Fluent Speech Prosody: Framework and Modeling

Chiu-yu Tseng[1*], Shao-huang Pin[1], Yehlin Lee[1], Hsin-min Wang[2**], Yong-cheng Chen[2]

[1]Phonetics Lab, Institute of Linguistics, Academia Sinica, Taipei

[2]Institute of Information Science, Academia Sinica, Taipei

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

Abstract

The prosody of fluent connected speech is much more complicated than concatenating individual sentence intonations into strings. Prosody framework and modeling should base on more understanding of both the production and perception of fluent speech. 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 effect in fluent speech. Subsequent cross-speaker and cross-speaking-rate acoustic analyses of identified speech paragraphs revealed systematic cross-phrase 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 that governs, constrains, and groups phrases to derive speech paragraphs and show how to account for the tune and rhythm characteristic to fluent speech prosody through cross-phrase specifications. A hierarchical multi-phrase framework is constructed to account for the governing effects, with complimentary perceptual evidence. The framework specifies phrasal intonations as subjacent sister constituent subject to higher specifications; while output fluent speech prosody is cumulative results of contributions from every prosodic layer. To test our framework, we further construct a modular prosody model of multiple phrase grouping with 4 corresponding acoustic modules and have begun testing the model with speech synthesis. Finally, we argue that development of unlimited TTS could benefit most appreciably by capturing cross-phrase relationships in prosody modeling.

1. Introduction

The prosody of fluent connected speech is much more complicated than concatenating individual sentence intonations into strings. However, by linguistic definition, syntactic structure has steered the studies of intonation, sentences are the accepted units of investigation, and intonation has been the focus of prosody investigation. Standard linguistic approach always starts from constructing text of sentences before collecting speech data, corpus of discreet sentences are usually collected, and discreet intonation patterns have been the focus of investigation. This approach regards fluent speech as a succession of independent sentences and much less attention has been paid to discourse effects to fluent speech prosody. Though the term “intonation group” has often been used in discourse and conversation analyses, no consistent operational definition could be found to implement such notion to modeling. With the syntax-specified intonation patterns best applicable to simple sentences, elaborations to accommodate complex sentences are still lacking. As a result, much attention has been given to the articulatory minutiae due to physiological constraints and look-ahead during speech production, for example, how segments adjust and modify when they are strung into larger units, while relatively little attention has been devoted to how articulatory adjustment must also be made to

reflect look-ahead across phrases in fluent speech and physiological constraints include breathing and cognitive limits during speech planning. Mandarin Chinese was no exception. To illustrate, we note that one linguistic approach and subsequent prosody modeling used short utterances of 5 syllables and adopted the same physiological perspective to account for cross-syllabic look-ahead in articulatory gestures with respect to tonal co-articulation [Xu, 2002], focusing on anticipatory effects in tone concatenation at the single-sentence level. Another linguistic approach and prosody modeling used short utterances of 8 syllables to study narrow focus in intonation with emphases on the interaction between tone and intonation [Shih, 1988; Shih, 2004]. A third study also used short yes-no questions produced as isolated units and reported an overall higher register than their declarative counterparts [Lin, 2002]. Perceptual studies also emphasized on tonal effects to intonation only [Yuan, 2004]. All of these approaches took sentence intonation as default prosody units and analyzed speech data either from bottom upward, focusing on between-syllable effects or on overall register height and tendency. Consequently, prosody studies for Mandarin Chinese have pretty much stopped at the level of phrase and/or simple sentences, while all of these findings yet to be tested on fluent speech data.

It was the development of unlimited Mandarin Chinese TTS (Text-to-Speech Synthesis) that brought a necessary shift of research orientation. The reasons may appear rather language specific to Chinese on the surface, though we believe the implications are definitely cross-linguistic. There are two reasons specific to Chinese. The first reason is that Chinese is often misunderstood as a mono-syllabic language because word boundaries are not reflected in writing. The syllable-based Chinese logographs require no spacing between words, word boundaries are shown in writing, and the co-existence of mono- and poly-syllabic words are therefore not reflected in text. The second reason is also text related due to lack of morphological affixations and less rigid punctuation requirements. Periods in text could be used to denote both the end of a complex sentence or a short paragraph, making it hard to distinguish between the two. As a result, the most convenient approach adopted for syntactic analyses of text corpora has been to treat commas and periods alike and analyze one phrase at a time, as practiced by the CKIP group [Chang & Chen, 1995 or http://rocling.iis.sinica.edu.tw/CKIP/]. Consequently, the most widely adopted approach to synthesize Chinese speech was to take mono-syllables as the basic units, and single phrases or sentences as prosody units. To this day, mono-syllable based speech synthesis and short-phrase based prosody simulation are still very much practiced by the Chinese research community. The question then is whether fluent connected speech is consisted of successions of multi-phrase speech paragraphs or independent unrelated phrases and sentences. In fact, simulation of a succession of short and often declination intonations in unlimited TTS did not produce satisfactory fluent speech prosody, and more systematic account of discourse effects reflected in between- and cross-phrase relationships and was called for. In spite of recent industrial technological development that produced better synthesis systems and prosody models using data-driven approaches, we believe the calls still remain largely unanswered by the linguistic community. Similar calls may also exist in languages other than Chinese for complex sentences and discourse prosody. Thus, we believe that shifts of research paradigm may be more cross-linguistic than language specific.

It has been the research focus of our group to address between- and cross-phrase prosodic characteristics from both speech production and speech perception. Methodologically, we adopted a corpus linguistic approach and planned from data collection, perceptual saliencies, annotation design, acoustic analyses, possible explanations and finally prosody modeling. For data collection, we began from collecting relatively large amount of speech data of read discourses instead of short sentences to better reflect cognitive planning constraints involved in fluent speech production. Over 9 sets of prosody-oriented speech corpus has been collected since 1997 [Tseng et al, 2003]. Spontaneous

speech and conversation were purposely avoided to reduce rapid shift of planning strategies. For perceptual saliencies, we began our analyses by listening to the corpora discourse by discourse instead of phrase by phrase, and looked for what overall tendencies and characteristics were consistently heard and identified. For annotation system, the capacity to transcribe speech data by perceived boundaries and in units above phrases/sentence became essential to the design [Tseng & Chou, 1999]. For acoustic analyses, we included every acoustic parameter involved, namely, F_0 patterns, syllable duration, intensity distribution, and in addition, boundary breaks instead of studying F_0 patterns only. Over time, characteristics of fluent speech began to emerge. We believe we are now able to account for fluent speech prosody and have developed a multi-phrase model for it as well.

This paper attempts to provide a comprehensive account of a multi-phrase framework of fluent speech prosody and its modeling. We will show that a simple framework of multiple-phrase-grouping emphasizing on cross-phrase prosodic specifications in addition to individual intonation patterns could quite adequately account for default fluent speech prosody because fluent speech prosody includes syntactic, semantic and discourse information. What the framework accounts for is basically how speech paragraphs are perceived in fluent speech via phrase grouping, and how it (phrase grouping) provides prosodic specifications to respective individual phrases or sentences under grouping in addition to phrasal intonation. These prosodic specifications involve all four acoustic correlates, namely, F_0 contours, syllable duration adjustment, intensity patterns, and boundary breaks, and can be treated independently in modeling. We will also show how to construct a modular acoustic model on the basis of the framework that how it can be applied to speech synthesis as well.

The paper is organized as follows: Section 2 describes perceptual and acoustic aspects of the framework of Phrase Grouping (PG) of fluent connected speech, including Section 2.1. F_0 specifications, Section 2.2. Duration patterns, Section 2.3. Intensity distribution, and Section 2.4. Boundary Pauses/breaks. Section 3 presents evidence of related perceptual studies. Section 4 shows how we built the framework into a prosody model. Section 5 briefly describes a Mandarin TTS system aimed at fluent prosody generation, and initial applications of our model. Finally, we will discuss implications and future directions in Section 6

2. Phrase Grouping--Organization and Framework

The speech data of under investigation consisted of speech corpora from 60 speakers, each of them read 600 paragraphs at around-500-character (syllables) apiece. Initial perceptual analyses of overall characteristics consistently included cross-listener parsing of multiple-phrase speech paragraphs inside each discourse on perceived boundaries [Tseng & Chou, 1999]. Perceptual parsing was arrived not by structure and meaning alone, but rather, by how these speech paragraphs were heard to begin and end. In particular, we found that the location of boundary breaks in speech flow did not always correspond to syntactic boundaries or punctuation marks, further suggesting multi-phrase prosody units functioning at least partially independent from structural specifications. We have studied some of these non-overlaps at the lower level, and showed how smaller lexical items formed larger prosodic items at the word level [Chen et al, 2004]. These identified speech paragraphs were termed Prosodic Phrase Group (PG) by us [Tseng & Chou, 1999]. It was evident to us that a framework of fluent speech prosody should include multi-phrase speech paragraphs instead of individual phrases, and to explain how speech paragraphs are formed through prosodic specifications. We have subsequently shown that systematic accounts of cross-phrase characteristics are essential to characterize the prosody for Mandarin Chinese fluent speech [Tseng et al, 2004a], and will discuss more evidence in

this paper. The premise is to accept multi-phrase units as necessary prosody units in fluent speech; the question afterwards is how to arrive at an explanation of how this unit operates.

The concept of phrase grouping is not is 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 [Selkirk, 1986; Shattuck-Hufnagel & Turk, 1996; Gussenhoven, 2004]. We proposed [Tseng et al, 2004a] that by adding another layer over the syntax hierarchy, Prosodic Phrase Grouping (PG) could be seen as a higher governing node above individual sentence whereby existing linguistic definitions still apply. Under a PG, phrases are immediate subjacent constituents, constrained by PG and therefore bear sister relationships. Note that in a relatively large spoken discourse formed by a succession of PGs, it is also important to specify how these PGs can be distinguished from one and other. We reported that boundary breaks indicating where PGs begin and end are most significant perceptually [Tseng, 2002] hence PG-related position and boundaries, most notably, PG-initial vs. PG-final, are most significant. The contrast with respect to PG specified positions is particularly important not only within PGs but also across them. Because amid a succession of PGs in a spoken discourse, the ending of a PG is always followed by the beginning of another PG, meaning PG-final characteristics are often followed by PG-initial characteristics in a spoken discourse where the sharpest contrast occurs. We will specify the unit that PG positions specify in subsequent discussions, as well as their respective features. In addition, PG specifications also include what happens to individual intonation of each phrase under grouping.

The multi-phrase framework presented below assumes an independent level and scope that most likely reflects the scope as well as threshold of discourse planning in the cognitive domain. Hence canonical and default global templates may exist for multiple phrases, and contribute to the look-ahead effects within and across phrases. The operation of such templates can be seen as additional to or over physiologically conditioned look-ahead during speech production so that cross-phrase look-ahead and anticipation can be added to physiologically-conditioned articulatory maneuvers at the segmental, tonal, and phrasal levels. By analogy to earlier work well-known to the Chinese linguistic community that describes the interacting and trading relationships between and tones and sentence intonation as "…*small ripples riding on large waves*" [Chao, 1968: 39], our framework assumes that larger and higher layer(s) may be superimposed over intonation and tones as tides over both waves and ripples. So the question then is what the tide is like and how ripples ride over waves and waves over tides.

