Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 46 (2005) 284–309
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Fluent speech prosody: Framework and modeling

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

[a] *Phonetics Lab, Institute of Linguistics, Academia Sinica, Taipei, Taiwan*
[b] *Institute of Information Science, Academia Sinica, Taipei, Taiwan*

Received 16 September 2004; received in revised form 10 March 2005; accepted 28 March 2005

Abstract

The prosody of fluent connected speech is much more complicated than concatenating individual sentence intonations into strings. 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 rather than sentence effects in fluent speech. Subsequent cross-speaker and cross-speaking-rate acoustic analyses of identified speech paragraphs revealed systematic cross-phrase prosodic patterns in every acoustic parameter, namely, F_0 contours, duration adjustment, intensity patterns, and in addition, boundary breaks. We therefore argue for a higher prosodic node that governs, constrains, and groups phrases to derive speech paragraphs. A hierarchical multi-phrase framework is constructed to account for the governing effect, with complimentary production and perceptual evidences. We show how cross-phrase F_0 and syllable duration patterns templates are derived to account for the tune and rhythm characteristic to fluent speech prosody, and argue for a prosody framework that specifies phrasal intonations as subjacent sister constituent subject to higher terms. Output fluent speech prosody is thus cumulative results of contributions from every prosodic layer. To test our framework, we further construct a modular prosody model of multiple-phrase grouping with four corresponding acoustic modules and begin testing the model with speech synthesis. To conclude, we argue that any prosody framework of fluent speech should include prosodic contributions above individual sentences in production, with considerations of its perceptual effects to on-line processing; and development of unlimited TTS could benefit most appreciably by capturing and including cross-phrase relationships in prosody modeling.

Keywords: Prosodic phrase grouping; Top–down; PG; Prosodic hierarchy; Multi-phrase; Cross-phrase; Constraints; Templates; Speech planning; Look-ahead; Global F_0 templates; Temporal allocations; Syllable duration patterns; Intensity distribution; Boundary breaks

* Corresponding author. Tel.: +886 2 27863300x222; fax: +886 2 2652 3133.
E-mail addresses: cytling@sinica.edu.tw (C. Tseng), whm@iis.sinica.edu.tw (H. Wang).

0167-6393/$ - see front matter © 2005 Published by Elsevier B.V.
doi:10.1016/j.specom.2005.03.015

1. Introduction

The prosody of fluent connected speech is much more complicated than concatenating individual sentence intonations as discreet units 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. As a result, much less attention has been paid to discourse effects reflected through 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 prosody modeling. With the syntax-specified intonation patterns best applicable to simple sentences, elaborations to accommodate complex sentences are still lacking. Perhaps inadvertently, 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. By comparison, relatively little attention has been devoted to how articulatory adjustment must also be made to reflect necessary look-ahead across phrases in fluent speech production and other physiological constraints such as breathing as well as cognitive limits in relation to speech planning.

Mandarin Chinese was no exception in terms of research paradigm and orientation. To illustrate, we note that one linguistic approach and subsequent prosody modeling used short utterances of five syllables and adopted the same physiological perspective to account for cross-syllabic look-ahead in articulatory gestures in tonal co-articulation (Xu, 2002), focusing on anticipatory effects in tone concatenation at the single-sentence level. Another linguistic approach and prosody modeling used well controlled short utterances or digit strings to study narrow focus in intonation with emphases on the interaction between tone and intonation (Shih, 1988, 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 intonations as default prosody units and analyzed speech data from bottom upward, focusing on between-syllable effects or on overall register height and tendency. The significance of these approaches not withholding, prosody studies of Mandarin Chinese have pretty much stopped short at the level of phrase and/or simple sentences. In fact, all of these findings remain yet to be tested on fluent speech data for more prosodic phenomena to be included and accounted for while much more knowledge of the structure of narratives or spoken discourse remains lacking in prosody modeling.

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, but 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 and 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 larger units of multi-phrase speech paragraphs or independent unrelated phrases and sentences. In fact, simulation of a succession of discreet and often declination intonations in unlimited TTS did not produce satisfactory fluent speech prosody, and in spite of improvements in commercial products, more systematic account of discourse effects reflected in between- and cross-phrase relationships was still called for. Similar calls may also exist in languages other than Chinese for complex sentences and discourse prosody. We believe the calls still remain largely unanswered by the linguistic community, and 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 to prosody modeling. For data collection, we began from collecting relatively large amount of speech data of read discourses instead of short or isolated sentences to better reflect planning units and cognitive constraints involved in fluent speech production. Over nine 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 from a top–down perspective by listening to the corpora by discourse instead of 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 and 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 which is essentially discourse prosody rather than sentence 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 cross-phrase prosodic specifications in addition to individual intonation patterns could quite adequately account for default fluent speech prosody. In particular, how discourse information is conveyed through prosody. 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 and 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 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 and Chou, 1999). Perceptual parsing was arrived not by syntactic structure and meaning alone, but rather, by how these speech paragraphs were heard and by what listeners look for when listening to fluent connected speech. In other words, what are some of the major perceptual features listeners use to parse spoken discourse? In addition, 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 and how prosodic words could be predicted from lexical words (Chen et al., 2004).

These identified speech paragraphs were termed prosodic phrase group (PG) by us (Tseng and Chou, 1999). It was evident to us that a framework of fluent speech prosody should include multi-phrase speech paragraphs in addition to individual phrases and sentences, and explain how speech paragraphs are formed through prosodic specifications to form the ongoing effects. 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., 2004), 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 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 and Turk, 1996; Gussenhove, 2004). We proposed (Tseng et al., 2004) 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 and threshold of discourse planning and on-line processing in the cognitive domain. Hence it is feasible to assume that 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 cognitively-conditioned look-ahead of speech planning over physiologically-conditioned look-ahead during speech production and implies cross-phrase look-ahead and anticipation can be added on top of 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), 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 tides are like, how ripples ride over waves, and how waves ride 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 and Chou, 1999) that annotated small to large boundaries with a set of five break indices (BI); i.e., B1–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 and Lee, 2004). The hierarchical governing and constraining functions of PG over phrases are illustrated schematically in Fig. 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–B5. These boundary breaks are not shown in Fig. 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.

