.2 | 3312.83 | 1.6119 | 0.0324 |
| FVT*FTn | 23 | 74873 | 3255.35 | 1.5839 | 0.0377 |
| Ton*PCT | 31 | 96176.4 | 3102.46 | 1.5095 | 0.0348 |
| VTy*FCT | 56 | 156598 | 2796.39 | 1.3606 | 0.0387 |
| Const | 1 | 2.31E+08 | 2.31E+08 | 1.12E+05 | <0.0001 |
| Error | 5118 | 1.05E+07 | 2055.23 | | |
| Total | 5654 | 2.40E+07 | | | |

Table5. ANOVA Table of Factors used in syllable layer, Duration TFS

| Source | df | Sums of Squares | Mean Square | F-ratio | Prob |
|---|---|---|---|---|---|
| CTy | 7 | 175279 | 25039.9 | 14.276 | <0.0001 |
| PVT | 9 | 74244.5 | 8249.38 | 4.7031 | <0.0001 |
| Ton*FTn | 20 | 163881 | 8194.05 | 4.6716 | <0.0001 |
| FCT | 8 | 56309.5 | 7038.69 | 4.0129 | <0.0001 |
| FVT*FTn | 26 | 178220 | 6854.63 | 3.9079 | <0.0001 |
| VTy | 7 | 45882.4 | 6554.62 | 3.7369 | 0.0005 |
| VTy*Ton | 26 | 146987 | 5653.35 | 3.2231 | <0.0001 |
| CTy*VTy | 43 | 226817 | 5274.82 | 3.0073 | <0.0001 |
| PTn | 4 | 18084.9 | 4521.23 | 2.5776 | 0.0356 |
| CTy*Ton | 26 | 116595 | 4484.44 | 2.5567 | <0.0001 |
| FCT*FTn | 29 | 128627 | 4435.42 | 2.5287 | <0.0001 |
| CTy*FTn | 35 | 122844 | 3509.82 | 2.001 | 0.0004 |
| PCT*PTn | 31 | 103806 | 3348.59 | 1.9091 | 0.0018 |
| PVT*PTn | 26 | 80244.6 | 3086.33 | 1.7596 | 0.01 |
| VTy*FVT | 49 | 148264 | 3025.8 | 1.7251 | 0.0013 |
| FCT*FVT | 50 | 146530 | 2930.6 | 1.6708 | 0.0022 |
| VTy*PCT | 56 | 148779 | 2656.77 | 1.5147 | 0.0081 |
| VTy*FTn | 28 | 72556.2 | 2591.29 | 1.4773 | 0.0504 |
| VTy*PVT | 49 | 119575 | 2440.3 | 1.3913 | 0.0373 |
| VTy*FCT | 56 | 136265 | 2433.31 | 1.3873 | 0.0301 |

| | | | | | |
|---|---|---|---|---|---|
| PTn*FCT | 39 | 92503.4 | 2371.88 | 1.3522 | 0.0709 |
| Const | 1 | 2.08E+08 | 2.08E+08 | 1.19E+05 | <0.0001 |
| Error | 5030 | 8.82E+06 | 1754.03 | | |
| Total | 5654 | 2.05E+07 | | | |

Table6. ANOVA Table of Factors used in syllable layer, Duration TMS

| Source | df | Sums of Squares | Mean Square | F-ratio | Prob |
|---|---|---|---|---|---|
| CTy | 7 | 417148 | 59592.5 | 14.966 | <0.0001 |
| Ton | 4 | 96536.5 | 24134.1 | 6.0609 | <0.0001 |
| CTy*VTy | 46 | 655573 | 14251.6 | 3.5791 | <0.0001 |
| FVT | 7 | 85042.7 | 12149 | 3.051 | 0.0033 |
| Ton*FTn | 16 | 184510 | 11531.9 | 2.896 | <0.0001 |
| VTy*Ton | 26 | 294887 | 11341.8 | 2.8483 | <0.0001 |
| FCT*FTn | 27 | 298044 | 11038.7 | 2.7722 | <0.0001 |
| Ton*FCT | 32 | 312549 | 9767.16 | 2.4529 | <0.0001 |
| FVT*FTn | 26 | 235986 | 9076.38 | 2.2794 | 0.0002 |
| FCT*FVT | 44 | 342737 | 7789.47 | 1.9562 | 0.0002 |
| CTy*Ton | 27 | 209032 | 7741.94 | 1.9443 | 0.0024 |
| FTn | 4 | 30036.5 | 7509.13 | 1.8858 | 0.11 |
| VTy*FTn | 28 | 197720 | 7061.43 | 1.7734 | 0.0073 |
| VTy*PTn | 33 | 223588 | 6775.41 | 1.7015 | 0.0075 |
| VTy*FCT | 56 | 371720 | 6637.86 | 1.667 | 0.0014 |
| FCT | 8 | 53079.5 | 6634.94 | 1.6663 | 0.1013 |
| Ton*PCT | 32 | 208780 | 6524.38 | 1.6385 | 0.0132 |
| VTy*PCT | 56 | 351417 | 6275.3 | 1.5759 | 0.0041 |
| PVT*PTn | 30 | 185741 | 6191.37 | 1.5549 | 0.0275 |
| PCT*PVT | 59 | 343430 | 5820.84 | 1.4618 | 0.0125 |
| VTy*FVT | 49 | 262980 | 5366.94 | 1.3478 | 0.0537 |
| Const | 1 | 3.91E+08 | 3.91E+08 | 98096 | <0.0001 |
| Error | 4865 | 1.94E+07 | 3981.94 | | |
| Total | 5482 | 3.89E+07 | | | |

Table7. ANOVA Table of Factors used in syllable layer, Duration BFS

**PW Layer**

Figures 3 to 5 demonstrate the regression coefficients of PW layer derived from the 3 sets of corpora. In the figures, each line represents the PW unit of different length, while the Y-axis of each point represents the derived coefficient at the specific position of a PW unit. These coefficients represent how much the syllable duration is lengthened or shortened in comparison to the prediction derived from lower layers. The general duration pattern of PW layer is 2-syllable contrast, where the second-from-last syllables was shortened and the last syllable lengthened in comparison with the predicted durations. Among the 3 sets of corpora, TFS has the smallest shorting/lengthening contrast, where the most significant difference between shortened and lengthened syllable is 19.68 ms in 4-syllable PW. The slowest speaker BFS has the biggest range of 39.52 ms in 3-syllable PW. However, all three speakers showed similar duration patterns.