The framework is based on the perceived units located inside different levels of boundary breaks across speech flow, and how these units and boundary breaks form multi-phrase speech paragraphs. The boundaries are annotated using a ToBI-based self-designed labeling system [Tseng & Chou, 1999] 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 [Tseng 2002, 2003, Tseng & Lee, 2004c]. The hierarchical governing and constraining functions of PG over phrases are illustrated schematically in Figure 1, whereby the framework can also be viewed as a tree-branching organization of multi-phrase prosody. Units used were perceived prosodic entities.

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. 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, the unit where intonation pattern applies 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. This also indicates a PG always ends and begins with a new breathing cycle. Viewed from bottom upward, the framework also accounts for how PG groups PPhs and other lower nodes.

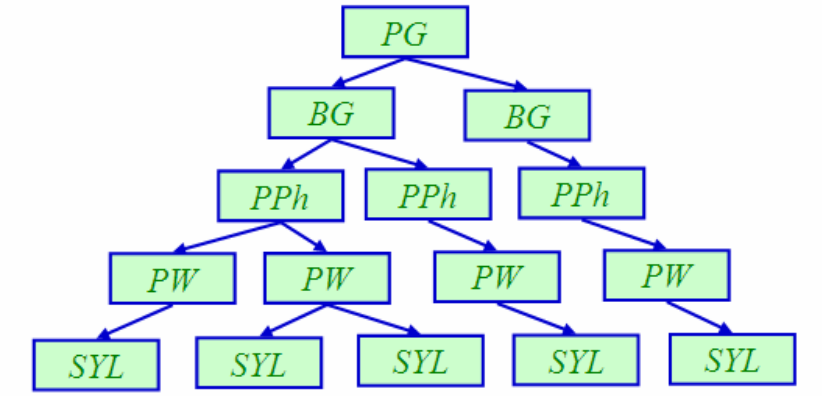


Figure1. A schematic representation of the hierarchical organization of multiple phrase grouping on perceived units and boundaries. Note that boundary breaks are not represented. However, the framework includes boundary breaks from B1 to B5. B1 is the perceived break between SYLs that may not correspond to a pause, B2 between PWs and from the PW layer up actual pauses, B3 between PPhs, B4 between BGs and B5 between PGs.

2. 1. Global F_0 patterns of PG

A canonical overall PG F_0 contour template from perceptual results was proposed to describe the overall tune of a multi-phrase speech paragraph [Tseng et al 2004a]. The unit of this template is PPh which can be either a phrase or a short sentence. In the same study, we also showed from our corpus analyses that a PG could be anywhere from 3 to 12 phrases. Figure 2 is a schematic representation of a PG of 5 PPhs, separated by major breaks B3 in the speech flow. Note that B4 usually occur only when a PG exceeds 5 phrases.

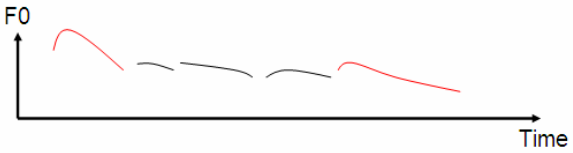


Figure2. Schematic illustration of the global trajectory of perceived F_0 contours of a 5-PPh PG. The units are PPhs separated by boundary breaks B3s. Note only the first and last PPhs (in red) into a PG possess identifiable declarative intonations. The two declination slopes in red are significantly different: the first PPh features a F_0 reset before declining rapidly, while final PPh features a lower F_0 reset before declining slowly. The declination of the first PPh does not reach to terminal end, nor does it trail while the declination of the last PPh is marked by final lengthening. The trajectories of

the three medial 3 PPhs (in black) do not possess distinct intonation patterns and are held flatter towards each of their respective ending boundaries.

Regarding perceived pitch patterns, the beginning and ending of a speech paragraph are signaled by a number of acoustic characteristics, including long pauses before F_0 reset, followed by pitch declination. The pitch contours of the first and last PPh of a PG could be described to possess distinct and identifiable intonation patterns, and in fact, are the ONLY positions where such intonations occur across pharses. The rest of the phrases in between do not possess distinct intonation patterns. We have termed them PG-initial PPh, PG-medial PPh(s) and PG-final PPh. However, note also how the two respective declarative intonations at PG-initial and PG-final positions are perceived differently. The intonation of the PG-initial PPh is marked by a F_0 reset before declining rapidly, but the decline stops short before reaching a terminal fall, nor is there final lengthening. Whereas the intonation of the PG-final PPh also possesses a F_0 reset, though not to the point of the PG-initial reset, then the contour trails to an ending with final lengthening. In the graphic display, the difference is represented by different slope of declination. As for the perceived contours of the three PG-medial PPhs, the F_0 contours are held somewhat flat, a feature related to an on-going effect. The F_0 reset and the non-terminal fall in PG-initial intonation indicates a new beginning to be followed by more speech, the less distinct and flat pattern indicate a continuing effect, while another but lower reset, together with the following gradual decline and final lengthening indicates the definite approaching of the overall terminal effect. With these three PG-specified positions, respective intonation patterns also represent cross-phrase relationship and global cross-phrase look-ahead of a speech paragraph, as does the global default melody of a multi-phrase speech paragraphs, marked by features of beginning, continuation, and termination. This is essentially why in fluent speech of a succession of PGs, a speech paragraph is always and easily perceived and the PG-initial intonation and the PG-final intonation are never confused by the listener. Taken one phrase at a time as intonation units, some phrases have distinct intonations and others don't, while those do have the same kind of intonations also differ in slope. The major features of a multi-phrase speech paragraph are where it begins, how long it is held, and when the end is finally approaching, and how it finally ends. In short, this is why concatenating independent and distinct phrasal intonations would not yield prosody of fluent speech. Hence, a multi-phrase F_0 template can be seen as a global trajectory or the *tide* over waves and ripples. Further, the F_0 template also represents how together these phrases form one prosodic unit above individual phases and how effect discourses can be achieved.

Therefore, when individual phrases are grouped into speech paragraphs, within-PG positions specify respective phrasal intonations to modify, along with PG-specified boundary breaks. Note that the template could easily be expanded to accommodate more than 5 PPhs by increasing the number of PG-medial PPhs only. Moreover, the relatively non-distinct contour patterns of these PG-medial PPhs explain why in fluent speech not all phrases possess identifiable intonation patterns, In addition, the default template could easily accommodate further implementation of emphasis, focus, and/or prominence and be further elaborated to studies of F_0 range. The question now is whether we could find proof for such a cross-phrase template in operation.

2.2. Duration patterns within and across phrases

This section of duration patterns presents results from corpora analyses of how each prosodic layer in the framework accounts for the duration pattern across syllables and contributes to the final duration outcome of phrases under grouping and how there exists an overall cross-phrase cadence

pattern. Syllable-cadence templates from each prosodic layer are derived to account for the rhythmic structure associated with prosody organization. In the discussion below, **duration** and **syllable duration** are used interchangeably.

The syllable is a more significant phonological unit of Mandarin Chinese: it is the unit of lexical tones and a temporal unit. Chinese is also a syllable-timed language instead of a stress-timed one. In other words, in terms of temporal structure, Mandarin Chinese should be treated in league with French than with English. The statistical method we used was a linear regression model that analyzes and predicts perceptually annotated speech corpora from the lower levels upward, specifying that residues that could not be accounted at a lower prosodic layer be moved up to the next higher layer. This method made it possible to account for the contribution from each respective, and to test how much the cumulative predictions from all prosodic layers involved could account for the final output.

Mandarin speech data representing 2 different speech rates, slower vs. faster speech, were used. The slower speech was recorded from 1 male untrained subject (hence SMS for Slower Male Speech) reading 595 paragraphs ranging from 2 to 180 syllables; the faster speech from 1 female radio announcer's relatively faster reading (hence FFS for Faster Female Speech) of 26 long paragraphs ranging from 85 to 981 syllables. 90% of the two sets of text overlap. A total of 22350 syllables of SMS and 11592 syllables of FFS were analyzed. Average syllable duration was 304.7ms for SMS and 199.75ms for FFS. Both sets of speech data were first labeled automatically for segments using the HTK toolkit and SAMPA-T notations [Tseng and Chou, 1999], then hand labeled for perceived prosodic boundaries by 3 trained transcribers for cross-listener consistencies. The HTK labeling was manually spot-checked; the manual perceptual labeling cross-checked for intra-transcriber consistency. Analyses were performed to (1.) compare duration variations with respect to different speech rates, and (2.) look for any possible interaction between speech rate and prosody units/levels.

Using a step-wise regression technique, a linear model with four layers [Keller and Zellner Keller, 1996] was developed and modified for Mandarin Chinese to predict speakers' tempo and rhythm with respect to the two different speech rates. Our framework of layered, hierarchical organization of prosody levels (the aforementioned system of boundaries and units) was used to classify prosodic units at levels of the syllable SYL, PW, PPh, and PG, with PG being the highest node of the hierarchy. Note that BG layer was not represented for lack of enough annotated data from the two speakers under investigation. Moving from the syllable 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.

2.2.1. Results

At the syllable layer, we examined the influence of segmental duration on syllable duration, the influence of preceding and following syllables on segmental duration and the possibility that tones may also interact with duration. Factors considered included 21 consonants, 39 vowels (including diphthongs), and 5 tones (including 4 lexical tones and 1 neutral tone). Classifications of segments were established to help simplify analyses of the speech data, which varied for the two different speech rates.

A Syllable-Layer Model was subsequently postulated as follows:

$$
\begin{aligned}
Dur\ (ms) = \quad & constant + CTy + VTy + Ton \\
& + PCt + PVt + PrT + FCt + FVt + FlT \\
& + 2\text{-}way\ factors\ of\ each\ factors\ above \\
& + 3\text{-}way\ factors\ of\ each\ syllable + \\
& + Delta\ 1
\end{aligned}
$$

CTy, VTy and Ton represent consonant type, vowel type and tone respectively. Prefix of P and F represent the corresponding factors of the preceding and following syllable. A total of 49 factors were considered. A linear model for discrete data was built using Data Desk with partial sums of squares (type 3). Factors with a p-value of under 0.5 were excluded from the analyses.

Table 1 shows benchmark values of the Syllable-Layer Model found in the two different speech rates. The residual error was 48.9% in SMS and 40.1% in FFS. In other words, the Model was able to account for 51.1% of syllable duration of the SMS and 59.6% of the FFS at the syllable layer. The residue that could not be accounted for at this layer was termed as Delta 1 and was dealt with at higher layers.