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., 2004). 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. Fig. 2 is a schematic representation of a PG of five PPhs, separated by

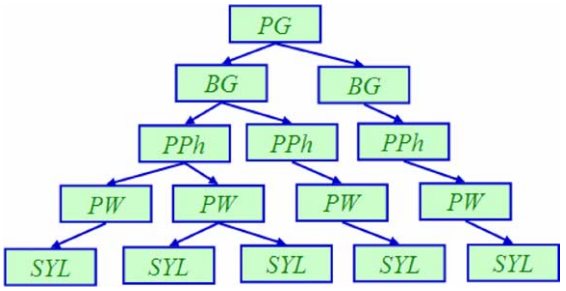

Fig. 1. A schematic representation of the hierarchical organization of multiple phrase grouping on perceived units and boundaries. Note that boundary breaks are not represented. However, the framework includes boundary breaks from B1 to B5. B1 is the perceived break between SYLs that may not correspond to a pause, B2 between PWs and from the PW layer up actual pauses, B3 between PPhs, B4 between BGs and B5 between PGs.

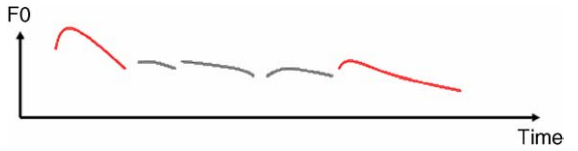

Fig. 2. Schematic illustration of the global trajectory of perceived F_0 contours of a 5-PPh PG. The units are PPhs separated by boundary breaks B3s. Note only the first and last PPhs (in red on the web) into a PG possess identifiable declarative intonations. The two declination slopes in red are significantly 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 three PPhs (in black on the web) do not possess distinct intonation patterns and are held flatter towards each of their respective ending boundaries.

major breaks B3 in the speech flow. Note that B4 usually occur only when a PG exceeds five phrases.

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 phrases. Their major function is to signal the beginning and ending of a PG rather than only their individual intonation identity. We have termed them PG-initial PPh, PG-medial PPh(s) and PG-final PPh. 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 ongoing effect. In fact, the more the number of PPhs in a PG are, the flatter the intonations of these PG-medial PPhs are held to signal the necessary ongoing effect. The flatter these PG-medial intonations are, the less distinct their identities become. However, note also how the two respective declarative intonations at PG-initial and PG-final positions are also perceived differently to signal their respective roles in PG. 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. Together, a speech paragraph is formed. 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 do not, 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, when the end is finally approaching, and how it finally ends. In short, this is why concatenating independent and distinct phrasal intonations without further modifications 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 5–12 PPhs in our corpora 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 two different speech rates, slower vs. faster speech, were used. The slower speech was recorded from one male untrained subject (hence SMS for slower male speech) reading 595 paragraphs ranging from 2 to 180 syllables; the faster speech from one female radio announcer's relatively faster reading (hence FFS for faster female speech) of 26 long paragraphs ranging from 85 to 981 syllables. Ninety percent of the two sets of text overlap. A total of 22 350 syllables of SMS and 11 592 syllables of FFS were analyzed. Average syllable duration was 304.7 ms for SMS and 199.75 ms 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 three 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 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 SYL layer upward to each of the higher prosodic units and levels, we examined each higher layer independently to see if it could account for residuals from one of the lower layers, and if so, how much was contributed by each level. All of the data were analyzed using DataDesk™ from Data Description, Inc. Two benchmark values were used in this study to evaluate the closeness of the predicted value to that of the original speech data: residual error (RE) 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 SYL 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 SYL-layer model was subsequently postulated as follows:

$$\begin{aligned} Dur\ (\text{ms}) &= \text{constant} + CTy + VTy + Ton \\ &\quad + PCty + PVty + PTon + FCty \\ &\quad + FVty + FTon \\ &\quad + \text{2-way factors of the above factor} \\ &\quad + \text{3-way factors of the above factor} \\ &\quad + \text{Delta 1} \end{aligned} \tag{1}$$

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 SYL-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 SYL layer. The residue that could not be accounted for at this layer was termed as Delta 1 and was dealt with at higher layers.

The same rationale was applied at the layer directly above the SYL 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:

$$\text{Delta } 1 = f(\text{PW length},\ \text{PW sequence}) + \text{Delta } 2 \tag{2}$$

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 three-syllable PW. The coefficient of each entry was calculated using linear regression techniques identical to those of the preceding layer. Fig. 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 five syllables were not considered, due to their under-representation in the data.

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 SYL layer. However, FFS showed the opposite: even when syllables of a PW were shortened, the final syllable maintained the duration predicted by the SYL 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.

The model was able to account for 6.7% of Delta 1 of SMS and 3.55% of FFS at the PW layer.

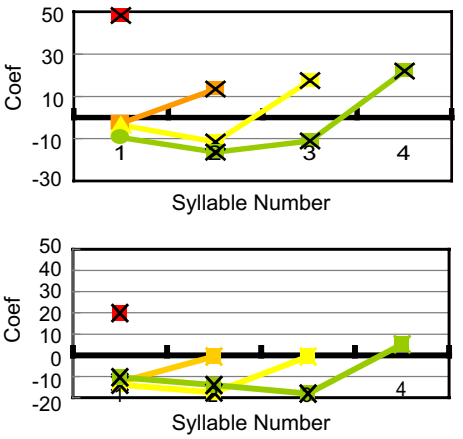

Fig. 3. Coefficients of syllable durations obtained for both speech rates using the PW model. The horizontal axis represents the position of each syllable within a PW; the vertical axis represents the coefficient values. The upper panel shows coefficients of FFS; the lower panel shows those of SMS. Positive 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.