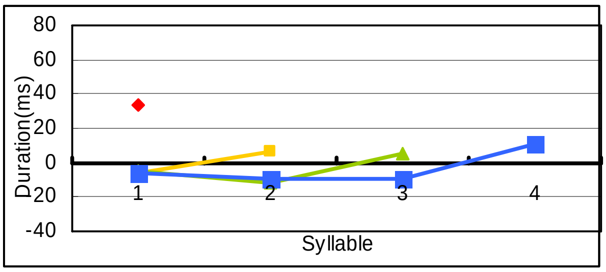


Figure 3. Regression Coefficients of PW layer, Duration, TFS

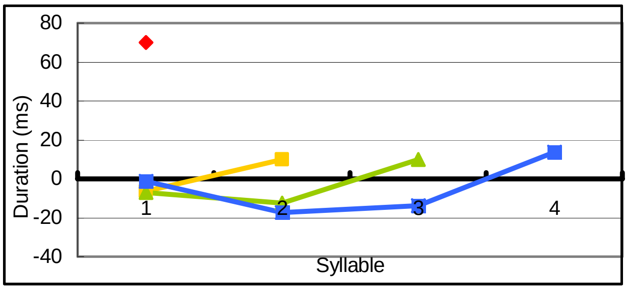


Figure 4. Regression Coefficients of PW layer, Duration, TMS

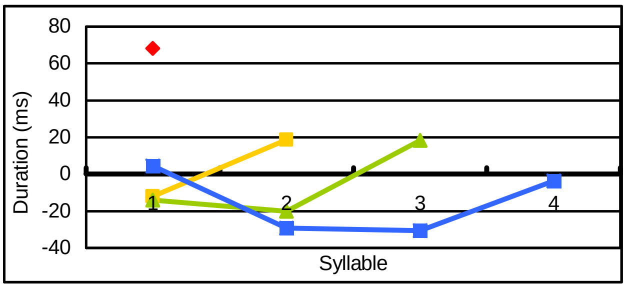


Figure 5. Regression Coefficients of PW layer, Duration, BFS

**PPh Layer**

Figures 6 to 8 demonstrate the regression coefficients of PPh layer derived from the 3 sets of corpora. The general pattern of PPh layer is found in the last four syllables of a PPh. While the fourth-from-last syllable is kept the same, the third-from-last syllable is shortened significantly to become the relatively shortest among the last four syllables. The last syllable is lengthened and usually recognized as phrase final effect. In other words, in graphic display an evident elbow shows the long-short contrast manifested in the last four syllables of PPh. All speakers exhibited the same general pattern. However, TMS and TFS showed the same pattern of final-syllable lengthening while BFS tends to only shorten the third-from-last syllable instead of lengthening the last syllable for PPhs over 5 syllables.

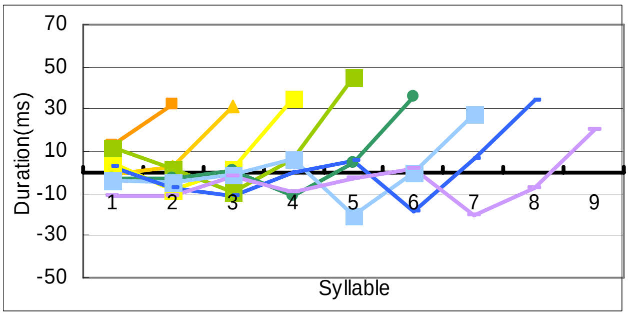


Figure 6. Regression Coefficients of PPh layer, Duration, TFS

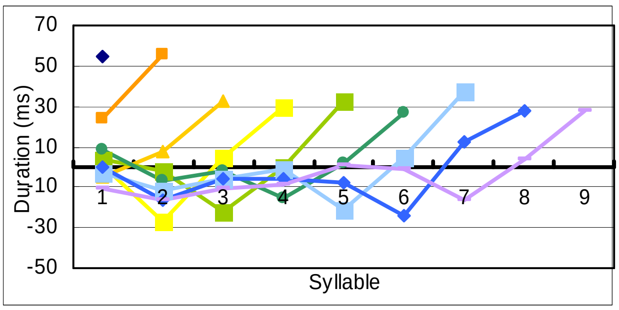


Figure 7. Regression Coefficients of PPh layer, Duration, TMS

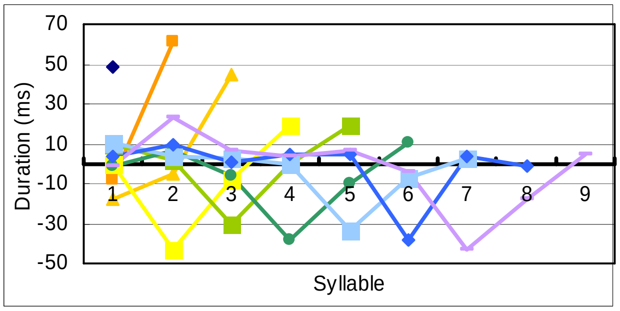


Figure 8. Regression Coefficients of PPh layer, Duration, BFS

**PG Layer**

Since in the data sets analyzed, most PGs consisted of one BG, the BG layer is in fact where the multi-phrase prosodic group PG is for the present study. As a result, relative PG positions were considered at the BG layer. PPhs were divided into three classes, PG-initial, -medial and –final.