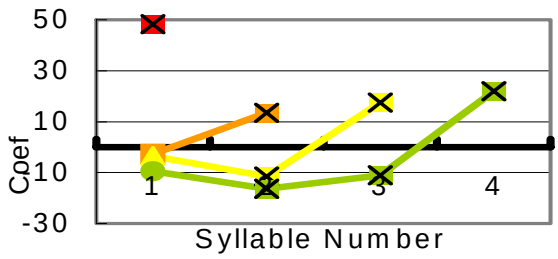
Table 1: Evaluation of duration predictions at the Syllable Layer

Test	SMS	FFS
R.E.	48.9%	40.1%
r	0.715	0.768

The same rationale was applied at the layer directly above the syllable layer, i.e., the PW layer, to investigate the possibility that a duration effect was caused by PW structure. Thus, the PW Layer Model can be written as follows:

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

Each syllable was labeled with a set of vector values; for example, (3, 2) denotes that the unit under consideration is the second syllable in a 3-syllable PW. The coefficient of each entry was calculated using linear regression techniques identical to those of the preceding layer. Figure 3 illustrates the coefficients of different PW durations for both speech rates. Positive coefficients represent lengthened syllable durations at the PW layer; negative ones represent shortened syllable durations. PWs over 5 syllables were not considered, due to their under-representation in the data.



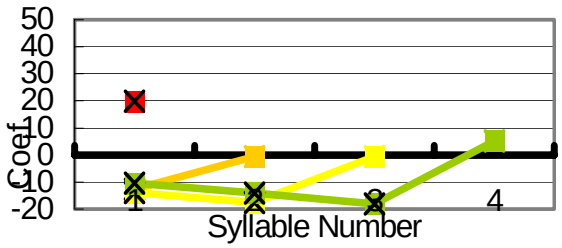


Figure 3: Coefficients of syllable durations obtained for both speech rates using the PW model. The horizontal axis represents the position of each syllable within a PW; the vertical axis represents the coefficient values. The upper panel shows coefficients of FFS; the lower panel shows those of SMS. Positive coefficients represent lengthened syllable durations at the PW layer; negative ones represent shortened syllable durations. The X labels in the figure mark coefficients of p-values smaller than 0.1.

Several interesting phenomena were observed: (1.) both speakers exhibit a pattern of PW-initial shortening followed by PW-final syllable lengthening relative to the other syllables considered; (2.) the longer the PW, the shorter the pre-final syllable is and longer the final syllable is lengthened and (3.) different speech rates contribute to different degrees of variation in syllable duration. At the PW Layer, SMS showed within-layer syllable shortening but final-syllable lengthening in comparison with lengthening predictions made at the Syllable Layer. However, FFS showed the opposite: even when syllables of a PW were shortened, the final syllable maintained the duration predicted by the Syllable Layer. These results could be used to characterize speaker-independent beat and tempo, and could be a major feature used to describe and characterize individual speaking style. Table 2 shows benchmark values of the PW Model.

Table 2: Evaluation of duration predictions at the PW Layer

Test	SMS	FFS
R.E.	93.3%	96.45%
T.R.E	45.6%	38.76%
r	0.737	0.778

The model was able to account for 6.7% of Delta 1 of SMS and 3.55% of FFS at the PW layer. The overall prediction was obtained by adding up the predicted value of both the syllable layer and the PW layer. The Total Residual Error (TRE) is the percentage of sum-squared residue over the sum-squared syllable duration. This result indicates that the residual error ratio cannot be accounted for by either layer discussed so far, and it will be dealt with at the next layer up.

The same rationale was applied to this layer. The linear regression model is thus formulated as follows.

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

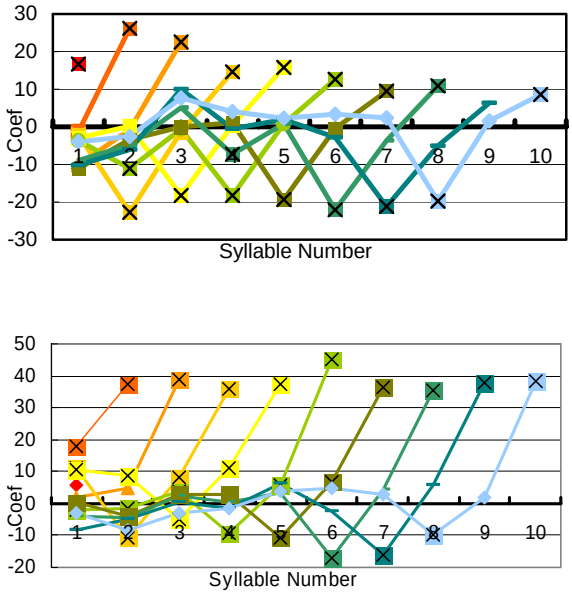


Figure 4: Coefficients of syllable durations obtained for both speech rates from the PPh model. The horizontal axis represents the position of each syllable within a PPh; the vertical axis represents the coefficient values. The upper panel shows the coefficients of FFS; the lower panel shows those of SMS. Positive coefficients represent lengthened syllable durations at the PPh layer; negative coefficients represent shortened syllable durations. The X labels in the figure mark coefficients of p-values smaller than 0.1.

Figure 4 shows the following results: (1.) A clear cadence phenomenon in PPh. (2.) That there is not only PPh-final syllable lengthening of the last two syllables, but also shortening of the antepenultimate syllable, which is an important feature of tempo structure in Mandarin Chinese (3.) Final-syllable lengthening at the PPh layer, which was twice as long for FFS, demonstrating the independent contribution of speech rate to tempo and rhythm, apart from individual speaker variation. (4.) A complementary effect of final-syllable lengthening between the PW Layer and the current PPh Layer, which may cause some trade-off in the final output. In other words, if the final syllable of a PW is lengthened, that same degree of final-syllable lengthening will NOT be found at the PPh level. Table 3 shows the evaluation of predictions at the PPh Layer.

Table 3: Evaluation of duration prediction at the PPh Layer.

Test	SMS	FFS
R.E.	93.0 %	86.5 %
T.R.E	42.4%	33.5%
r	0.760	0.814

The current PPh layer could account for only 13.5% of FFS and 7% of SMS Delta 2, where the correlation coefficient r is 0.814. The remaining residue that could not be accounted for was termed as Delta 3, which was dealt with in the layer directly above.

In order to investigate the influence of syllable duration on breath-group effect (the longer pause created by breathing which follows a PG), we studied the residue from the PPh Layer (Delta 3) at the PG Layer. Duration differences were found to occur more often at the initial and the final portions of a PPh. The initial-, medial-, and final- PPhs within a PG were also influenced by syllable duration patterns differently. We postulate that PG exerts duration effects on the initial and final portions of each PG-internal PPh, but not on the middle portion. More importantly, PG-internal positions constrain higher prosodic layers only. Table 11 summarizes the results of these evaluations.

The PG layer could account for 2.2% of delta 3 in SMS and 5.2% in FFS. The overall prediction correlates with the original corpus at the correlation coefficient r = 0.766 for SMS and 0.825 for FFS, an encouraging outcome for the current investigations.

Table 4: Evaluation of duration predictions at the PG Layer.

Test	SMS	FFS
R.E.	97.8 %	94.8%
T.R.E	41.52%	31.7%
r	0.766	0.825

The effect from the PG Layer on the next layer down (the PPh) is shown in Figure 5. Each figure illustrates the influences on the duration of the PPh under 6 syllables. Influences on the first and the last 3 syllables of PPh over 6 syllables were calculated and are shown in purple. Both speech rates showed lengthening by 10 to 20ms on the first and last syllables. In other words, duration adjustments are quite
pronounced for PG-initial PPhs.

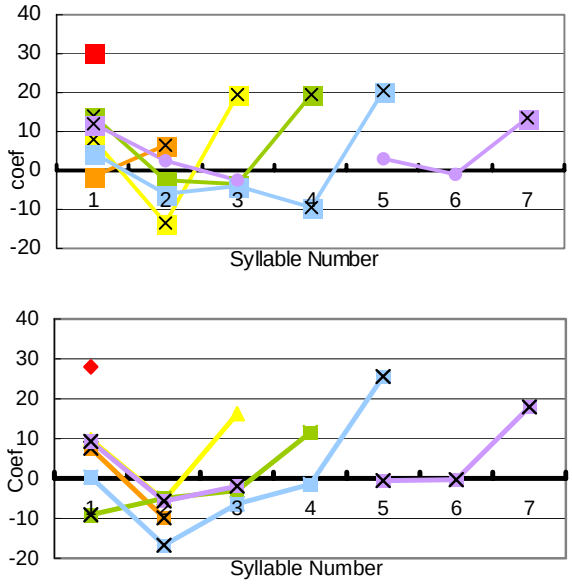


Figure 5: Illustration of the coefficients of the initial PPhs at the PG Layer. The upper panel shows coefficients of syllable durations of SMS; the lower panel shows coefficients of FFS.

Figure 6 shows effects of the PG layer on PG-medial PPhs. The first syllable is shortened while the final one is lengthened for the PG-medial PPhs considered, although this influence is more

pronounced in FFS than in SMS. However, duration adjustments for PG-medial PPhs are not as pronounced as BG-initial ones.

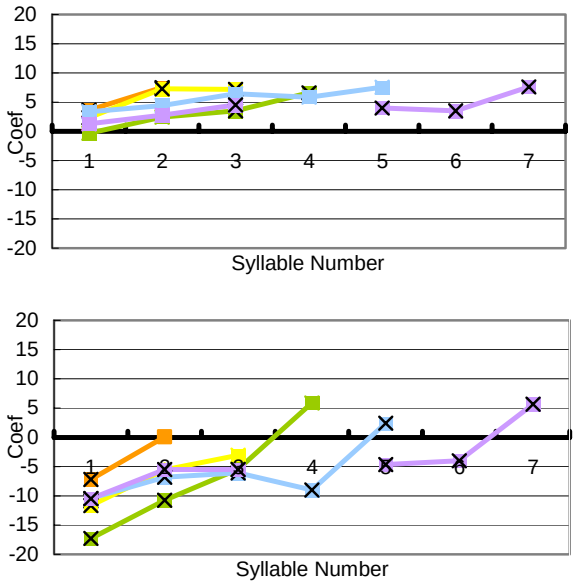
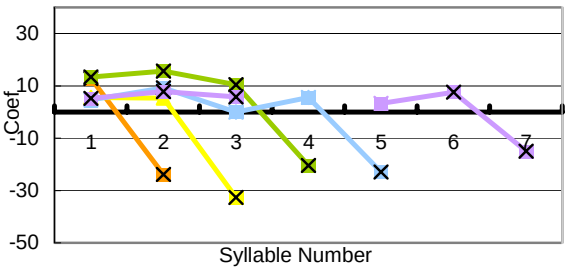


Figure 6: Illustration of the coefficients of medial PPhs at the PG Layer. The upper panel shows coefficients of syllable durations of SMS; the lower panel shows coefficients of FFS.

Figure 7 illustrates the coefficients of final PPhs. In contrast with initial PPhs, the final syllable of final PPhs is shortened. Note that the overall effect of final-syllable lengthening at the PG Layer is still present. The negative coefficients reflect a clear distinction between PG-initial and PG-final prosodic phrases.

Duration adjustments with respect to position provide further evidence of how prosodic units and layers function as constraints on syllable duration in speech flow and how higher-level prosodic units may be constrained by factors that differ from those constraining lower-level units.



