Speech Prosody: Issues, Approaches and Implications

Chiu-yu Tseng, Shaohuang Pin and Yeh-lin Lee

Abstract

In the development of unlimited Mandarin Chinese TTS, two fundamental problems have been the greatest barriers to progress and the most difficult to solve: (1.) how to predict prosody from text, and (2.) how to achieve more natural speech output. These impediments are not simply technology-related problems at the application end, nor are they uniquely Chinese, but rather reflect a major gap in our understanding of speech prosody. In other words, it is our conceptualization of the characteristics of speech flow that has been lacking. This inadequacy of understanding is at least twofold: one aspect being the overall organization of the prosody of connected speech, another being speech prosody in terms of its relationship to speech planning. The present paper discusses the source of these problems, some possible solutions and some promising directions for future research. We propose that prosody research on speech flow in Mandarin must include the following in the scope of its investigation: (1.) the characteristics of connected speech that constitute the prosodic properties of speech flow, *i.e.*, units, boundaries, F_0 contour patterns, tempo and intensity, (2.) a wider scope with respect to the type of speech data collected, *i.e.*, paragraphs instead of isolated or unrelated sentences, (3.) consideration of prosody in relation to speech planning at levels other than that of the individual phrase, *i.e.*, information at lexical, syntactic and higher levels, and (4.) development of a framework for the overall organization of prosody in speech flow, *i.e.*, one that incorporates and accounts for the above-mentioned features. In addition, this paper will propose a base form of F_0 contour patterns for Mandarin speech prosody and a framework for the temporal organization within those patterns.

1. Introduction

Traditionally, phonetic investigation has focused mainly on establishing phonological systems for the world's languages and on comparisons of those systems, with an emphasis on how speakers produce the sounds of particular linguistic systems rather than on how speakers actually speak to each other. Consequently, speech sounds were often elicited and analyzed outside of any larger

context. This bottom-up approach has led to the choice of isolated speech units such as words, short phrases or sentences as the only units of investigation.

By linguistic definition, syntactic structure governs how these units are organized into phrases and sentences, and studies of intonation have often been similarly grounded in principles of syntactic structure. This standard linguistic approach has inadvertently resulted in treating speech flow as strings of concatenated phrases or sentences, with no higher governing structure. In other words, the units of syntactic investigation have become the units of prosody research; phrases and sentences have become the units of intonation, and uninterrupted speech has come to be regarded as a succession of such phrases/sentences. Though the term "intonation group" has often been used in analyses of discourse and conversation, there has been no consistent operational definition of how such groups are formed in connected speech, whether they bear any relationship to one other, or what kind of overall prosodic grouping characteristics might be derived from those groups. The intonation contours developed under current theory seem applicable only to simple sentences or short phrases. Connected speech has been viewed as the outcome of the concatenation of such phrases and sentences, with much "local" attention paid to how the fragments are strung together. So, it comes as no surprise that no intonation models for complex sentences have been developed. The substantial literature in segmental coarticulation is one example of this research orientation, with its exclusive focus on how segments are concatenated within a single phrase; this approach has also been applied to studies of tonal coarticulation (e.g. Xu, 2002). As a result, prosody studies have pretty much stopped at the phrase and/or simple-sentence level for Mandarin Chinese (a language in which short phrases occur frequently), and not much attention has been paid to whether there may be some kind of a "global" higher constraint, which organizes phrases and sentences into larger units within a discourse. In short, the scientific study of sounds and sound systems had focused little attention on how speech sounds form larger units in speech flow, until the development of speech technology, particularly unlimited TTS (Text-to-Speech Synthesis), demanded a more systematic analysis of the characteristics of fluent speech.

Research in the development of Mandarin Chinese TTS has demonstrated that predicting prosody from text by using punctuation marks fails to produce natural speech output. Instead, this strategy produces a series of the right sentence or phrase intonations in short and choppy unrelated sequences. Hence, it is clear that predicting prosody from punctuated text requires a more detailed system of prosodic specifications. In the sections to follow, we will show that in order to learn more about prosody in relation to speech flow, we must begin by collecting speech data of read paragraphs instead of phrases or sentences in isolation. The next step would be to identify and systematically characterize speech units at levels higher than the individual phrase. In other words, the characteristics of running speech will emerge only through higher-level analysis of continuous speech data within the context of speech flow. Consequently, a notation system with the capacity to transcribe speech flow in chunks above the sentence level is essential to the labeling system design; within such a system, it is also necessary to describe these larger chunks as a single prosodic unit.