**PG-Initial PPh**

Figures 9 to 11 demonstrate the regression coefficients of PG-initial PPh derived from the 3 sets of corpora. The general duration pattern contributed by the PG-initial PPh is similar among the 3 speakers, namely, final lengthening. PG-initial PPh exhibited relative larger degree of final lengthening to the prediction of the PPh layer. The final syllable of the PG-initial PPh was lengthened from 23.85(TFS) to 29.55 ms (BFS).

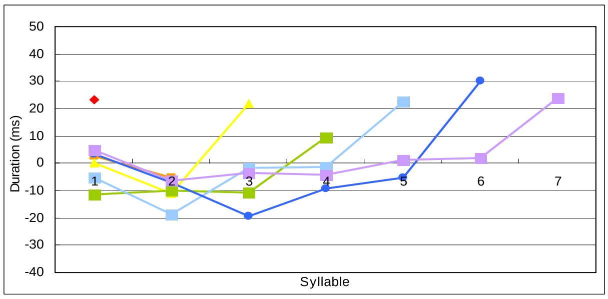


Figure 9. Regression Coefficients of Initial PPH in BG layer, Duration, TFS

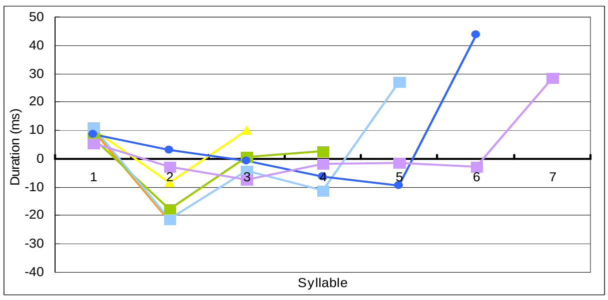


Figure 10. Regression Coefficients of Initial PPH in BG layer, Duration, TMS

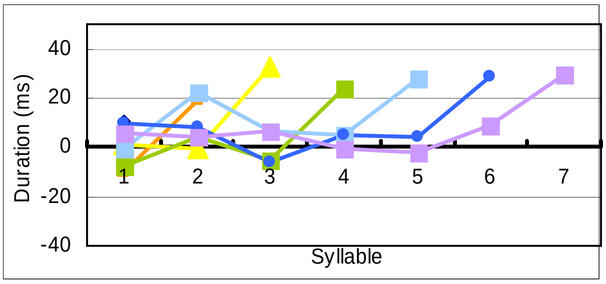


Figure 11. Regression Coefficients of Initial PPH in BG layer, Duration, BFS

### PG-Medial PPh

Figures 12 to 14 demonstrate the regression coefficients of PG-medial PPh derived from the 3 sets of corpora. In addition to PPh final lengthening, the PG-medial PPhs were further lengthened. Note that the following differences existed between the BG-initial and -medial PPhs: (1.) The PG-initial PPh has a greater lengthening effect at phrase final, which is about 10ms more than the medial ones. (2) The first syllable of PG-medial PPh was shortened by 10ms. Note that first syllable shortening was not found in the PG-initial PPh.

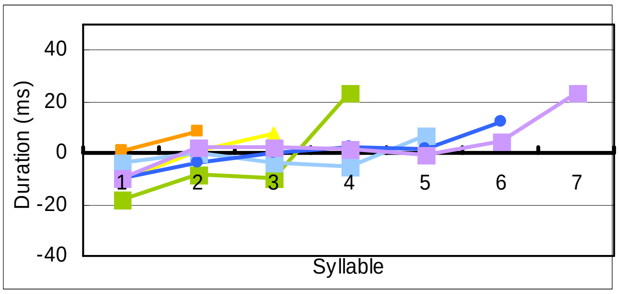


Figure 12. Regression Coefficients of Medial PPh in BG, Duration, TFS

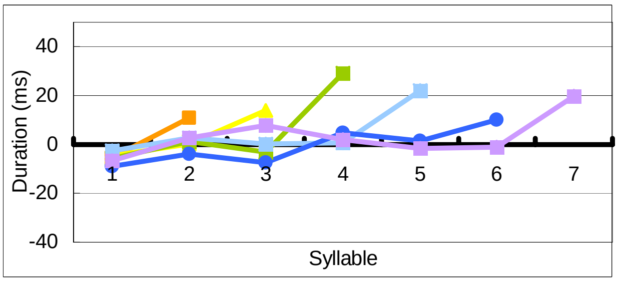


Figure 13. Regression Coefficients of Medial PPh in BG layer, Duration, TMS

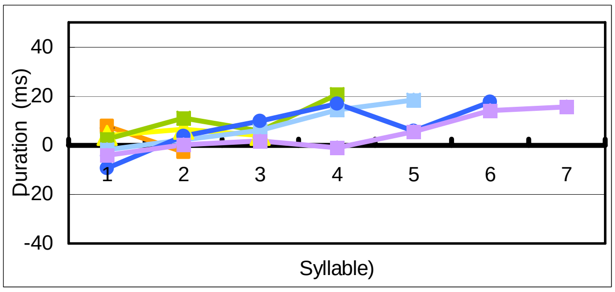


Figure 14. Regression Coefficients of Medial PPh in BG layer, Duration, BFS

### PG-Final PPh

Figures 15 to 17 demonstrate the regression coefficients of BG-final PPh derived from the 3 sets of corpora. Note that the last syllable was not lengthened as in PG-initial and -medial PPh. Instead, the last syllable was shortened significantly. For PG-final PPh over 7 syllables, it was shortened by 31.3 ms in TFS, 45.55 ms in TMS, and 18.8 ms in BFS. However, by definition the overall effect on a BG final syllable is the summation of the prediction outcome, which is still lengthening.