Table 1
Evaluation of duration predictions at the SYL layer

Test	SMS	FFS
RE	48.9%	40.1%
r	0.715	0.768

Table 2
Evaluation of duration predictions at the PW layer

Test	SMS	FFS
RE	93.3%	96.45%
TRE	45.6%	38.76%
r	0.737	0.778

The overall prediction was obtained by adding up the predicted value of both the SYL 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 the PPh layer. The linear regression model is thus formulated as follows:

$$\text{Delta } 2 = f(\text{PPh length},\ \text{PPh sequence}) + \text{Delta } 3 \tag{3}$$

Fig. 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.

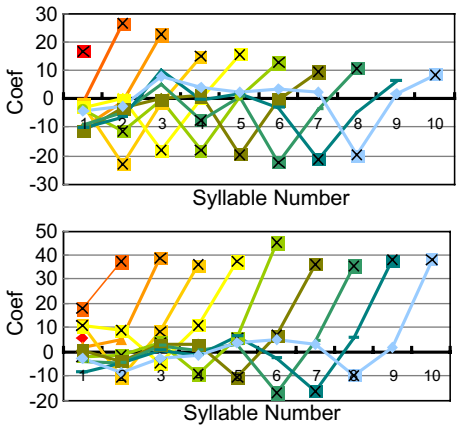

Fig. 4. Coefficients of syllable durations obtained for both speech rates from the PPh model. The horizontal axis represents the position of each syllable within a PPh; the vertical axis represents the coefficient values. The upper panel shows the coefficients of FFS; the lower panel shows those of SMS. 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.

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 4 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

Table 3
Evaluation of duration prediction at the PPh layer

Test	SMS	FFS
RE	93.0%	86.5%
TRE	42.4%	33.5%
r	0.760	0.814

Table 4
Evaluation of duration predictions at the PG layer

Test	SMS	FFS
RE	97.8%	94.8%
TRE	41.52%	31.7%
r	0.766	0.825

correlation coefficient $r = 0.766$ for SMS and 0.825 for FFS, an encouraging outcome for the current investigations.

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

Fig. 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 PG-initial ones.

Fig. 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.

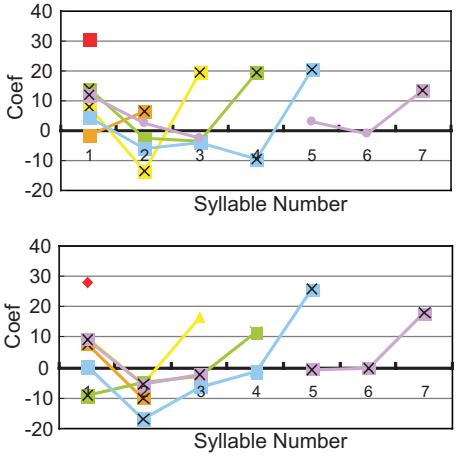

Fig. 5. Illustration of the coefficients of the initial PPhs at the PG layer. The upper panel shows coefficients of syllable durations of SMS; the lower panel shows coefficients of FFS.

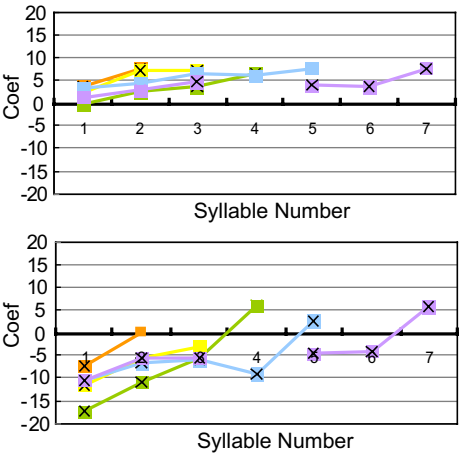

Fig. 6. Illustration of the coefficients of medial PPhs at the PG layer. The upper panel shows coefficients of syllable durations of SMS; the lower panel shows coefficients of FFS.

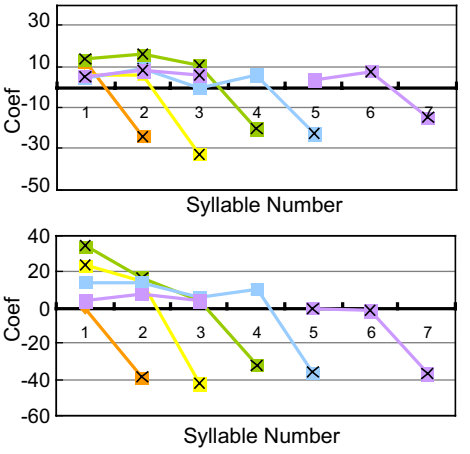

Fig. 7. Illustration of the coefficients of final PPhs at the PG layer. The upper panel shows coefficients of syllable durations of SMS; the lower panel shows coefficients of FFS.

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.