Once paragraphs of speech data have been collected and labeled for higher-level prosodic phenomena, it will be possible to look at individual intonation contours in relation to one another within the larger units, and to further study the acoustic, as well as the perceptual phenomena involved. In addition to analyzing the intonation contours, chunking and breaks/pauses of speech flow, we could analyze patterns of phrase grouping into larger discourse units, characteristics of the temporal alignment of speech flow, and patterns of fluctuation in intensity. Clearly, speech flow exhibits perceptible overall F_0 contours as well as rhythmic patterns. Thus, developers of any model of speech flow must consider the following questions: How should F_0 contour patterns be described and represented? How is the rhythm of speech executed in terms of duration? How should temporal allocation across speech flow be characterized? What is the best way to predict fluctuations in intensity? In particular, how should these physical characteristics be organized into a coherent system of description, which is able to predict their distribution in fluent speech?

In the present paper, we will demonstrate the following: how we have segmented speech data of read paragraphs into perceptible chunks; how we have categorized intonation contours related to phrase grouping and derived a possible framework for the organization of those groupings; how we have derived a basic framework of temporal allocation within and across syllables, and the relation of that framework to prosodic organization; and how the aforementioned information will contribute to understanding the prosody of speech flow.

Our prosodic framework is actually quite simple; it concatenates phrases and sentences into strings and labels them using notation that specifies their position in relation to other phrases and sentences within and across prosodic groups. Using this "top-down" approach to study Mandarin Chinese speech prosody, we found acoustic evidence for the existence of a prosodic unit larger than an individual phrase/sentence. In addition, we found that the location of breaks in speech flow does not always correspond to punctuation marks. These prosodic features may be particularly prominent in Mandarin, but they are definitely not language-specific. The following English example, which also illustrates this point, was taken from a study of the relationship between speaker's intention and speech planning (Bratman, 1988). (Note that punctuation, which is much more strictly constrained in English than in Mandarin, was modified to reflect the speaker's intention to produce a larger unit): *"I am about to go running; I hear on the radio that the pollen count is high; I recall my general policy of not running when the pollen count is high; and so I decide not to run."*

In this light, it is revealed that speech flow is governed by a higher prosodic organization, which groups phrases into larger units. These units were introduced in previous studies as the Prosodic Group (PG) (Tseng & Chou, 1999; Tseng, 2002). How the structure of a given PG relates to syntactic as well as semantic information and constraints is still under investigation. In this paper, we will present discussions of the prosody-related phenomena we have studied so far, using analyses of both text and speech data to illuminate our theory of prosody in fluent speech as a structural, hierarchical grouping of phrases and sentences.

Postulation of the existence of a hierarchical prosodic structure requires the development of a possible model, as well as a set of constraints on that model. It requires a precise description of how phrases are grouped and how many phrases a prosodic group (PG) can include; it also requires the development of a set of constraints on those groupings. Any viable model for speech prosody should fulfill these requirements; it should be able to model not only individual phrasal and sentential intonations, but also the grouping of those intonations (Tseng, 2003). The analysis of two sets of data will be presented to illustrate this point. A third set will provide an account of speech rhythm in terms of syllable duration within the proposed prosody framework. We believe that the collective preliminary findings have already accounted for some, if not most of the prosodic characteristics unique to Mandarin Chinese, and that these findings can be applied to the generation of Mandarin speech prosody. We also believe that output naturalness of synthesized speech can be improved significantly by adopting the proposed framework, although more studies are needed for confirmation. In addition, we will briefly discuss promising directions for future research to further our understanding of speech flow.

This paper will present 3 experiments, which will illustrate the following points: 1. Human prediction of prosody had more than one outcome, which suggests that we must further explore the range of factors affecting human predictions, 2. Analysis of speech data showed perception of phrase groupings to be consistent across speakers, even when variations in prosody output had occurred, which highlights that the same difficulties exist in text processing by hand as in prosody simulation, 3. Basic patterns of temporal allocation for Mandarin Chinese can also be modeled. Our framework for Mandarin prosodic organization has derived a base form of prosodic phrase grouping, which provides an account for all of the above.