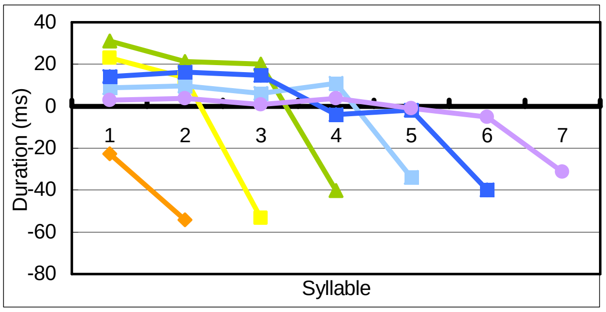


Figure 15. Regression Coefficients of Final PPh in BG layer, Duration, TFS

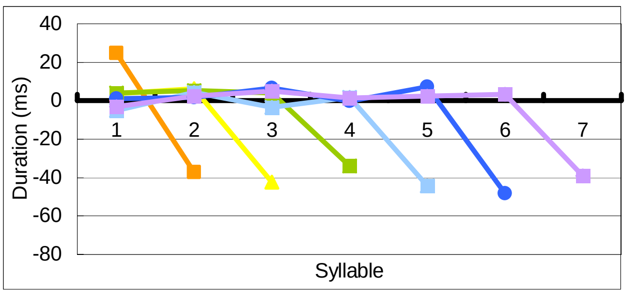


Figure 16. Regression Coefficients of Final PPh in BG layer, Duration, TMS

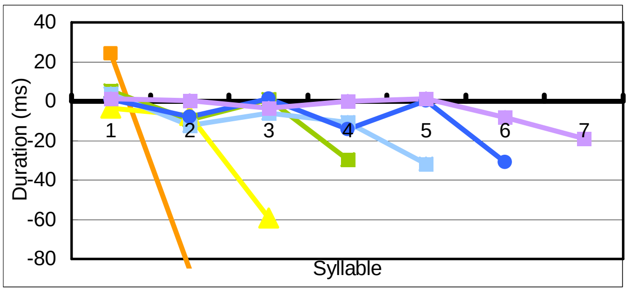


Figure 17. Regression Coefficients of Final PPh in BG layer, Duration, BFS

It is shown in BG layer that different PPhs contributed its own duration pattern over the already predicted PPh layer. The overall prediction evaluation was listed in Table8. The correlation coefficient r of the overall prediction is 0.806 in TFS, 0.842 in TMS, and 0.819 in BFS. Figure 18 plotted layered contributions in the lower column and cumulative predictions in comparison with original speech data. Note that how after trade-off at the final position; the predictions are very close to original speech data.

| Corpus / Layer | TFS | | TMS | | BFS | |
|---|---|---|---|---|---|---|
| | T.R.E. | r | T.R.E. | r | T.R.E | r |
| Syllable | 43.88% | 0.749 | 43.11% | 0.754 | 49.97% | 0.709 |
| PW | 42.43% | 0.759 | 39.87% | 0.776 | 43.22% | 0.755 |
| PP | 38.07% | 0.789 | 33.41% | 0.822 | 36.13% | 0.803 |
| BG | 35.13% | 0.806 | 29.37% | 0.842 | 33.56% | 0.819 |

Table8. Overall Evaluations on Duration Prediction by speakers.

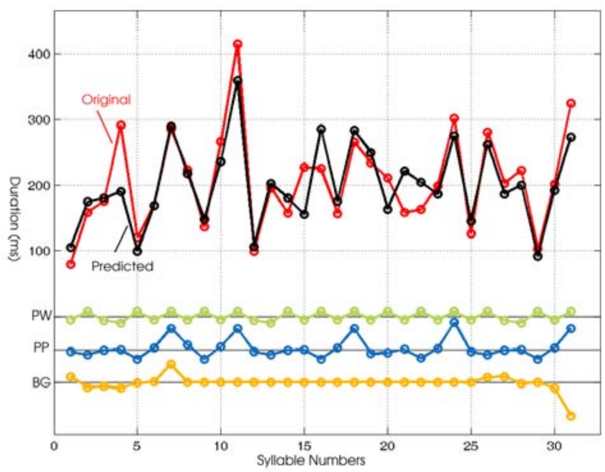


Figure 18. Comparison between cumulative predictions to original speech data of TFS. The lower column shows layered predictions at the PW, PPh and BG layers; the upper column shows cumulative predictions in black and original speech data in red.