Finally, by adding up the predictions of each prosodic layer, we can derive a total prediction

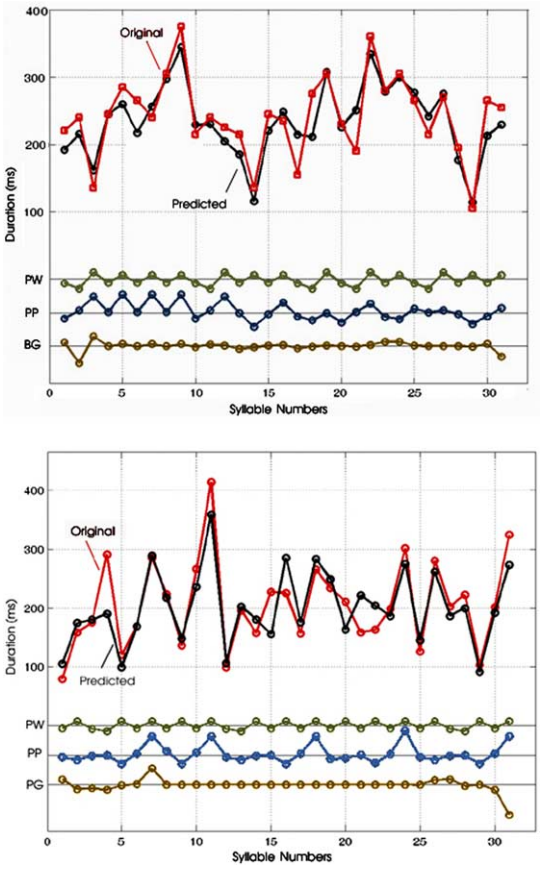

Fig. 8. The upper portion shows a comparison of derived predictions from all prosodic layers combined (in black on the web) to the original speech data (in red on the web). The lower portion shows the prediction generated at each prosodic layer. The upper panel shows coefficients of syllable durations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

of temporal allocation across phrases under grouping. Fig. 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 SYL 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.

2.2.2. Discussion

Figs. 5–7 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 Figs. 5 and 6). PG-final PPhs exhibit a reverse pattern of final-syllable shortening (shown in Fig. 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, 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 and Lee, 2004) from each prosodic layer cannot account for the final duration output independently. In particular, duration prediction at the SYL layer was only around chance level, but cumulatively, over 90% of the duration output was accounted for. It is apparent why concatenating syllables with only lower level (such as lexical) 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 six speakers and two 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. Figs. 9–13 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. Fig. 9 shows intensity distribution at the PW layer where PWs from one to four syllables were analyzed. Fig. 10 shows intensity distribution at the PPh layer where PPhs from one to nine syllables were analyzed. Figs. 11–13 show

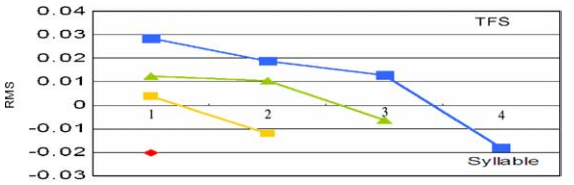

Fig. 9. Regression coefficients of intensity distribution at the PW layer where PWs from one to four syllables were analyzed. A gradual decline of intensity occurred over time. Longer PWs require more energy initially.

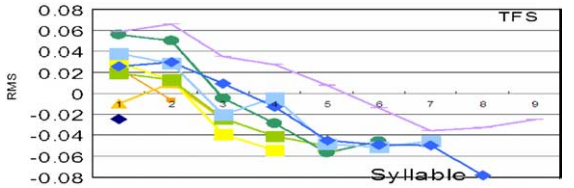

Fig. 10. Regression coefficients of intensity distribution at the PPh layer where PPhs from one to nine syllables were analyzed. A gradual decline of intensity occurred over time. Note how the energy level begins high and declines gradually over time and how the longer a PPh is, the more energy it requires.

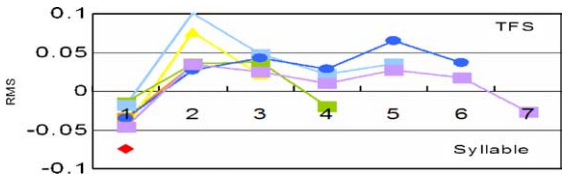

Fig. 11. Regression coefficients of intensity distribution of the PG-initial PPhs at the PG layer where PPhs from one to seven syllables were analyzed. The energy level is low at the first syllable, increases sharply at the second syllable, and declines with variations.

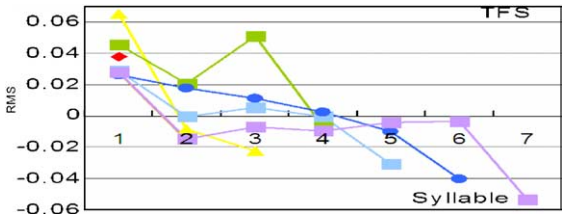

Fig. 12. Regression coefficients of intensity of the PG-medial PPhs at the PG layer where PPhs from one to seven syllables were analyzed. Note how energy level begins high and declines over time.

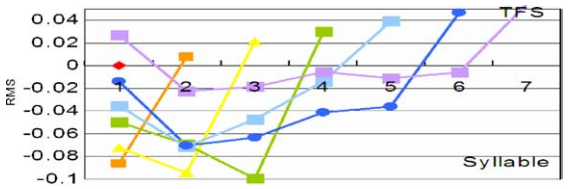

Fig. 13. Regression coefficients of intensity distribution of the PG-final PPhs at the PG layer where PPhs from one to seven syllables were analyzed. Note how the pattern reverses compared with the patterns found in PG-initial (Fig. 11) and PG-medial (Fig. 12) PPhs. A distinct increase of energy occurred at the final syllable.

intensity distribution at the PG layer where PG-initial, PG-medial and PG-final PPhs from one to seven syllables were analyzed.