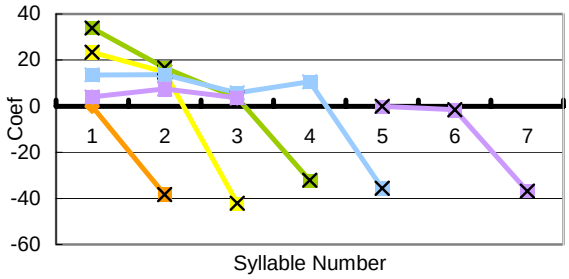


Figure 7: Illustration of the coefficients of final PPhs at the PG Layer. The upper panel shows coefficients of syllable durations of SMS; the lower panel shows coefficients of FFS.

Finally, by adding up the predictions of each prosodic layer, we can derive a total prediction of temporal allocation across phrases under grouping. Figure 8 shows comparisons between the model's prediction and the original speech data. Its prediction is quite close to the original speech data, for both fast and slow speech rates. Since the model's prediction at the syllable layer was only slightly above chance level (see Table 1), the final cumulative predictions indicate that patterns of temporal allocation in Mandarin speech flow can be accounted for only by including all levels of prosodic information. Moreover, these results can also be seen as evidence of prosodic organization in operation.

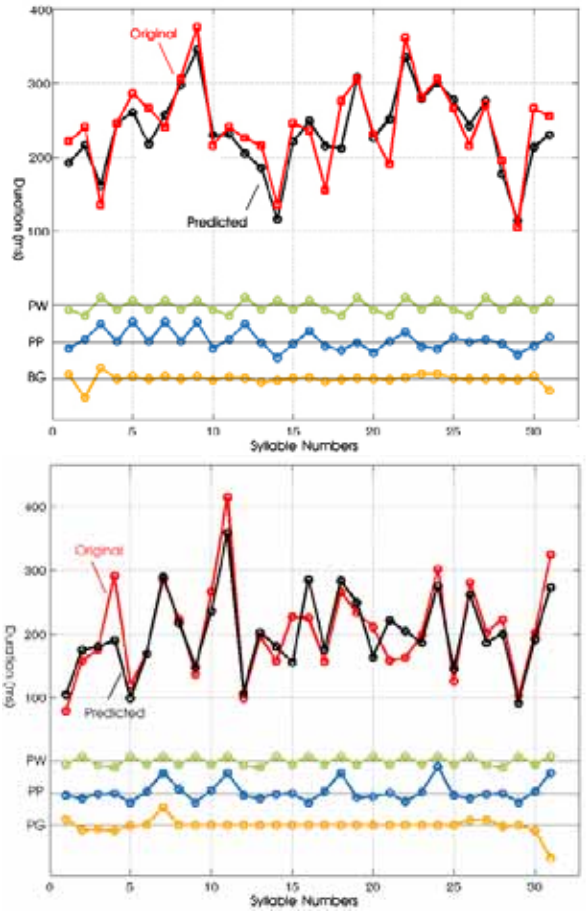


Figure 8: The upper portion shows a comparison of derived predictions from all prosodic layers combined (in black) to the original speech data (in red). The lower portion shows the prediction generated at each prosodic layer. The upper panel shows coefficients of syllable durations of SMS; the lower panel shows coefficients of FFS.

2.2.2. Discussion

Figures 5 to 7 show that at the highest PG layer, 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. 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 5 and 6). PG-final PPhs exhibit a reverse pattern of final syllable shortening (shown in Figure 7). However, by adding information from each layer, trade-offs occur and the PG-final lengthening is still achieved. These duration templates are also complimentary to PG-position related characteristics in the F_0 templates (See Section 2.1.) 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, a similar but larger-scale tidal effects over waves and ripples from the highest layer are found in adjustment of syllable duration and temporal allocation.

Furthermore, respective contribution [Tseng et al, 2004C] from each prosodic layer cannot account for the final duration output independently. In particular, duration prediction at the syllable 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) specifications of duration adjustment is insufficient. In summary, we have shown that syllable-cadence exist at each prosodic level, and believe they are cognitively based. The respective cadence patterns show that distinct rhythmic patterns exist at each prosodic layer, and explain why the rhythm of fluent speech could not be achieved unless information from each and every prosodic layer is available. We believe the cadences very likely represent cognitive templates used in both the planning and parsing of fluent speech, with the upper level templates accounting for the global look-ahead involved in speech production as well as strategies developed for parsing and processing speech.

From our results, it is quite clear why concatenating isolated phrases without higher level duration specifications simply would not yield desirable 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, these duration templates in our study are language specific to Mandarin Chinese, and different syllable-cadence templates exists in every language. Results obtained also lead us to argue that in fluent speech 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. Consequently, any modeling of fluent speech 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 were performed by speaker for corpora of 6 speakers and 2 speaking rates, and 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 to illustrate the points. Figures 8 through 12 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 8 shows intensity distribution at the PW layer where PWs from 1 to 4 syllables were analyzed. Figure 9 shows intensity distribution at the PPh layer where PPhs from 1 to 9 syllables were analyzed. Figures 11 to 13 show intensity distribution at the PG layer where PG-initial, PG-medial and PG-final PPhs from 1 to 7 syllables were analyzed.

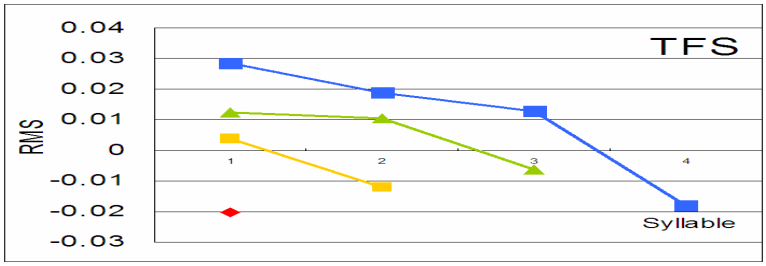


Figure 9. Regression Coefficients of Intensity distribution at the PW layer where PWs from 1 to 4 syllables were analyzed. A gradual decline of intensity occurred over time. Longer PWs require more energy initially.

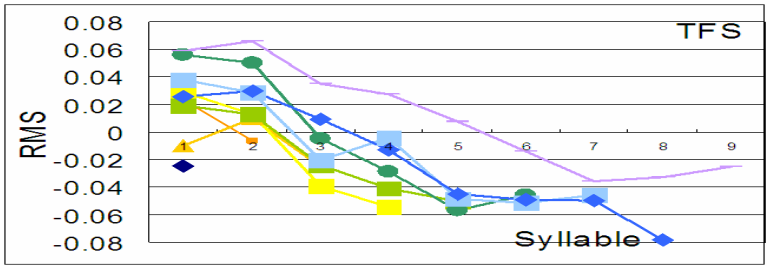


Figure10. Regression Coefficients of intensity distribution at the PPh layer where PPhs from 1 to 9 syllables were analyzed. A gradual decline of intensity occurred over time. Note how the energy level begins high and declines gradually over time and how the longer a PPh is, the more energy it requires.

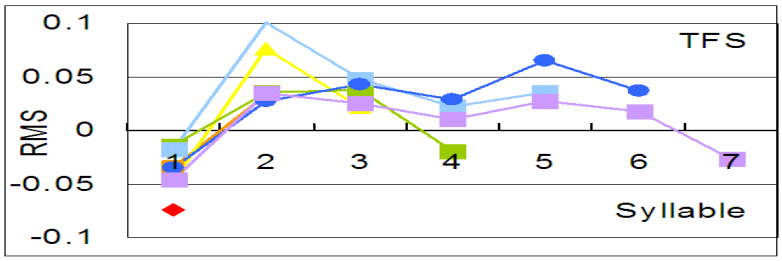


Figure11. Regression Coefficients of Intensity distribution of the PG-initial PPhs at the PG layer where PPhs from 1 to 7 syllables were analyzed. The energy level is low at the first syllable, increases sharply at the second syllable, and declines with variations.

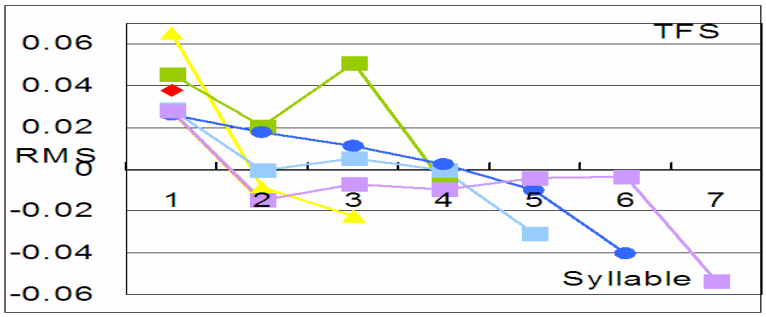


Figure12. Regression Coefficients of Intensity of the PG-medial PPhs at the PG layer where PPhs from 1 to 7 syllables were analyzed. Note how energy level begins high and declines over time.

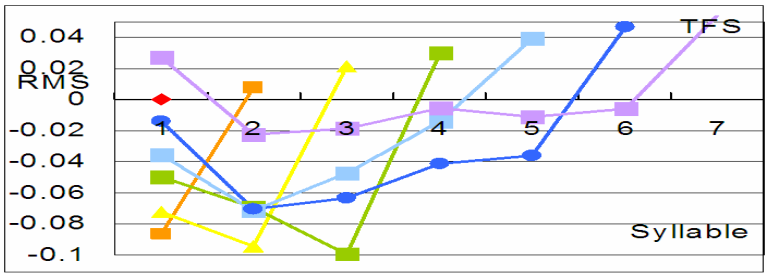


Figure13. Regression Coefficients of Intensity distribution of the PG-final PPhs at the PG layer where PPhs from 1 to 7 syllables were analyzed. Note how the pattern reverses compared with the patterns found in PG-initial (Figure10) and PG-medial (Figure11) PPhs. A distinct increase of energy occurred at the final syllable.

The results presented above showed that distinct patterns of intensity distribution are found to be associated with each prosodic layer. Figures 9 and 10 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 it initially. At the PG level, once again PG-relative positions show different intensity patterns as shown from Figures 11 to 13 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 Figures 11 and 12 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 PG-final positions.

Results of percentage of contribution from each prosodic layer were also obtained, as with duration patterns. At the syllable 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 he 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. Methodologically, it also indicates that fluent speech operates in bigger prosodic units. Lifting fragments from fluent speech and analyzing microscopic phonetic or acoustic details will not recover prosody information contained. We then argue further that any prosody organization and modeling should incorporate language specific patterns of intensity distribution in addition to F_0 contour patterns tempo/rhythmic patterns with respect to prosody organization.