### 2.2.2. Discussion

Figures 9 to 17 show that each prosodic layer possesses distinct duration allocation patterns and contribute to overall duration output. At the highest PG layer (collapsed with the BG layer here) the PPhs at each of the three respective positions, i.e., PG-initial, PG-medial and PG-final are characterized by three different cross-syllable cadence patterns, thus providing evidence for higher level information from the discourse. Our interpretation is that PG specified positions define respective syllable-cadence templates across phrases under grouping. Final lengthening of the last syllable occurs at both PG-initial and PG-medial PPhs but in different degrees (shown in Figures 9 though14). PG-final PPhs exhibit a reverse pattern of final syllable shortening (shown in Figures 15 to 17). However, by adding information from each layer, trade-offs occur and the PG-final lengthening is still achieved (Figure 18). These duration templates are also complimentary to PG-position related characteristics in the $F_0$ templates (Tseng et al, 2004b) where PG-initial and PG-final PPhs possess distinct intonation patterns while PG-medial PPhs do not, but their respective patterns differ. The fact that each PG-position signals different overall effect of a speech paragraph is also exhibited through duration analyses. In other words, similar but larger-scale tidal effects over waves and ripples from the highest layer are found in adjustment of syllable duration and temporal allocation patterns. Furthermore, intrinsic segment durations with respective tonal effects only are inadequate to generate prosody duration. Inherent speech rhythm exists in prosody; it requires speech units of the last three or four syllables depending on prosodic layer. Respective contribution [14] from each prosodic layer cannot account for the final duration output independently. In particular, duration prediction at the SYL layer was only around chance level, but cumulatively, over 90% of the duration output was accounted for. It is apparent why concatenating syllables with only lower level (such as lexical) [15] specifications of duration adjustment are insufficient.

The results also indicate that at the higher discourse lever, Mandarin Chinese spoken in Taiwan and Putonghua spoken in Beijing share similar duration patterns and overall tendencies. Whatever difference in speech rhythm between the two dialects would most likely be due to lower level information such as tone range and stress patterns.

In summary, we have shown that syllable-cadence exists at each prosodic level thus illustrating how higher level rhythm exists in fluent speech. We believe these templates are used for production planning and perception processing. The respective cadence patterns show that distinct rhythmic patterns exist at each prosodic layer, and explain why speech rhythm is an important prosodic feature of fluent speech. Further, we also show why output speech rhythm could not be achieved unless information from each and every prosodic layer is available. The above results also indicate clearly why concatenating isolated phrases without higher level duration specifications would not yield the rhythm of fluent speech, and why lower level (lexical and syntactic) specifications are insufficient to account for the dynamics of cross-phrase phenomena.

Of course, it is reasonable to assume that these duration templates in our study are language specific to Mandarin Chinese. But the point is higher level speech rhythm exists in every language and CAN be derived. Results obtained also lead us to argue that in narrative prosody duration patterns are as important as $F_0$ contour patterns since the former accounts for the tempo and rhythm of fluent speech while the latter the overall melody. Either one without the other is incomplete. Consequently, any modeling of narrative prosody should include language-specific cross-phrase tempo/rhythmic patterns in addition to $F_0$ contour patterns. Any prosody framework should be better enhanced by including tempo specifications. We believe these templates could also be used to construct forecasting models in speech recognition as well.

### 2.3. Intensity distribution

The same rationale for duration analyses was used to investigate intensity distribution by calculating RMS values from the lower prosodic level upward. The same linear regression analyses of speech corpora from the same 2 speakers TFS and TMS (COSPRO 05) of the above 3 speakers were performed; intensity patterns for each speaker were obtained. Similar patterns were also found across speakers and speaking rates, as with duration patterns (See Section 2.2. above). However, the following presentation reports statistical results from one speaker TFS whose duration data was presented in Figures 3, 6, 9, 12, 15 and 18) to illustrate the points. Figure 19 to 23 show derived patterns of RMS distribution of the same speech corpora used for duration analyses. Each line in the figures represents the corresponding regression coefficient of a syllable at the specific position at the specified prosodic level. Figure 19 shows intensity distribution at the PW layer where PWs from 1 to 4 syllables were analyzed. Figure 20 shows intensity distribution at the PPh layer where PPhs from 1 to 9 syllables were analyzed. Figure 21 to 23 show intensity distribution at the PG layer where PG-initial, PG-medial and PG-final PPhs from 1 to 7 syllables were analyzed.

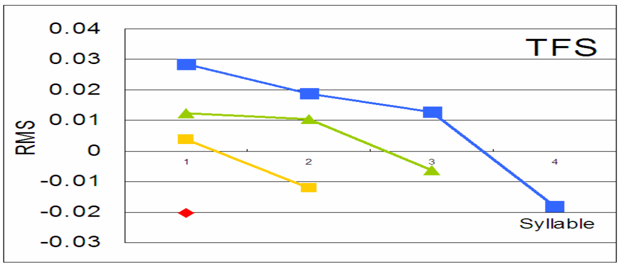


Figure 19. Regression coefficients of intensity distribution at the PW layer where PWs from 1 to 4 syllables were analyzed. A gradual decline of intensity occurred over time. Longer PWs require more energy initially.

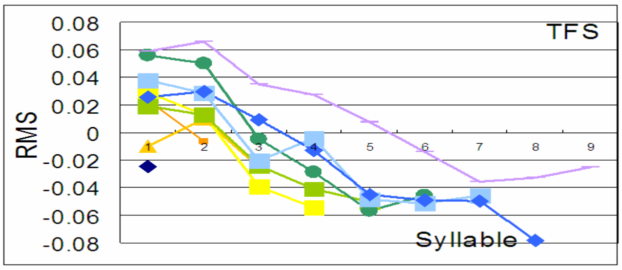


Figure 20. Regression coefficients of intensity distribution at the PPh layer where PPhs from 1 to 9 syllables were analyzed. A gradual decline of intensity occurred over time. Note how the energy level begins high and declines gradually over time and how the longer a PPh is, the more energy it requires.

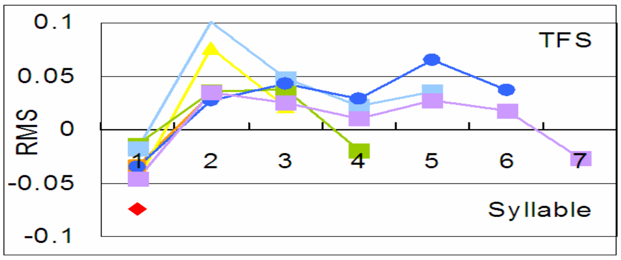


Figure 21. Regression coefficients of intensity distribution of the PG-initial PPhs at the PG layer where PPhs from 1 to 7 syllables were analyzed. The energy level is low at the first syllable, increases sharply at the second syllable, and declines with variations.