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

Results of percentage of contribution from each prosodic layer were also obtained, as with duration patterns. At the SYL level, segmental identity accounted for 51% of intensity distribution. At the PW level, the contribution of intensity is insignificant, although the gradual declination exists. However, at the PPh level, the contribution of intensity is accounted for 14% more of intensity distribution, indicating that the prosodic phrase is a more significant unit for amplitude distribution patterns for fluent speech than prosodic words. Moreover, the shorter final PPhs had a wider coefficient range. We believe the different cross-phrase pattern of intensity distribution is closely associated with the perceived result of the terminal end of a speech paragraph in addition to F_0 contours and duration patterns. 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 and 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 and Chou, 1999) that specified 5 deg 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 three 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 two males and four 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 (two males and

two females) who read at speaking rate of 224, 362, 275 and 306 ms/syllable, respectively. Two speakers were radio announcers who read relatively faster at the speaking rate of 234 and 236 ms/syllable, respectively. We observed some distinct differences between untrained speakers and radio announcers. One was in speaking rate and another was the number and type of pauses/breaks used. In general, radio announcers used faster speaking rate (235 ms/syllable) than untrained speakers (292 ms/syllable), paused less during speaking (less B2's and B3's), and change breath more often (more B4's). Whereas the untrained speakers tend to speak slower, used more minor breaks (more B2's and B3's), but did not seem to change breath nearly as often (less B4's). The results may be representative of trained public speaking style vs. untrained informal way of speech production in pause and breathing style, with the untrained speakers sounding more halting than the announcers. Table 5 presents cross-speaker, cross-speaking-rate comparisons. Since the text each speaker read varied slightly, only overlapped text were compared. The purpose was to see how many degrees of boundary breaks exist within and across speaking rates in fluent speech.

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

Table 5
(A–F) Comparison of perceptual labeling of six speakers' breaks of overlapped portion of read speech

	B1	B2	B3	B4	B5
(A) Speaker F001 (speaking rate: 224 ms/syllable)					
Number	25369	14425	3163	71	1646
μ/σ	8/16	14/21	215/158	407/114	NULL
(B) Speaker F01S (speaking rate: 362 ms/syllable)					
Number	11084	4698	3630	193	473
μ/σ	2/11	11/27	342/245	717/212	799/434
(C) Speaker F03S (speaking rate: 275 ms/syllable)					
Number	12672	4888	4202	132	574
μ/σ	4/11	14/23	276/194	649/136	NULL
(D) Speaker M02S (speaking rate: 306 ms/syllable)					
Number	12409	5046	4303	250	546
μ/σ	1/9	10/26	315/264	742/234	949/242
(E) Speaker M051P (speaking rate: 234 ms/syllable)					
Number	6663	3327	1207	270	130
μ/σ	0/2	3/10	249/207	520/124	619/110
(F) Speaker F051P (speaking rate: 236 ms/syllable)					
Number	6645	3352	1157	287	150
μ/σ	2/6	6/13	215/152	332/164	399/226

Speakers F001, F01S, F03S and M02S were untrained native speakers; M051P and F051P radio announcers. For each speaker, the number of each perceived break was presented, where μ is mean duration of the break and σ standard deviation. Note that speakers F001 and F03S were given read relatively shorter paragraphs instead of longer text, the end of each paragraph was a complete recording unit, therefore, PG-final breaks (B5) were not available for measurement and hence was labeled NULL.

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 deg 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 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. Ten 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 1–6 s) in duration. All of the chosen PGs ended in yes–no questions without phrase-final mono-

syllable question particles. Among the 10 speech samples, five ended in two-syllable PWs; the other five in three-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. Four subjects, one male and three females, 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. Ten 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 2.5–6 s) in duration. Among the 10 declarative speech samples, eight ended in two-syllable PWs; two in three-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 four 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. Ten 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 2.2–5.7 s) in duration. Two of these 10 yes–no questions ended in two-syllable PWs; three in three-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 four 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 three 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 shows the results of Experiment 1, i.e., perceptual identification of humming of yes–no question intonation without question particles. Fig. 14 plotted the same results.

Results from Experiment 1 show that correct identification was best when the entire PG contour was presented. Identification begins to decay to below chance at one syllable edited off from the terminal end, and is the worst when two syllables were edited off. However, note that identification improved when three 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 7 shows the results of Experiment 2, i.e., perceptual identification of humming of declarative intonation as well as yes–no question intonation without question particle. Fig. 15 plotted the same results.

Table 8 shows the results of Experiment 3, perceptual identification of humming of declarative intonation, yes–no question intonation without question particles, and yes–no question with question particles. Fig. 16 plotted the same results.

Table 6
Correct identification rates of yes–no questions without question particles

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

Ss: subjects, S1–S4 represent subjects 1–4. *A*: tokens of full PG, *B*: tokens without last syllable, *C*: tokens without last two syllables, *D*: tokens without last three syllables.

Table 7
Correct identification of declarative vs. yes–no questions without question particles

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

Ss: subjects, S1–S4 represent subjects 1–4. *A*: tokens of full PG, *B*: tokens without last syllable, *C*: tokens without last two syllables, *D*: tokens without last three syllables.

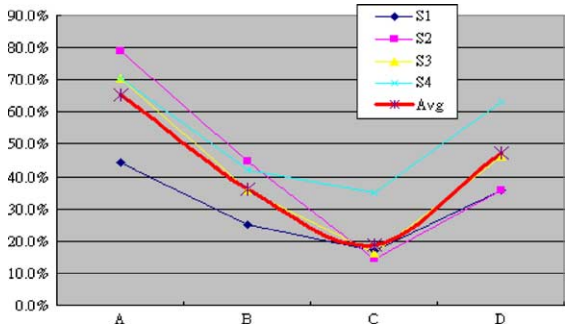

Fig. 14. Correct identification rates of yes–no questions without question particles by listeners.