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 [Tseng & Chou, 1999] 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 3 males and 3 females to look for cross-speaker patterns. 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 slowly at the average speaking rate of around 300 ms/syllable; two were radio announcers who read relatively faster at the average of 200 ms/syllable. One major difference observed was the number and type of pauses/breaks used between slower and faster speaking rates. In general, the faster the speaking rate was (200 ms/syllable), fewer minor breaks (almost no B2's and less B3's) occurred, but the speaker took breath more often (more B4's). Whereas the slower the speaking rate was (300 ms/syllable), more minor breaks (more B2's and B3's) occurred but the speaker did not seem to change breath nearly as often. The results may be representative of more stressed vs. relaxed way of speech production in pause and breathing style. Table 5 is an example of cross-speaker, cross-speaking-rate comparison. Since the text each speaker read varied slightly, only overlapped text were compared. We did not comparison of B2's for lack of incognizant amount in the speech data of faster speaking rate. The purpose was to see is how many degrees of boundary breaks exist within and across speaking rates in fluent speech.

Table 5. Comparison of perceptual labeling of two speakers' breaks of overlapped portion of read speech. F051 is the data of a female speaker in faster speaking rate; F03 in slower speaking rate. The first row of the left column denote the total number of syllables produced, the second to fourth rows numbers of B3, B4, and B5 manually annotated, respectively where B3s are boundary breaks after PPhs, B4 after BG and B5 after PG. The last 3 rows of the left column show the mean duration of boundary breaks B3, B4 and B5 and standard deviation of pauses of each break. The other two columns are speech data from two speakers F051 and F03 of faster vs. slower speaking rates. Note that the duration of speaker F03's PG-final breaks was not available and hence NA.

	F051(faste	F03(slowe

	r speaking)	r speaking)
Total # of Syllables produced	9771	9638
Total # of B3	1044	1775
Total # of B4	191	46
Total # of B5	119	333
B3(μ / σ)ms	230/154	267/189
B4(μ / σ)ms	363/182	629/126
B5(μ / σ)ms	342/277	NA

Results from Table 5 show identified breaks across speakers. Our analyses showed that at least 3 levels of breaks were consistently maintained across speakers and speaking rate. We believe that at least 3 levels of boundary breaks instead of 2 are necessary to accommodate multiple phrase grouping, namely, minor break, major break and PG break. Together with intensity specification, the make-up of the rhythm and tempo of fluent speech prosody could be constructed. We have incorporated this feature it into our prosody framework and differs our framework from the commonly adopted 2 degrees of boundary breaks, usually referred to as minor and major breaks.

2.5. Summary

From the evidences presented in this section, we argue that a prosody organization of fluent connected speech should accommodate discourse effects 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 and intensity distribution with reference to overall cross-phrase relationships merit reconsideration. Boundary breaks also require further understanding. From the above evidence of syllable cadence templates derived, it is quite evident that cross-phrase duration patterns with respect to 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 fluent speech prosody, reflecting also 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 forecasting apparatus for processing in 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, look-ahead, forecasting, matching, and filtering could also be built into fluent speech recognition as well.

3. Perceptual roles of F_0 patterns and phrasal intonation

In this section, we report perceptual investigation on the role of phrasal intonation in the organization of speech prosody. By our account in Section 2, only PG-initial and PG-final phrases possess distinct F_0 patterns, with the PG-final phrase corresponding best to phrasal intonation defined by sentence types. Whereas the PG-medial phrases are required to be held flatter towards each boundary to withhold the non-terminal effect. In other words, all phrasal intonations under phrase grouping undergo modifications and as a result some of them would lose their intonation identities. The goal in this section is to see whether these PG-final phrasal intonations, the best preserved intonations in phrase groups, are consistently identifiable. If so, whether they are identified by overall F_0 contour patterns as their roles in languages like English, or by other features instead. Moreover, whether there exists a default intonation for Mandarin Chinese, and whether the role of phrasal intonations, default or otherwise, is as significant as the literature suggests.

Three perception tests were performed to test the following hypotheses, namely, (1.) phrasal intonations exist in Mandarin Chinese, but do not play as much a role as they do in non-tonal languages. (2.) The utterance-final syllable question particles play a more significant role than overall intonation contour patterns in Mandarin Chinese. (3.) Phrasal F_0 contours lose their intonation characteristics when the final syllable is removed irrespective of their POS, thereby further shows the less significant role of overall intonation pattern at the phrase/sentence level. And (4.) default intonation for Mandarin is the declarative. We will present three experiments below.

3.1. Perception Experiments

Three auditory perception experiments were conducted to test the above hypothesis.

3.1.1. Experiment 1

3.1.1.1. Methodology

10 PG samples of male microphone read speech were chosen from a speech database of 599 read discourses collected in sound proof rooms. These PG samples ranged from 8 to 24 characters/syllables (or approximately 1 to 6 secs) in duration. All of the chosen PGs ended in yes-no questions without phrase-final mono-syllable question particles. Among the 10 speech samples, 5 ended in 2-syllable PWs; the other 5 in 3-syllable PWs. Backward editing of these PGs was performed, removing the last one, last two and last three syllables of the PG respectively. A total of 40 PGs were generated. Using the PRAAT software, the segmental information of these 40 PGs were removed and then replaced by humming while the overall F_0 patterns were extracted and retained. A total of 40 humming tokens were created to serve as stimuli of Experiment 1. Four repetitions of the tokens were randomized, making up a total of 160 test tokens of the experiments.

3.1.1.2. Subjects.

4 subjects, 1 male and 3 female, participated in Experiment 1. All of the subjects were college educated native speakers of Mandarin Chinese spoken in Taiwan with no hearing impairment.

3.1.1.3. Procedures.

Perception identification tests were administered in sound proof rooms over headsets. Each subject received different randomization results. Subjects were asked to identify if they heard yes-no question intonation.

3.1.2. Experiment 2

3.1.2.1. Methodology

For Experiment 2, 20 PG samples of male microphone read speech from the same speech data base were chosen. 10 of the PG samples were the same samples from Experiment 1, namely, PGs end in yes-no questions without phrase-final mono-syllable question particles. Another 10 PGs were samples that ended in declarative phrases. These declarative-ending PG samples ranged from 10 to 24 characters/syllables (or approximately 2.5 to 6 secs) in duration. Among the 10 declarative speech samples, 8 ended in 2-syllable PWs; 2 in 3-syllable PWs. The same backward editing of these PGs was performed, removing the last one, last two and last three syllables of the PG respectively. A total of 80 PGs were generated. Using the PRAAT software to remove segmental information but retaining overall F_0 patterns, a total of 80 humming tokens were created to serve as stimuli of Experiment 2. Four repetitions of the tokens were randomized, making up a total of 320 test tokens of the experiments.

3.1.2.2. Subjects.

The same 4 subjects participated in Experiment 2 on a different day.

3.1.2.3. Procedures.

The same perceptual identification tests were administered in sound proof rooms over headsets. Each subject received different randomization results. Subjects were asked to identify if they heard declarative intonation.

3.1.3. Experiment 3

3.1.3.1. Methodology

For Experiment 3, 30 PG samples of male microphone read speech from the same speech database were chosen. 10 more PG samples were added to the samples chosen for Experiment 2. That is, in addition to 10 PGs ended in yes-no questions without phrase-final mono-syllable question particles and 10 PGs ended in declarative phrases, another 10 PGs of yes-no questions with phrase-final mono-syllable question particles were chosen. These last 10 question-ending PG samples ranged from 9 to 23 characters/syllables (or approximately 2.2 to 5.7 secs) in duration. Two of these 10 yes-no questions ended in 2-syllable PWs; three in 3-syllable PWs. The same backward editing of these PGs was performed, removing the last one, last two and last three syllables of the PG respectively. A total of 120 PGs were generated. Using the PRAAT software to remove segmental information but retaining overall F_0 patterns, a total of 120 humming tokens were created to serve as stimuli of Experiment 3. Four repetitions of the tokens were randomized, making up a total of 480 test tokens of the experiments.

3.1.3.2. Subjects.

The same 4 subjects participated in Experiment 3 on a different day.

3.1.3.3. Procedures.

The same perceptual identification tests were administered in sound proof rooms over headsets. Each subject received different randomization results. Subjects were asked to identify if they heard declarative intonation.

3.2. Results.

The following tables summarize results of correct percentage of identification by subjects of the above 3 perceptual identification experiments; the accompanying figures are plotted display of the same results. Correct identification is defined as follows: For both yes-no questions with and without utterance final question particle, only the complete utterance intonation is defined as question intonation, except for Experiment 1. All edited tokens were treated as declarative intonation by default. Table 6 show the results of Experiment 3.1.1, i.e., perceptual identification of humming of yes-no question intonation without question particle. Figure 14 plotted the same results.

Table 6. Correct identification rates of yes-no questions without question particles. Ss: subjects, S1 to S4 represent subjects 1 through 4. A: Tokens of full PG, B: Tokens without last syllable, C: Tokens without last 2 syllables, D: Tokens without last 3 syllables

Ss	A	B	C	D
S1	44.4%	25.0%	17.1%	35.9%
S2	78.9%	44.7%	14.3%	35.9%
S3	70.3%	35.9%	16.2%	45.9%
S4	70.3%	42.1%	35.1%	63.2%
Avg	65.4%	36.3%	18.8%	47.1%

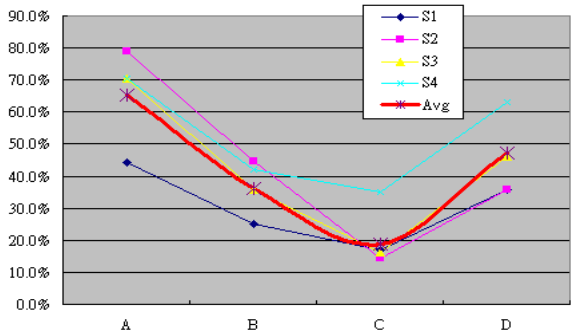


Figure13. Correct identification rates of yes-no questions without question particles by speakers.

Results from Experiment 1 shows that correct identification was best when the entire PG contour was presented. Identification begins to decay to below chance at 1 syllable edited off from the terminal end, and is the worst when 2 syllables were edited off. However, note that identification improved when 3 syllables were edited off, but is till at chance level. Since we balanced the number of PW syllables at PG-final positions, we could not offer any explanation at this point. Nevertheless,

since the test tokens ranged from 8 to 24 syllables, the results suggest that the overall F_0 contour of the final PPh is not as significant.

Table 8 show the results of Experiment 2, i.e., perceptual identification of humming of declarative intonation as well as yes-no question intonation without question particle. Figure 15 plotted the same results.

Table 8. Correct identification of declarative vs. yes-no questions without question particles. Ss: subjects, S1 to S4 represent subjects 1 through 4. A: Tokens of full PG, B: Tokens without last syllable C: Tokens without last 2 syllables, D: Tokens without last 3 syllables.

Ss	A	B	C	D
S1	75.0%	66.3%	60.0%	66.3%
S2	65.0%	61.3%	36.3%	47.5%
S3	60.0%	61.3%	50.0%	60.0%
S4	58.8%	56.3%	46.3%	58.8%
Avg	64.7%	61.3%	48.1%	58.1%

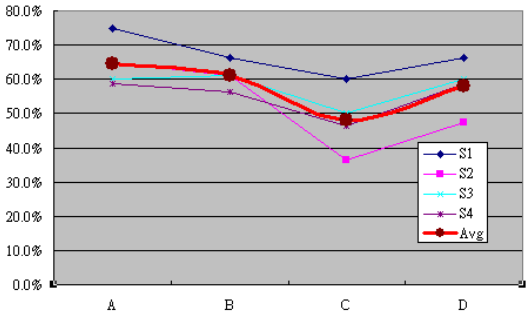


Figure15. Correct identification of declarative vs. yes-no questions without question particles by speakers.