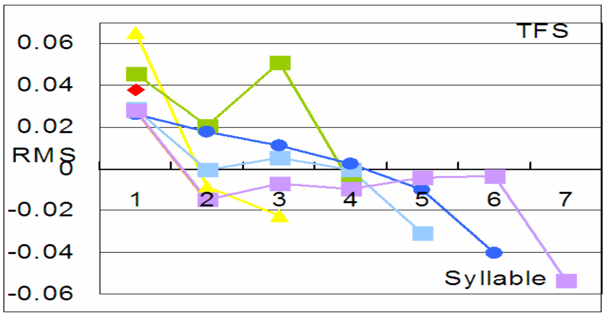


Figure 22. Regression coefficients of intensity of the PG-medial PPhs at the PG layer where PPhs from 1 to 7 syllables were analyzed. Note how energy level begins high and declines over time.

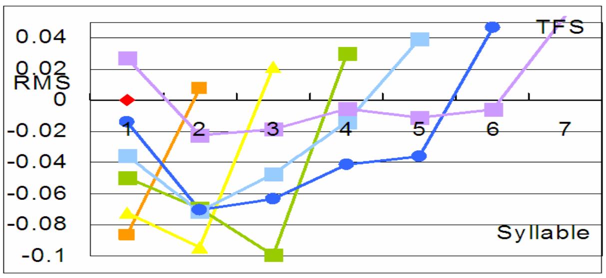


Figure 23. Regression coefficients of intensity distribution of the PG-final PPhs at the PG layer where PPhs from 1 to 7 syllables were analyzed. Note how the pattern reverses compared with the

patterns found in PG-initial (Figure 11) and PG-medial (Figure 12) PPhs. A distinct increase of energy occurred at the final syllable.

The results presented above showed that distinct patterns of intensity distribution were found to associate with each prosodic layer. Figures 19 and 20 show that at both the PW and PPh levels, a gradual decline of intensity occurs over time. In addition, the longer the unit is (more numbers of syllables in the unit) the more energy it requires at the onset. At the PG level, once again PG-relative positions show different intensity patterns as shown from Figure 21 to 23 and are in accordance with duration results. For both PG-initial and PG-medial PPhs, intensity declines in different degrees as the respective slopes in Figure 21 and 22 show. But the PG-final PPh shows a reverse pattern, with a distinct increase of energy at the final syllable. By adding information from each layer, trade-offs account for the PG-final decline of intensity, as with final lengthening found in duration patterns and $F_0$ trailing-off, and the significant role of the terminating effect occurred only at PG-final positions.

Results of percentage of contribution from each prosodic layer were also obtained, as with duration patterns. At the SYL level, segmental identity accounted for 51% of intensity distribution. At the PW level, the contribution of intensity is insignificant, although the gradual declination exists. However, at the PPh level, the contribution of intensity is accounted for 14% more of intensity distribution, indicating that the prosodic phrase is a more significant unit for amplitude distribution patterns for fluent speech than prosodic words. Moreover, the shorter final PPhs had a wider coefficient range. We believe the different cross-phrase pattern of intensity distribution is closely associated with the perceived result of the terminal end of a speech paragraph in addition to $F_0$ contours and duration patterns.

### 2.4. Boundary pauses/breaks

We have stated in Section 1 that the multi-phrase prosody framework is based on perceived unit located inside different levels of boundary breaks across the flow of fluent speech. These boundaries were annotated with a ToBI-based self-designed labeling system [3] that specified 5 degrees of break indices (BI) (See Section 2). Thus, it is important that both intra- and inter-transcriber consistencies be maintained for manually annotated speech corpora. The speech data were first automatically aligned with initial and final phones using the HTK toolkit, and then manually labeled by trained transcribers for perceived prosodic boundaries or break indices (BI). All of the corpora used were manually labeled by 3 trained transcribers independently. Intra- and inter-transcriber comparisons were obtained weekly. Corpora were considered annotated when over 85% of inter-transcriber consistencies were maintained. $F_0$, duration, and intensity analyses were performed on annotated corpora subsequently. We have analyzed speech corpora of 2 males and 4 females to look for cross-speaker patterns (COSPRO 1 & 5). Each speaker read the paragraphs of discourses in slightly various editions at around 500 syllables/characters per paragraphs, producing speech corpora of around 12000 syllables each. Four of the speakers were untrained speakers (2 males and 2 females) who read at speaking rate of 224, 362, 275 and 306 ms/syllable, respectively. Two speakers were radio announcers who read relatively faster at the speaking rate of 234 and 236 ms/syllable, respectively. We observed some distinct differences between untrained speakers and radio announcers. One was in speaking rate and another was the number and type of pauses/breaks used. In general, radio announcers used faster speaking rate (235 ms/syllable) than untrained speakers (292 ms/syllable), paused less during speaking (less B2's and B3's), and change breath more often (more B4's). Whereas the untrained speakers tend to speak slower, used more minor breaks (more B2's and B3's), but did not seem to change breath nearly as often (less B4's). The results may be representative of trained public speaking style vs. untrained informal way of speech production in pause and breathing style, with the untrained speakers sounding more halting than the announcers. Table 9 presents cross-speaker, cross-speaking-rate comparisons. Since the text each speaker read varied slightly, only overlapped text were compared. The purpose was to see how many degrees of boundary breaks exist within and across speaking rates in fluent speech.