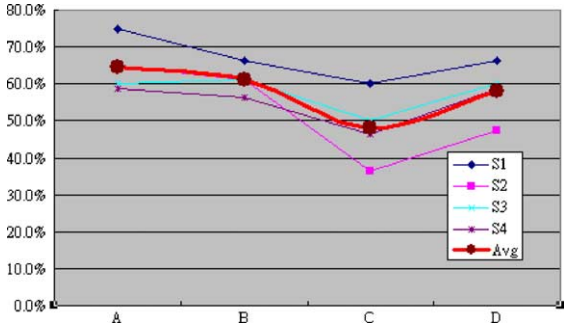

Fig. 15. Correct identification of declarative vs. yes–no questions without question particles by listeners.

Table 8
Correct identification of declarative utterances vs. yes–no questions vs. yes–no questions with question particles

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

Ss: subjects, S1–S4 represent subjects 1–4. *A*: tokens of full PG, *B*: tokens without last syllable, *C*: tokens without last two syllables, *D*: tokens without last three syllables.

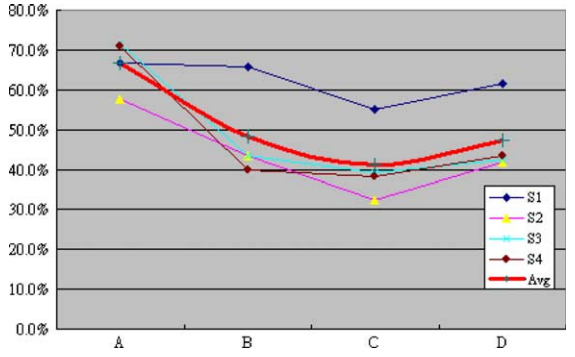

Fig. 16. Correct identification of declarative utterances vs. yes–no questions vs. yes–no questions with question particles by listeners.

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 did not end right. Second, they looked for final lengthening as an indicator for question intonation. In other words, the ending was not 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 cues the listener is looking for to process input of connected speech on-line, 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 and perceived. 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. Listening to solo singing is also different from chorus singing. 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 is 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 three-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 Fig. 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.

As expected, the F_0 contours of identical target syllables exhibit different patterns of overall contours as well as register (tone height). They also differ in syllable duration and intensity (not shown in the figure). All of these variations are related to PG positions. Note that each of the three PPh in the carrier sentence appears to possess identifiable intonation patterns. However, the 29-character/syllable carrier sentence is a lot shorter than the pieces of text used to collect our fluent speech data, one breathing cycle is sufficient to produce it. This brings back our earlier discussion (Section 1) of how the type of speech data (isolated or relatively short sentences) collected may inadvertently affect results yielded and interpretation thereafter, and PG-medial intonations with respect to PG lengths adjust (Section 2.1). Without our earlier data, it would be no surprise to analyze the prosody of such sentences by three intonations with two boundary breaks. However, we reiterate here that the PG effect is most notably marked by how it

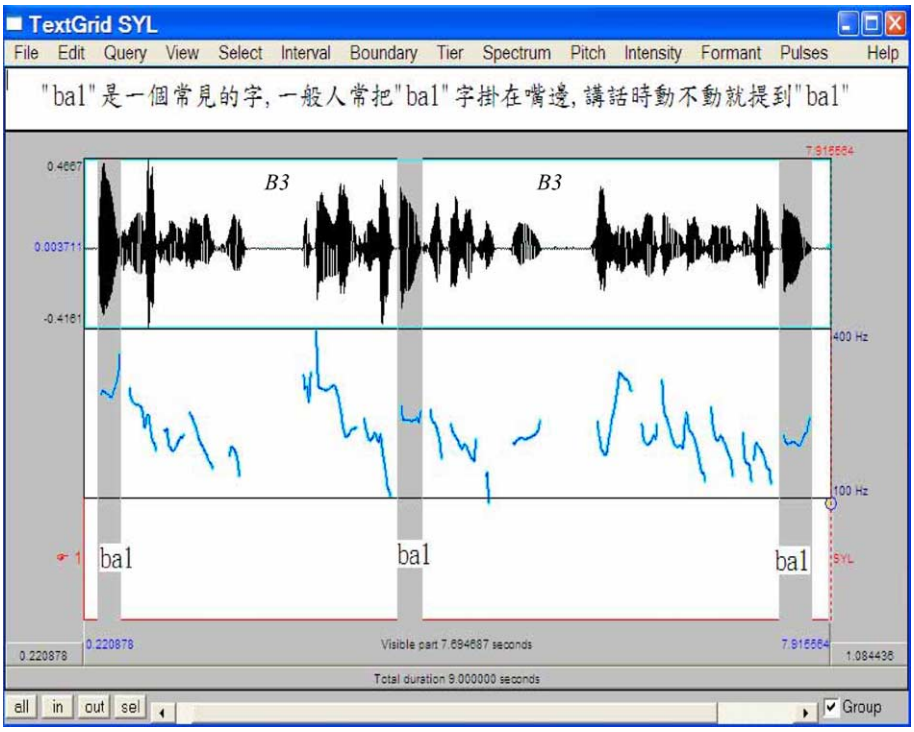

Fig. 17. The waveform and F_0 tracking of a carrier sentence with three target syllables "ba1" embedded in three darkened positions. ("ba1" is a frequently used syllable, people say "ba1" very often, often times when people speak, they would use "ba1"). The target syllable "ba1" occurred in three PG positions, namely, PG-initial, PG-medial, and PG-final to provide PG-related information. Note how high-level tone 1 in the target syllable 'ba' exhibits three different F_0 patterns in the three positions.

begins, whether it ends, and finally where and how it ends, as shown from our perceptual experiments (Section 3). Note also how the PG-medial PPh did not decline at its end.