Table 9 show the results of Experiment 3, perceptual identification of humming of declarative intonation, yes-no question intonation without question particle, and yes-no question with question particle. Figure 16 plotted the same results.

Table 9. Correct identification of declarative utterances vs. yes-no questions vs. yes-no questions with question particles. Ss: subjects, S1 to S4 represent subjects 1 through 4. A: Tokens of full PG, B: Tokens without last syllable C: Tokens without last 2 syllables, D: Tokens without last 3 syllables

Ss	A	B	C	D
S1	66.7%	65.8%	55.0%	61.7%

S2	57.5%	43.3%	32.5%	41.7%
S3	71.7%	43.3%	39.2%	42.5%
S4	70.8%	40.0%	38.3%	43.3%
Avg	66.7%	48.1%	41.3%	47.3%

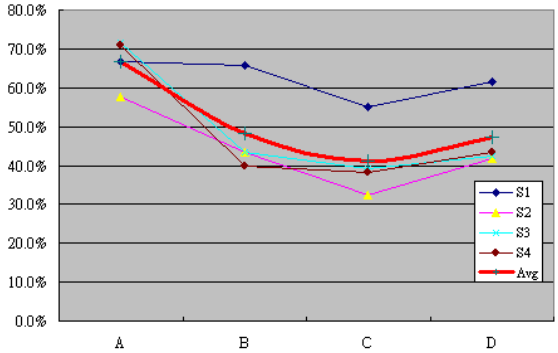


Figure16. Correct identification of declarative utterances vs. yes-no questions vs. yes-no questions with question particles by speakers.

3.3. Discussion

For Mandarin Chinese, perceptual results of humming F_0 contours indicated the following: (1.) the role of utterance final syllable was crucial to correct identification of intonations instead of the intonation patterns; overall phrasal or sentential intonation contour pattern was relatively less significant even at the best preserved positions. (2.) The syllable-cadence template and intensity patterns (Sections 2.2 and 2.3) also depend crucially on the final syllable to maintain, suggesting that impaired rhythm also impaired correct identification. (3.) Edited yes-no questions with or without sentence final question particles were identified as declarative intonation, indicating that declarative is the default intonation. (4.) Compared with the final syllable, the general higher register exhibited in yes-no questions without utterance final question particles was not the most salient cue to signal question intonation, suggesting that listeners did not pay much attention to register height or intonation in general. In summary, the perceptual results suggest that phrasal intonations and overall register height of tone languages may not as be nearly as significant as their counterpart roles in intonation languages.

The reasons are not only based on the statistics shown in this section, but also from reports from the subjects. Subjects reported how difficult the tests were, and what strategies they used during testing. Two strategies were used for identification of question intonation across subjects. First, they reported that a large number of intonations ended abruptly, making it impossible for them to associate with questions. In other words, it didn't end right. Second, they looked for final lengthening as an indicator for question intonation. In other words, the ending wasn't long enough. The first strategy suggests that abruptness overrides overall intonation pattern, and both strategies focused on how each test token ended rather than how it proceeded over time. Both strategies also imply that subjects

somehow associated question intonation with an ending duration patterns instead of F_0 contours, thereby suggesting that the role of speech tempo patterns are as important as intonation contours, if not more. These reports are particularly important. The subjects were in fact telling us what they were looking for during these experiments. It was how the humming ended, not how they went over time. Choosing the default intonation did not necessarily mean they considered the intonation default. From the viewpoint of our framework, these reports were hardly surprising.

The above results contradict with most other perceptual studies on Mandarin intonation, both in overall contours and global question. We note that almost all of documented perceptual studies echoed the existence of universal phrasal intonation by syntactic types [Ho, 1976; Shen, 1985; Chang, 1998; Lin, 2002; Yuan, 2004], and stressed on how lexical tone and intonation interact [Shih, 1988; 2004; Yuan, 2004]. However, note that all of these studies employed relatively short sentences produced as discreet units. In other words, all of the utterances in these studies were produced and perceived WITHOUT context. Emphases on the interaction between tone and intonation also assumed that the lexical tone of each syllable was always produced with distinct patterns, a similar assumption as with intonation patterns. Nevertheless, note that when producing fluent speech a speaker is equipped with other available linguistic knowledge and alternatives linguistic resources than intonation alone to convey meaning, just as strategies available to the listener to process speech signals are many layered as well. The linguistic knowledge of the speaker results in many and various forms of missing information in speech production, and the speech signals produced may very well be incomplete or distorted. This is particularly the case with spontaneous speech. The same or similar knowledge is used in processing to reconstruct the distorted signals to successfully derived meaning intended. So the question is what kind of cue the listener is looking for to process speech prosody, whether it is the entire contour, the overall tendency or the characteristics associated with the very end. When short utterances were produced as unrelated units one at a time, that is, without adjacent sister phrases to help supply contextual information; the information load of intonation increases and hence the best or least distorted form produced. Whereas when producing a succession of utterances to form narratives and/or discourses, the respective individual identity of each and every phrase is reduced while a different overall effect achieved, and the listener may very well be looking for cues other than the overall tendency. By analogy, solo singing is distinctly different from chorus singing. The former stresses individual interpretation without a conductor's baton, while the latter team works and harmony with everyone's attention on the conductor's baton. Researches in unlimited TTS have long demanded the speech community to come up with more systematic account of the choral aspect of fluent speech production. Approaching phrases one at a time would be like responding to questions of choral singing with solos only.

4. Modeling Mandarin fluent speech prosody

Based on the prosody organization discussed in Section 2 and the perceptual evidence, the hierarchical PG structure of fluent speech was adopted to model the speech prosody of multiple phrase groups. To further test both the framework and the model, we have also implemented a Mandarin TTS system using a syllable-token database at this stage (See Section 5). Since a speech database of PWs is still under construction and manual annotation time consuming, and since there are only 1292 distinct tonal syllables, we, too, choose syllables as the concatenate units to test the model for the present. Needless to say, we could test the model as soon as our database of prosodic units is constructed and annotated. However, our syllables are collected in accordance with our framework feature. We designed a 29-syllable 3-phrase complex carrier sentence to record target syllable tokens in order to

solicit the same syllable produced in three distinct PG-related positions. An example is shown in Figure 17. Each target syllable was embedded in the carrier sentence at the initial, medial, and final positions, respectively. Therefore, the database consists of 1292*3 Mandarin tonal syllable tokens.

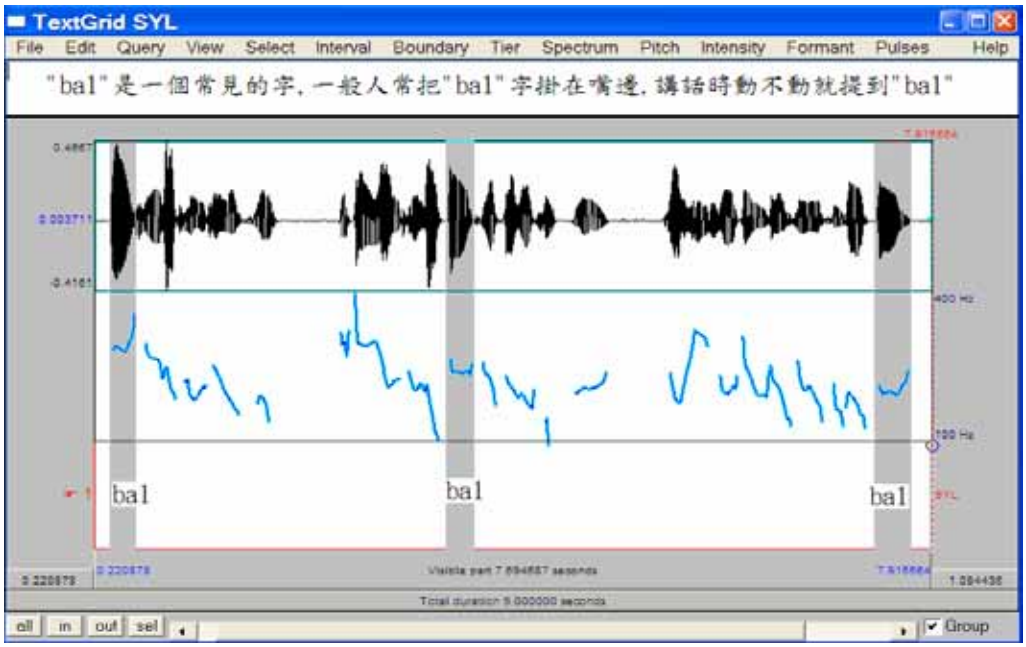


Figure17. The waveform and F_0 tracking of a carrier sentence with three target syllables "ba1" embedded in three darkened positions. ("ba1" is a frequently used syllable, people say "ba1" very often, often times when people speak, they'd use "ba1"). The target syllable "ba1" occurred in three PG positions, namely, PG-initial, PG-medial and PG-final to provide PG related information.

Our model adopted a modular approach to model F_0 contours, duration patterns, intensity patterns and break predictions in separate modules. Since temporal allocations and rhythmic structure in speech flow are carefully dealt with in addition to F_0 patterns, the TTS system is capable of converting long paragraph of text input into more natural synthesized speech output.

4.1. F_0 modeling

There are many existing F_0 models of sentence/phrasal intonation around. In fact, our framework could adopt any F_0 model at the PPh level and further adjust each respective F_0 contour pattern with specifications from higher node(s) to generate multiple phrase F_0 output. We adopt the well-known Fujisaki model as the production model of F_0 [Fujisaki & Hirose, 1984; Fujisaki, 2002. The model connects the movements of cricoid's cartilage to the measurements of F_0 and is hence based on constraints of human physiology. Therefore, it is reasonable to assume that the model could accommodate F_0 output of different languages. In fact, successful applications of the model on many language platforms have been reported, including Mandarin [Mixdorff et al, 2003; Mixdorff, 2004].

In the case of Mandarin Chinese, phrase commands were used to produce intonation at the phrase level while accent commands were used to produce lexical tones at the syllable level [Mixdorff, 2000]. Phrasal intonations are superimposed on sequences of lexical tones. Therefore, interactions between the two layers cause modifications of F_0 to produce the final output. The superimposing of a higher level onto a lower level leaves room for even higher level(s) of F_0 specification to be superimposed and built. Thus, we decided to implement our PG framework of phrase/intonation-grouping on the Fujisaki model by adding a PG layer over phrases. In other words, after generating phrasal intonations for each phrase, PG specifications were then superimposed onto phrase strings subsequently. By adding one higher level of PG specification, the F_0 patterns of phrase grouping could be achieved.