Table 9 (A through F) presents comparison of perceptual labeling of 6 speakers' breaks of overlapped portion of read speech. Speakers F001, F01S, F03S and M02S were untrained native speakers; M051P and F051P radio announcers. For each speaker, the number of each perceived break was presented, where $\mu$ is mean duration of the break and $\sigma$ standard deviation. Note that speakers F001 and F03S were given read relatively shorter paragraphs instead of longer text, the end of each paragraph was a complete recording unit, therefore, PG-final breaks (B5) were not available for measurement and hence was labeled NULL.

| F001 | B1 | B2 | B3 | B4 | B5 |
|---|---|---|---|---|---|
| Number | 25369 | 14425 | 3163 | 71 | 1646 |
| $\mu$ / $\sigma$ | 8 / 16 | 14 / 21 | 215 / 158 | 407 / 114 | NULL |

(A) Speaker F001 (speaking rate: 224ms/syllable)

| F01S | B1 | B2 | B3 | B4 | B5 |
|---|---|---|---|---|---|
| Number | 11084 | 4698 | 3630 | 193 | 473 |
| $\mu$ / $\sigma$ | 2 / 11 | 11 / 27 | 342 / 245 | 717 / 212 | 799 / 434 |

(B) Speaker F01S (speaking rate: 362ms/syllable)

| F03S | B1 | B2 | B3 | B4 | B5 |
|---|---|---|---|---|---|
| Number | 12672 | 4888 | 4202 | 132 | 574 |
| $\mu$ / $\sigma$ | 4 / 11 | 14 / 23 | 276 / 194 | 649 / 136 | NULL |

(C) Speaker F03S (speaking rate: 275ms/syllable)

| M02S | B1 | B2 | B3 | B4 | B5 |
|---|---|---|---|---|---|
| Number | 12409 | 5046 | 4303 | 250 | 546 |
| $\mu$ / $\sigma$ | 1 / 9 | 10 / 26 | 315 / 264 | 742 / 234 | 949 / 242 |

(D) Speaker M02S (speaking rate: 306ms/syllable)

| M051P | B1 | B2 | B3 | B4 | B5 |
|---|---|---|---|---|---|
| Number | 6663 | 3327 | 1207 | 270 | 130 |

| | | | | | |
|---|---|---|---|---|---|
| μ / σ | 0 / 2 | 3 / 10 | 249 / 207 | 520 / 124 | 619 / 110 |

(E) Speaker M051P (speaking rate: 234ms/syllable)

| F051P | B1 | B2 | B3 | B4 | B5 |
|---|---|---|---|---|---|
| Number | 6645 | 3352 | 1157 | 287 | 150 |
| μ / σ | 2 / 6 | 6 / 13 | 215 / 152 | 332 / 164 | 399 / 226 |

(F) Speaker F051P (speaking rate: 236ms/syllable)

Table 9 Number of hand labeled boundary breaks by speaker where $\mu$ is mean duration of the break and $\sigma$ standard deviation .

From the mean durations, it is clear that the degrees of breaks were maintained from perceptual labeling. Moreover, when B5s were available in the data, they are longer than B4s. This indicates that in order to accommodate multiple phrase grouping of longer discourses, at least 3 levels of breaks, i.e., B3, B4 and B5 or minor break, major break and PG break, are needed for narratives of fluent speech. In other words, we believe that 2 levels of breaks, namely, minor break and major break, are inadequate to fluent running speech. We have incorporated the boundary break features into our prosody framework. Together with correlative intensity distribution patterns, the make-up of the rhythm and tempo of narrative prosody could be constructed.

#### 2.5. Summary and Discussion

The above results demonstrate that in fluent speech, higher level information is involved in the planning of speech production; speech units are no longer discrete intonation units, larger multi-phrase prosodic units reflecting higher level discourse organization are in operation during the production of fluent speech. Hence methodologically, lifting fragments from fluent speech and analyzing microscopic phonetic or acoustic details will not reveal or recover prosody information contained. We then argue further that any prosody organization and modeling should incorporate language specific patterns of duration allocation pattern and intensity distribution in addition to $F_0$ contour patterns and with respect to prosody organization. A mathematical modular model corresponding to our framework was subsequently constructed [5, 16, 17].

From the evidences presented in this section, we argue that a prosody organization of fluent connected speech should accommodate higher level discourse information above phrases and sentences, and account for the dynamic cross-phrase relationship that derives phrase groups corresponding to perceived speech paragraphs. All three acoustic correlates, namely, $F_0$, duration and amplitude, should be accounted for with respect to phrase grouping, along with at least 3 degrees of boundary breaks. $F_0$ contour patterns alone are not necessarily the most significant prosody feature, and are insufficient to characterize the major part of speech prosody. Rather, the roles of syllable duration adjustment with respect to temporal allocation over time and intensity distribution with reference to overall cross-phrase relationships merit further deliberation. Boundary breaks also require further understanding. It is quite evident from the above evidence of syllable cadence templates derived that cross-phrase duration patterns with respect to higher level prosody organization are just as important as cross-phrase $F_0$ modifications, whereas intensity patterns is also more distinct at the higher prosodic PPh layer. We believe that together with boundary breaks, these features account for the major part of melody and tempo in narrative prosody, reflecting the domain, unit and to quite an extent strategy of speaker's planning of fluent speech. In other words, these template and boundary breaks are used by the speaker for planning in speech production, and as processing apparatus by the listener as well. In summary, what is intended by the speaker through these vehicles available in prosody maneuvering are also significant to the listener's expectations during processing. Cross-phrase as well as overall template fitting, matching, and filtering could also be built into fluent speech recognition as well [18]

## 3. From paragraph to discourse

In a separate paper for this conference[19], we also discuss preliminary evidence how between paragraphs units associate paragraphs to form discourse. These in-between units are termed discourse markers (DM) and prosodic fillers (PF) for the time being, as shown in Figure 1. That is, paragraphs are discourse units; another higher node exists to relate discourse information between and among paragraphs.