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 and 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 11 592 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*:

$$\begin{aligned} Ap = {} & \text{constant} + \text{coeff}1 \\ & \times \text{Pause length before phrase command} \\ & + \text{coeff}2 \\ & \times \text{Accumulated previous phrase command response} \\ & + \text{coeff}3 \times F_0\,\text{min in the Fujisaki model} \\ & + f(\text{Phrase command position in PPh}) \end{aligned} \tag{4}$$

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 as 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:

$$AccF_0 = \sum_{\text{prev } Ap} Ap \cdot \alpha^2 \cdot (t - T_{0i}) \cdot \mathrm{e}^{(-\alpha(t-T_0))} \tag{5}$$

$AccF_0$ could then represent previous accumulated intonation due to Eq. (5).

4.1.2. Building the PG intonation model

As discussed in (Tseng et al., 2004), the PG intonation has significant effects in the first and last PPh units only. Therefore, the parameter *Ap* in Eq. (4) in the intonation model can be modified as:

$$\begin{aligned} Ap = {} & \text{constant} + \text{coeff}1 \\ & \times \text{Pause length before phrase command} \\ & + \text{coeff}2 \\ & \times \text{Accumulated previous phrase command response} \\ & + \text{coeff}3 \times F_0\,\text{min in the Fujisaki model} \\ & + f(\text{Phrase command position in PPh}) \\ & + f(\text{Phrase position in PG(Initial, Medial, Final)}) \end{aligned} \tag{6}$$

Thus we considered the prediction of *Ap* in a layered perspective. Individual prosodic phrases are using the phrasal intonation model and the global effects are superimposed onto the phrasal intonation model in the last term of Eq. (6).

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

$$\begin{aligned}&= \text{constant} + CTy + VTy + Ton + PCty \\ &\quad + PVty + PTon + FCty + FVty + FTon \\ &\quad + \text{2-way factors of the above factor} \\ &\quad + \text{3-way factors of the above factor}\end{aligned} \tag{7}$$

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 Fig. 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 cannot be explained at the SYL layer can be further explained by the PW layer. Accordingly, the syllable duration is postulated as:

$$\begin{aligned}DurS\ (\text{ms}) &= \text{Syllable intrinsic duration} \\ &\quad + f_{\text{PW}}(\text{PW length, position in PW})\end{aligned} \tag{8}$$

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 Fig. 1 and vice versa.

The PPh layer affects the syllable duration in a similar way as the PW layer (shown in Fig. 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:

$$\begin{aligned}DurS\ (\text{ms}) &= \text{Syllable intrinsic duration} \\ &\quad + f_{\text{PW}}(\text{PW length, position in PW}) \\ &\quad + f_{\text{PPh}}(\text{PPh length, position in PPh}) \\ &\quad + f_{\text{IFPPh}}(\text{Initial/Final PPh length, position in PPh})\end{aligned} \tag{9}$$

where *DurS* means the modeled syllable duration, $f_{\text{PW}}(\cdot)$, $f_{\text{PPh}}(\cdot)$ and $f_{\text{IFPPh}}(\cdot)$ 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:

$$\begin{aligned}IntS\ (\text{NRMS}) &= \text{Syllable intrinsic intensity} \\ &\quad + f_{\text{PW}}(\text{PW length, position in PW}) \\ &\quad + f_{\text{PPh}}(\text{PPh length, position in PPh}) \\ &\quad + f_{\text{IFPPh}}(\text{Initial/Final PPh length, position in PPh})\end{aligned} \tag{10}$$

where *IntS* means normalized average syllable intensity, RMS value, $f_{PW}(\cdot)$, $f_{PPh}(\cdot)$ and $f_{IFPPh}(\cdot)$ 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 three-phrase carrier sentence (a PG of three PPhs) in initial, medial, and final positions, respectively (Fig. 17 with syllable "ba1" embedded in it shows 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 soundproof 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) (Charpentier and 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 three 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 is 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:

$$G_p(t) = \begin{cases} \alpha^2 t \cdot \exp(-\alpha t) & \text{for } t \geqslant 0 \\ 0 & \text{for } t < 0 \end{cases} \tag{11}$$

In Eq. (11), the time to reach its maximum is $1/\alpha$, since the maximum phrase value, say P, will be:

$$P = Ap \times \alpha \times \exp^{(-1)} \tag{12}$$

where $1/\alpha$ is substituted into t in Eq. (11). In Eq. (12), 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 from the carrier sentence and the first F_0 peak, denoted as P_p, of predicted PG intonation results the adjustment of the predicted Ap:

$$\Delta Ap = \widehat{Ap} - Ap = (P_c - P_p) \times \exp \times \alpha^{-1} \tag{13}$$

$\widehat{Ap}$ 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 ith syllable is given by

$$DurS_i^* = \begin{cases} OriDur(S_i), & i = 1, m/2, m \\ OriDur(S_i) - DF_i, & 1 < i < m/2, m/2 < i < m \end{cases} \tag{14}$$

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

$$\begin{aligned} DF_i = {} & M_{\mathrm{TC}}/M_{\mathrm{MC}} \\ & \times [f_{\mathrm{PW}}(\text{PW length, position in PW}) \\ & - f_{\mathrm{PW}}(2,1) \\ & + f_{\mathrm{PPh}}(\text{PPh length, position in PPh}) \\ & - f_{\mathrm{PPh}}(11,6) \\ & + f_{\mathrm{IFPPh}}(\text{Initial/Final PPh length, position in PPh})] \end{aligned} \tag{15}$$

where M_{TC} and M_{MC} are, respectively, the mean of syllable duration of the TTS corpus and the training corpus, and $f_{\mathrm{PW}}(\cdot)$, $f_{\mathrm{PPh}}(\cdot)$ and $f_{\mathrm{IFPPh}}(\cdot)$ are the same as that in Eq. (9), 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 to 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 ith syllable is given by

$$IntS_i^* = \begin{cases} OriInt(S_i), & i = 1, m/2, m \\ OriInt(S_i) - DF_i, & 1 < i < m/2, m/2 < i < m \end{cases} \tag{16}$$