4.1.1. Building the phrasal intonation model

The corpus used for training the F_0 model was a female read speech data of 26 long paragraphs or discourses in text, or a total 11592 syllables (or Chinese characters). 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). We first proceeded with automatic parameter extraction, and then used the extraction results to build a statistical phrasal intonation model. A linear model is adopted for the Fujisaki model's phrase command Ap:

Ap = constant + coeff1 ⨯ Pause length before phrase command
+ coeff2 ⨯ Accumulated previous phrase command response
+ coeff3 ⨯ F_0min in the Fujisaki model
+ f(Phrase command position in PPh) (1)

where pause is the speechless portion in relation to a following phrase command, accumulated previous phrase command response is the accumulated response of previous phrase commands at the response of the current phrase command reaches to its peak, F_0min is the minimum fundamental frequency of the utterance, and f (phrase command position in PPh) is a function of the PW position in PPh where the PW is related to the phrase command. The accumulated response of previous phrase commands at time t is calculated as:

$$AccF0 = \sum_{prev\ \mathrm{Ap}} A_p \cdot \alpha^2 \cdot (t - T_{0i}) \cdot e^{(-\alpha(t-T_0))} \qquad (2)$$

AccF0 could then represent previous accumulated intonation due to equation 2.

4.1.2. Building the PG intonation model

As discussed in Tseng et al 2004a, the PG intonation has significant effects in the first and last PPh units only. Therefore, the parameter Ap in equation (1) in the intonation model can be modified as:

Ap = constant + coeff1 ⨯ Pause length before phrase command
+ coeff2 ⨯ Accumulated previous phrase command response
+ coeff3 ⨯ F_0min in the Fujisaki model
+ f(Phrase command position in PPh)
+ f(Phrase position in PG(Initial, Medial, Final)) (3)

Thus we considered the prediction of Ap in a layered perspective. Indvidual prosodic phrases are using the phrasal intonation model and the global effects are superimposed onto the phrasal intonation model in the last term of equation (3).

4.1.3. Application to the TTS system

The constructed PG intonation model can be applied to our Mandarin TTS system to produce F_0 contours. Since the higher level of prosodic unit is taken into account, more fluent and natural intonation can be obtained. The details of adjusting the F_0 output will be described in Section 5.2.

4.2. Duration modeling

Our duration model of the rhythmic patterns in Mandarin speech flow (Section 2.2.) reveals that the syllable duration is not only affected by the syllable constitution itself, but also affected by the upper layer prosodic structures, namely PW, PPh, BG, and PG, respectively.

The same speech database used for training the F_0 model was used for training the duration model.

4.2.1. Intrinsic statistics of syllable duration

A layered model is used to estimate the syllable's duration. At the SYL-layer, a linear model is adopted:

Syllable intrinsic duration = constant + CTy + VTy + Ton + PCTy + PVTy + PTon + FCTy + FVTy + FTon
+ 2-way factors of the above factor
+ 3-way factors of the above factor (4)

The constant was set to 185 ms, which was dependent on the corpus. CTy, VTy, and Ton represent the offset values corresponding to the consonant type, vowel type and tone of the current syllable, respectively. Prefix P and F represent the corresponding factors of the preceding and following syllables respectively. The 2-way factors consider the joint effect of two single-type factors. There are C_2^9(=36) 2-way factors in total. The 3-way factors consider the joint effect of three single-type factors. The 3-way factors with a negligible influence on the syllable duration were excluded from consideration. Only three 3-way factors were left, they are the combination of consonant type, vowel type and tone of the preceding, current, and following syllables, respectively. As a result, a total of 49 factors were considered. The 21 consonants and 39 vowels (including diphthongs) of Mandarin were, respectively, grouped into 7 and 9 categories according to their measured mean duration. Notice that the SYL-layer model is independent of the prosodic structure. The SYL-layer model can explain about 60% of syllable duration.

4.2.2. The effect of layered prosodic structure

As depicted in Figures 3, the syllable duration is affected by its position within a PW. Note that the PW final syllable tends to be lengthened compared to other syllables. The residue error that can not be explained at the SYL-layer can be further explained by the PW-layer. Accordingly, the syllable duration is postulated as:

DurS (ms) = Syllable intrinsic duration + f_{PW}(PW length, position in PW). (5)

Since the syllable intrinsic duration is the duration controlled by the SYL-layer, the PW-layer has its effect of speeding the rhythm by subtracting a value derived from Figure 1 and vice versa.

The PPh-layer affects the syllable duration in a similar way as the PW-layer (shown in Figure 4). As to the BG-layer or above, the length of the prosodic unit gets longer and complicated, the perceived significance exists only in the initial and final PPh units. Therefore, we model PG-layer's effect as the effect in the initial and final PPhs in the PG-layer. The overall model is thus formulated as:

DurS (ms) = Syllable intrinsic duration
+ f_{PW}(PW length, position in PW)
+ f_{PPh}(PPh length, position in PPh)
+ f_{IFPPh}(Initial/Final PPh length, position in PPh) (6)

Where DurS means the modeled syllable duration, fPW, fPPh, fIFPPh mean the portions of syllable duration which are affected by the function of length and position of PW, PPh and PG respectively.

4.2.3. Application to the TTS system

The results of these duration patterns are not only evidence of interaction between syllable duration adjustment and prosodic level units, but also a useful duration prediction method. Therefore, temporal allocation is implemented in our TTS system. The details will be described in Section 5.3.

4.3. Intensity modeling

Segmental RMS values were first derived using an ESPS toolkit. For each initial and final phone in syllable, 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 intensity modeling is much the same way like modeling in durations:

IntS (NRMS) = Syllable intrinsic intensity
+ f_{PW}(PW length, position in PW)
+ f_{PPh}(PPh length, position in PPh)
+ f_{IFPPh}(Initial/Final PPh length, position in PPh) (7)

Where IntS means normalized average syllable intensity, rms value, fPW, fPPh, fIFPPh mean the portions of syllable intensity which are affected by the function of length and position of PW, PPh and PG respectively.

The TTS corpus is designed as carrier sentence, which the initial, medial, and final syllables have fixed preceding and following syllable. The absolute intensity predicted by the intensity model should be adjusted, while the stress pattern in the PG organization should be kept.

5. The TTS system

5.1. Speech database

Both the duration and F_0 models described above are built based on the PG structure. Therefore, we have specially designed our database such that the TTS system can be implemented to use these models.

The database is made of 1292*3 Mandarin tonal syllable tokens. Each of the 1292 syllables was embedded in a phrase of a 3-phrase carrier sentence (a PG of 3 PPhs) in initial, medial, and final positions, respectively (see Figure 16 with syllable "ba1" embedded in it show the associated wave, the F_0 contour, and the time stamps of the target syllable "ba1".) The speech data were recorded by a native female speaker in a sound-proof room. The target syllable tokens were listened to and manually edited from the carrier sentence by trained transcribers. In our TTS system, the time-domain pitch-synchronous overlap-add (TD-PSOLA) [Carpentier & Stella, 1986] method is employed to perform prosody modification. The pitch marks were first automatically estimated, and then manually repaired by trained transcribers.

For each syllable, there are 3 tokens, respectively, collected from the initial, medial, and final positions of a PG. Since the prosodic models were trained using a different speaker's speech, the models need to be adapted to satisfy the condition indicated by the initial, medial, and final syllables of a PG to be synthesized. In other words, the TTS system will only adjust the duration and F_0 of the other syllables using the modified prosodic models but keep those of these 3 syllables unchanged.

5.2. F_0 adjustment

The speech intonation of our TTS system is predicted by the Fujisaki model in our PG framework. The reason why we have to adjust the predicted output is because it's a redundant process to alter the intonation of the target syllable which has already been in the correct position. Since the target syllables are having their own intonation embedded in original carrier sentences, we have to level up or down the predicted results according to the difference between them.

In the implementation of adjustment, the comparison is confined between the first F_0 peak of predicted PG intonation and the average F_0 of the first syllable from the carrier sentence. Based on the equation of the Fujisaki model's phrase commands:

$$G_p(t) = \begin{cases} = \alpha^2 t \cdot \exp(-\alpha t), & \text{for } t \geq 0 \\ = 0, & \text{for } t < 0 \end{cases} \tag{8}$$

In equation (8), the time to reach its maximum is $1/\alpha$, since the maximum phrase value, say P, will be:

$$P = Ap \times \alpha \times \exp^{(-1)} \tag{9}$$

where $1/\alpha$ is substituted into t in equation (8). In equation (9), we found P is proportional to Ap, while α remains constant.

The difference between average F_0, denoted as P_c, of the first syllable form the carrier sentence and the first F_0 peak, denoted as P_p, of predicted PG intonation results the adjustment of the predicted Ap:

$$\Delta Ap = \hat{A}p - Ap = (P_c - P_p) \times \exp \times \alpha^{-1} \tag{10}$$

Where $\hat{A}p$ is the value after adjustment, and Ap is the value of original prediction. Thus every phrase command has to adjust to its new value according to ΔAp. After the adjustment, the shape of intonation is not changed but the level of it is changed according to the carrier sentence database.

5.3. Duration adjustment

Since the TTS database was from a different speaker, the absolute duration predicted by the duration model should be adjusted, while the rhythmic patterns in the PG organization should be kept.

Because the initial, medial, and final syllables are originally collected from the same positions of a PG, their durations should not be changed. The durations of the rest syllables, which were originally the first syllable of a PW at the medial position of a medial PPh of a 3-PPh PG, should be modified to satisfy the rhythmic pattern in the PG organization. In this way, to synthesis a PG of m characters (or syllables), the duration of the i-th syllable is given by

$$DurS_i^* = \begin{cases} OriDur(S_i) & , i = 1, m/2, m \\ OriDur(S_i) - DF_i & , 1 < i < m/2, m/2 < i < m, \end{cases} \tag{11}$$

where $OriDur(S_i)$ is the corresponding syllable-token's original duration and DF_i is an offset factor, which is calculated by

$$\begin{aligned} & DF_i \\ & = M_{TC} / M_{MC} \times [f_{PW}(PW\ length, position\ in\ PW) - f_{PW}(2,1) \\ & + f_{PPh}(PPh\ length, position\ in\ PPh) - f_{PPh}(11,6) \\ & + f_{IFPPh}(Initial / Final\ PPh\ length, position\ in\ PPh)], \end{aligned} \tag{12}$$

where M_{TC} and M_{MC} are, respectively, the mean of syllable duration of the TTS corpus and the training corpus, and $f_{PW}(\ )$, $f_{PPh}(\ )$ and $f_{IFPPh}(\ )$ are the same as that in equation (6), which were calculated from the training corpus.