## 4. Conclusions

This paper has demonstrated that one of the most important prosodic characteristics of fluent Mandarin Chinese speech cannot be obtained at the level of intonations units, but rather, reveals itself only in the examination of fluent speech of narratives or spoken discourses. The operating unit essential to the execution of fluent Mandarin speech is the multi-phrase speech paragraph; a higher-level unit which combines individual phrase and sentence intonations into a corresponding governing prosodic unit PG. PG possesses a global canonical $F_0$ template for intonation modification, a cadence template for duration adjustments, an intensity pattern for amplitude distribution, and break/pause patterns. It specifies the subjacent phrases or sentences as sister prosodic constituents, and assigns their roles with respect to PG positions. The prosody templates are the tides for the subjacent prosodic constituents to ride over, thus triggering the waves and ripples to modify accordingly. We believe that such spoken discourse effect is also cross-linguistic. Specific to Mandarin Chinese and perhaps other tone languages is that phrasal intonations are not as significant as they are in intonation languages. Language specific questions may very well be within- and cross-phrase cadence patterns since they are most likely to differ from one language to another.

Research seeking to describe explain Mandarin Chinese prosody has focused mostly on the intonation units of phrases or sentences produced and analyzed in isolation, but it remains to be seen how these units interact with higher level prosodic organization in fluent speech materials. Our perception motivated multi-phrase PG model offers at least in part a knowledge base and viable framework for formulating theories of higher prosodic organization manifested through speech prosody. We presented evidence to show why more understanding of higher level information as in discourse effects to fluent speech is essential, and how cross-phrase templates of prosody-related melodic as well as rhythmic cadence, intensity and boundary patterns may together account for the necessary speech planning in text reading and spoken discourses. We believe our framework is also capable to adopt and accommodate any discrete intonation model at the PPh level. Furthermore, the idea should also enhance technological and computational applications, in particular, speech synthesis, and may be adapted to constructing canonical complex sentence intonation for other languages.

In summary, higher level organization must be accounted for; hierarchical framework is feasible and necessary; prosodic units must accommodate and correspond to higher nodes and higher information; intonation unit is subordinate and subjacent prosody units, and finally, paragraphs are discourse prosody units as well. Hence we argue that discourse prosody is higher level prosody and is systematically accountable.

## 5. References


[1] Chang, L., Chen, K., 1995. The CKIP part-of-speech tagging system for modern Chinese texts. In: *Proceedings of 1995 International Conference on Computer Processing of Oriental Languages*, pp. 172-175.

[2] Tseng, C., Cheng, Y., Lee, W., and Huang, F., 2003. Collecting Mandarin speech databases for prosody investigations. In: *Proceedings of the Oriented COCOSDA 2003*.

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

[4] 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.

[5] Tseng, C., Pin, S., Lee, Y., Wang, H. and Chen, C. 2005a. Fluent speech prosody: Framework and modeling, *Speech Communication* (Special Issue on Quantitative Prosody Modeling for Natural Speech Description and Generation), Vol. 46:3-4, 284-309.

[6] Shattuck-Hufnagel, S., Turk, A., 1996. A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguist Research*, 25(2): 193.

[7] Gussenhove, C. 1997. Types of Focus in English? In Daniel Buring, Matthew Gordon and Chungming Lee (eds.) *Topic and Focus: Intonation and Meaning: Theoretical and Crosslinguistic Perspectives*. Dordrecht: Kluwer

[8] Selkirk, Elisabeth. 2000. The interaction of constraints on prosodic phrasing. In Prosody: Theory and Experiment, ed. Merle Horne, 231-262. Dordrecht: Kluwer Academic Publishing

[9] Tseng, C., 2002. The prosodic status of breaks in running speech: examination and evaluation. In *Proceedings of Speech Prosody 2002*, pp. 667-670.

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

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

[12] Chao, Y. R., 1968. *A Grammar of Spoken Chinese*. University of California Press, Berkeley and Los Angeles, California.

[13] Keller. E., Zellner Keller, B., 1996. A timing model for fast French. *York Papers in Linguistics, 17*, University of York. 53-75.

[14] Tseng, C. and S. Pin, 2004b. Mandarin Chinese prosodic phrase grouping and modeling - method and implications. In *Proceedings of International Symposium on Tonal Aspects of Languages - with Emphasis on Tonal Languages (TAL 2004)*, pp. 193-197.

[15] Chen, K., Tseng, C. Peng, H., Chen, C., 2004. Predicting prosodic words from lexical words - a first step towards predicting prosody from text. In: Proceedings of International Symposium on Chinese Spoken Language Processing (ISCSLP2004), pp. 173-176.

[16] Pin, S., Lee, Y., Chen, Y., Wang, H., Tseng, C., 2004. A Mandarin TTS system with an integrated prosodic model. In: *Proceedings of International Symposium on Chinese Spoken Language Processing (ISCSLP2004)*, pp. 169-172.

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

[18] Tseng, C. 2006b. Recognizing Mandarin Chinese Fluent Speech Using Prosody Information—An Initial Investigation" The 3rd International Conference on Speech Prosody 2006, May 2-5, 2006, Dresden, Germany

[19] Tseng, C., Su, Zh., Chang, C. and Tai, 2006a C. Prosodic filers and discourse markers—Discourse prosody and text prediction. *(TAL 2006)*