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

$$\begin{aligned} DF_i = {} & M_{\mathrm{TC}}/M_{\mathrm{MC}} \\ & \times [f_{\mathrm{PW}}(\text{PW length, position in PW}) \\ & - f_{\mathrm{PW}}(2,1) \\ & + f_{\mathrm{PPh}}(\text{PPh length, position in PPh}) \\ & - f_{\mathrm{PPh}}(11,6) \\ & + f_{\mathrm{IFPPh}}(\text{Initial/Final PPh length, position in PPh})] \end{aligned} \tag{17}$$

where M_{TC} and M_{MC} are, respective, the mean of syllable intensity of the TTS corpus and the training corpus, and $f_{\mathrm{PW}}(\cdot)$, $f_{\mathrm{PPh}}(\cdot)$ and $f_{\mathrm{IFPPh}}(\cdot)$ are the same as that in Eq. (10), 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, three 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

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.

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 why more understanding of the prosodic structure of discourse effects to fluent speech is essential, and how cross-phrase templates of prosody-related pitch, cadence, intensity and boundary patterns may together account for the necessary look-ahead in narratives and spoken discourses during on-line production and perception. We believe our framework is also capable to adopt and accommodate any discreet intonation model at the PPh level. 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. Furthermore, we have 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.

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

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.

References

Charpentier, M.J., Stella, M.G., 1986. Diphone synthesis using an overlap-add technique for speech waveforms concatenation. In: Proc. ICASSP'86, pp. 2015–2018.

Chang, L., Chen, K., 1995. The CKIP part-of-speech tagging system for modern Chinese texts. In: Proc. 1995 Internat. Conf. on Computer Processing of Oriental Languages, pp. 172–175.

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

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

Chen, K., Tseng, C. Peng, H., Chen, C., 2004. Predicting prosodic words from lexical words—a first step towards predicting prosody from text. In: Proc. Internat. Sympos. on Chinese Spoken Language Processing (ISCSLP2004), pp. 173–176.

Fujisaki, H., 2002. Modeling in the study of tonal feature of speech with application to multilingual speech synthesis. In: Proc. SNLP-O-COCOSDA 2002.

Fujisaki, H., Hirose, K., 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. Jour. Acoust. Soc. Jpn. (E) 5 (4), 233–241.

Gussenhove, C., 2004. Types of focus in English? In: Buring, Daniel, Gordon, Matthew, Lee, Chungming (Eds.), Topic and Focus: Intonation and Meaning: Theoretical and Crosslinguistic Perspectives. Kluwer, Dordrecht.

Ho, A.-T., 1976. Mandarin tones in relation to sentence intonation and grammatical structure. Jour. Chin. Linguist. 4, 1–13.

Keller, E., Zellner Keller, B., 1996. A timing model for fast French. York Papers in Linguistics, 17, University of York, pp. 53–75.

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., 2000. A novel approach to the fully automatic extraction of Fujisaki model parameters. In: Proc. ICASSP2000, pp. 1281–1284.

Mixdorff, H., 2004. Quantitative tone and intonation modeling across languages. In: Proc. Internat. Sympos. on Tonal Aspects of Languages—with Emphasis on Tonal Languages (TAL2004), pp. 137–142.

Mixdorff, H., Hu, Y., Chen, G., 2003. Towards the automatic extraction of Fujisaki model parameters for Mandarin. In: Proc. Eurospeech 2003, pp. 873–876.

Selkirk, E., 1986. Derived domains in sentence phonology. Phonol. Yearbook 3, 371–405.

Shattuck-Hufnagel, S., Turk, A., 1996. A prosody tutorial for investigators of auditory sentence processing. Jour. Psycholinguist. Res. 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., Wang, L. (Eds.), Beijing Yuyin Shiyanlu (Working Papers in Experimental Phonetics). Beijing University Press, Beijing, pp. 73–130.

Shih, C., 1988. Tone and intonation in Mandarin. Working Papers of the Cornell Phonetics Laboratory 3, pp. 83–1009.

Shih, C., 2004. Tonal effects on intonation. In: Proc. Internat. Sympos. on Tonal Aspects of Languages—with Emphasis on Tonal Languages (TAL2004), pp. 163–168.

Tseng, C., 2002. The prosodic status of breaks in running speech: examination and evaluation. In: Proc. Speech Prosody 2002, pp. 667–670.

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

Tseng, C., Chou, F., 1999. A prosodic labeling system for Mandarin speech database. In: Proc. ICPhS'99, pp. 2379–238.

Tseng, C., Lee, Y., 2004. Speech rate and prosody units: evidence of interaction from Mandarin Chinese. In: Proc. Speech Prosody 2004, pp. 251–254.

Tseng, C., Cheng, Y., Lee, W., Huang, F., 2003. Collecting Mandarin speech databases for prosody investigations. In: Proc. Oriented COCOSDA 2003.

Tseng, C., Pin, S., Lee, Y., 2004. Speech prosody: issues, approaches and implications. In: Fant, G., Fujisaki, H., Cao, J., Xu, Y. (Eds.), From Traditional Phonology to Mandarin Speech Processing, Foreign Language Teaching and Research Process, pp. 417–438.

Xu, Y., 2002. Articulatory constraints and tonal alignment. In: Proc. Speech Prosody 2002, pp. 91–100.

Yuan, J., 2004. Perception of Mandarin intonation. In: Proc. Internat. Sympos. on Chinese Spoken Language Processing (ISCSLP2004), pp. 45–48.