5.4 Intensity adjustment

Because the initial, medial, and final syllables in TTS corpus keep the characteristic in a PG, their intensity should not be modified while they are initial, medial, and final syllable of synthesized utterance. According our unit selection method, the intensity of rest syllables, which were originally the first syllable of a PW at the medial position of a medial PPh of a 3-PPh PG, in the synthesized utterance, should be changed to satisfy the stress pattern in the PG organization. In this principle, if m characters (or syllables) need to be synthesized, the intensity of the i-th syllable is given by

$$IntS_i^* = \begin{cases} OriInt(S_i) & , i = 1, m/2, m \\ OriInt(S_i) - DF_i & , 1 < i < m/2, m/2 < i < m, \end{cases} \tag{13}$$

where OriInt(S_i) is the corresponding syllable-token's original intensity and DF_i is an offset factor, where is calculated by

$$\begin{aligned} & DFi \\ & = M_{TC} / M_{MC} \times \left[f_{PW}(PW\ length, position\ in\ PW) - f_{PW}(2,1) \right. \\ & + f_{PPh}(PPh\ length, position\ in\ PPh) - f_{PPh}(11,6) \\ & \left. + f_{IFPPh}(Initial / Final\ PPh\ length, position\ in\ PPh) \right], \end{aligned} \tag{14}$$

where M_{TC} and M_{MC} are, respective, the mean of syllable intensity of the TTS corpus and the training corpus, and $f_{PW}()$, $f_{PPh}()$ and $f_{IFPPh}()$ are the same as that in Eq. 5, which were calculated from the training corpus

5.5. Break prediction

The prosodic boundaries and break indices are predicted by analyzing the syntactic structure of the text to be synthesized (Chen et al, 2004). As discussed in Section 2.4., 3 levels of breaks relative to speaking rate are incorporated into our model to accommodate multiple phrase grouping.

5.6. System flowchart

Given a piece of text, first of all, the prosodic boundaries and break indices will be predicted based on the analysis of syntactic structure. The PG hierarchical structure and the pronunciations (the syllable sequence associated with the text) will be generated as well. Then, the durations of all syllables will be assigned by the duration model, while the F_0 contours of all phrases will be generated by the intonation model. All the outputs of text processing will be stored in a predefined XML document. Finally, the TD-PSOLA method is employed to perform prosody modification, and the TTS system will output the concatenate waveform.

5.7. Discussion

Our TTS system aims at synthesizing fluent speech in long paragraphs. Because long speech paragraphs are perceived with its significant initial and final PPhs, modeling this phenomenon will signal output topics clearer in multiple phrases groups and to avoid a succession of short and choppy phrases. The duration model was clear in each layer, thus a straightforward linear model was sufficient to model durational effect of every prosodic unit. The F_0 model based on the Fujisaki model is more complicated but we used the extensible ability of the Fujisaki model to extend the F_0 model to the overall intonation of PG. We argue from collective evidences that a prosody framework of multiple phrase grouping could better account for the make-up of fluent speech prosody.

6. Conclusions

Up to this point, research seeking to describe and predict Mandarin Chinese prosody has focused mostly on the intonation of phrases or sentences in isolation, but it remains to be seen how these effects interact with higher prosodic levels in fluent speech materials. These studies have yielded

detailed information about intonation in sentences of 10 syllables or less, which were produced in isolation, under the tacit assumption that fluent speech would be a concatenated version of such sentences.

This paper has demonstrated, hopefully in a nutshell, that one of the most important prosodic characteristics of fluent Mandarin Chinese speech cannot be seen at the level of single-sentence intonations, but rather, reveals itself only in the examination of larger chunks of fluent speech. The operating unit essential to the execution of fluent Mandarin speech is a higher-level unit, which combines individual phrase and sentence intonations into the governing prosodic group (PG), often manifested as multi-phrase speech paragraphs in narration and/or spoken discourse. Consequently, subjacent phrase- or sentence-intonation contours are often less significant and should be seen as sister prosodic constituents within a canonical prosodic form, whose characteristics and manifestation will vary according to their relative positions within the PG.

The present study demonstrates that to account for fluent speech prosody, discourse effect must be included and multi-phrase units must be examined. More research efforts irrespective of language under investigation should be directed to the following: (1.) larger units for overall speech output planning must be taken into consideration. (2.) Cross-phrase prosodic relationship need to be specified. (3.) Patterns of temporal allocation, speech rate, and tempo/rhythm are also related to prosodic organization, and their characteristics can be predicted by examining higher-level prosodic patterns. (4.) How a speech paragraph starts and ends, an in particular, the ending patterns, are crucial in fluent speech. Specific to Mandarin Chinese and perhaps other tone languages is that phrasal intonations are not as significant as they are in intonation languages.

Moreover, the perception motivated multi-phrase PG model offers at least in part a knowledge base and viable framework for formulating theories of prosodic organization in other syllable-timed languages. We presented evidence to show that more understanding of the prosodic structure of fluent speech is essential, and how templates of prosody-related pitch, cadence, and boundary patterns may together account for cross-phrase look-ahead in fluent speech. As for technological and computational applications, we have also illustrates in Section 4 how mono-syllables could be collected to offer more prosody information, in compliment with speech database of prosodic units at the same time. We have also implemented an initial version of this framework into current TTS system because it is our belief that identifying and simulating speech paragraphs are the key to solve output naturalness for TTS. An integrated prosodic model that organizes phrase groups into related prosodic units to form speech paragraphs will be instrumental to improve output naturalness for unlimited TTS. The implications and applications are without doubt not language specific to Chinese only. We believe our model should fit in nicely the needs for any concatenate TTS system, and may be adapted to constructing canonical complex sentence intonation for other languages as well.

Future works include expanding the framework to accommodate focus and prominence in relation to F_0 range, investigating boundary breaks in relation to perceived pitch resets in more detail, building the TTS system on larger amount of more varied speech data and prosodic units, and using the model with synthesis as tools for perception studies that aims at establishing concrete measures for output naturalness.

7. References

Charpentier, M.J. and M.G. Stella, “Diphone Synthesis using an Overlap-Add Technique for Speech Waveforms Concatenation”, Proceeding of ICASSP86, pp. 2015-2018.

Chang, L. and Chen, K. “The CKIP Part-of-speech Tagging System for Modern Chinese Texts.” Proceedings of the 1995 International Conference on Computer Processing of Oriental Languages (ICCPOL), 1995, Hawaii, pp. 172-175 or *http://rocling.iis.sinica.edu.tw/CKIP/*

Chang, Y. 1998 “les indices acoustiques et perceptifs des questions totales en Mandarin parle de Taiwan”, Cahiers de Linguistique Asie Orientale, 1998, pp. 51-78.

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

Chen, K., C. Tseng, H. Peng, and C. Chen, “Predicting Prosodic Words from Lexical Words - A First Step towards Predicting Prosody from Text”, Proceedings of the International Symposium on Chinese Spoken Language Processing 2004, December 15-18, Hong Kong, pp. 173-176.

Fujisaki, H. and K. Hirose, 1984 “Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese”, Journal of the Acoustical Society of Japan (E), 5(4): pp. 233-241.

Fujisaki, H. 2002 “Modeling in the Study of Tonal Feature of Speech with Application to Multilingual Speech Synthesis”, Proceedings of SNLP-O-COCOSDA 2002. Hua Hin, Thailand, D-1

Gussenhove, C. “Types of Focus in English?” In Daniel Buring, Matthew Gordon & Chungming Lee (eds.) Topic and Focus: Intonation and Meaning: Theoretical and Crosslinguistic Perspectives. Dordrecht: Kluwer.

Ho, A.-T. 1976 Mandarin tones in relation to sentence intonation and grammatical structure”, Journal of Chinese Linguistics, 4, 1976, pp. 1-13.

Keller. E., and B. Zellner Keller, “A Timing model for Fast French”, York Papers in Linguistics, 17, University of York. 53-75. 1996

Lin, M-C, 2002 “Hanyu yunlyu jiegou han gongneng yudiao (Mandarin prosody organization and functional intonations, in Chinese)”, Report of Phonetic Research 2002, Phonetics Laboratory, Institute of Linguistics, Chinese Academy of Social Sciences pp. 7-23.

Mixdorff, H. “A Novel Approach to the Fully Automatic Extraction of Fujisaki Model Parameters”, Proceedings of ICASSP2000, pp. 1281-1284.

Mixdorff, H., Y. Hu, and G. Chen, “Towards the Automatic Extraction of Fujisaki Model Parameters for Mandarin”, Proceedings of Eurospeech 2003, September 1-4 , 2003, Geneva, Switzerland, pp. 873-876.

Mixdorff, H. 2004 “Quantitative Tone and Intonation Modeling across Languages”, Proceedings of International Symposium on Tonal Aspects of Languages- with Emphasis on Tone Languages (TAL 2004), pp. 137-142.

Pin, S., Y. Lee, Y. Chen, H. Wang and C. Tseng, “A Mandarin TTS System with an Integrated Prosodic Model.” Proceedings of the 2004 International Symposium on Chinese Spoken Language Processing (ISCSLP-2004), December 15-18, 2004, Hong Kong, pp. 169-172.

Selkirk, E. “Derived Domains in Sentence Phonology”, Phonology Yearbook 3: 371-405.

Shattuck-Hufnagel, S. & A. Turk, 1996 “A Prosody Tutorial for Investigators of Auditory Sentence Processing” Journal of Psycholinguist Research 25(2): 193.

Shen, J. 1985 “Beijinghua shengdiao de yinyu he yudiao (Pitch range of tone and intonation in Beijing dialect, in Chinese)”, in Lin T. and Wang L eds. Beijing Yuyin Shiyanlu (Working Papers in Experimental Phonetics), Beijing, Beijing University Press, 1985, pp. 73-130.

Shih, C. “Tone and Intonation in Mandarin” Working Papers of the Cornell Phonetics Laboratory 3, 1988, pp. 83-1009

Shih, C. “Tonal Effects on Intonation” Proceedings of International Symposium on Tonal Aspects of Languages—with Emphasis on Tonal Languages (TAL 2004), pp. 163-168.

Tseng, C. F. Chou, 1999 “A Prosodic Labeling System for Mandarin Speech Database”, Proceedings of ICPhS99, pp. 2379-238.

Tseng, C. 2002 “The prosodic status of breaks in running speech: Examination and evaluation”, Speech Prosody 2002, 11-13 April, 2002 Aix-en-Provence, France, pp. 667-670.

Tseng, C. 2003 “Towards the organization of Mandarin speech prosody: Units, boundaries and their characteristics”, XIV International Congress of Phonetics Science, Aug.1-9, 2003, Barcelona, Spain, pp.

Tseng, Chiu-yu, Cheng, Yun-ching, Lee, Wei-shan and Huang, Feng-lan (2003). “Collecting Mandarin speech databases for prosody investigations,” Proceedings of the Oriented COCOSDA 2003, (Oct. 1-3, 2003), Sentosa, Singapore.

Tseng, C, S. Pin, and Y. Lee, 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, Beijing, China, pp. 417-438.

Tseng, C. and S. Pin, 2004b “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), pp. 193-197.

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

Tseng, C. and S. Pin, 2004d “Modeling Prosody of Mandarin Chinese Fluent Speech via Phrase Grouping”, Proceedings of ICSLT-O-COCOSDA 2004. November 17-10, 2004, Delhi, India.

Xu, Y. 2002 “Articulatory constraints and tonal alignment” Speech Prosody 2002, 11-13 April, Aix-en-Provence, France, pp. 91-100.

Yuan, J. “Perception of Mandarin Intonation”. Proceedings of the 2004 International Symposium on Chinese Spoken Language Processing (ISCSLP-2004), December 15-18, 2004, Hong Kong, pp.45-48..