2. Experiments

2.1 Experiment 1

Using punctuation marks as indicators of text processing, and therefore possible markers of speech unit boundaries in fluent speech, the aim of Experiment 1 was to explore (1.) the range of individual differences in native Mandarin Chinese speakers' chunking of text, (2) the role punctuation marks play in Chinese as cues to syntactic structure as well as delineators of semantic domains, vis-à-vis the role played by punctuation marks in alphabetic and inflectional languages and (3.) the ways in which a Mandarin sentence, as denoted by the punctuation mark "period", is formed by grouping short phrases into complex sentences, which are long enough to be short paragraphs. The simple and short sentences often found in syntactic studies were not identified as the largest units.

2.1.1 Methodology

Two native speakers of Mandarin Chinese participated in the experiment. Both subjects had received Master's degrees and were fluent readers of Chinese. Test materials consisted of two pieces of text, which were 1118 and 1075 characters in length, respectively. Both pieces of text were composed of very high-frequency words, and all punctuation marks were removed. Each subject was asked to independently punctuate the unpunctuated text. Their results were analyzed and compared. The task may appear quite simple and simple-minded on the surface. However, it is worth noting that unlike alphabetic and inflectional writing systems, a logographic writing system such as Chinese does not provide sufficient morphological indicators to syntactic structures, thereby making punctuation a relatively free reference to syntactic structures as well as to semantic domains.

2.1.2 Results

The punctuation task results were analyzed by two factors namely, commas and equivalent intermediate markings; and periods and their equivalent terminal markings (e.g. question marks and exclamation points). Table 1 shows the results: the number of punctuations used by type and by Ss, as well as the number of commas used before a period.

Table 1: Results: number of periods used by each S and the mean number of commas before each period. There are 1118 characters in Text 1, 1075 in Text 2.

	S1			S2		
	# of periods	M of commas	Var.	# of periods	M of commas	Var.
Text1	27	2.96	8.83	18	3.21	5.39
Text2	28	2.60	1.73	24	2.88	4.11

The performance of the two subjects differs both in the total number of periods used and the number of commas before each period, indicating individual variation in processing units. S1 used a total of 107 punctuation marks for Text 1 and 104 for Text 2, whereas S2 used 77 punctuations for Text 1 and 93 for Text 2. Note that S1 used more periods than S2 for both pieces of text, indicating that constraints varied on the choice of the largest unit. The mean numbers of within-period commas also varied, indicating that variation was present in the number of phrases within the largest unit. For S1, who used more periods, the total number of complete sentences generated increased, while fewer commas were used within each sentence. In summary, comparison of these simple results demonstrates the following: (1) Ss vary considerably in their grouping of syntactic as well as

semantic units. (2.) Ss vary considerably in their grouping of phrases into complex sentences, though overlap does occur. S1 tended to group 3 phrases into a unit, whereas S2 tended toward 3+ phrases. (3.) Neither subject showed any inclination toward choosing simple sentences; both subjects invariably chose to group phrases into complex sentences. In addition, preliminary analysis indicates that preferred complex sentences consist of at least 3 phrases. These results show why categorizing intonation using only simple-sentence types is insufficient, which had been previously discovered in the development of unlimited TTS. These results also highlight the necessity for identifying intonation groups that reflect more complex groupings.

Table 2 shows the results of matching punctuation marks across Ss. In this case, matching is defined as placement of identical punctuation marks at identical locations across the text between Ss. Of the 107 punctuation marks used by S1 for Text 1, only 53 (49%) matched S2's. Of the 104 punctuation marks used for Text 2, only 64 (61%) matched those of S2. If identical placement of punctuation marks is taken as evidence of an identical grouping strategy, a relatively low correspondence was found.

Table 2: Comparison of matched *vs.* mismatched punctuations. S1 used 107 punctuation marks for Text 1, 104 for Text 2. S2 used 77 punctuation marks for Text 1, 93 for Text 2.

Text 1	Match (53 in total)		Mismatch (59 in total)		
	Commas	Periods	Same location different mark	S1-only Marks	S2-only Marks
#	40	13	19	35	5
%	75.4%	24.6%	32.2%	59.3%	8.5%
Text 2	Match (64 in total)		Mismatch (54 in total)		
#	48	16	15	25	14
%	75%	25%	27.8%	46.3%	25.9%

These results suggest that prosody prediction from punctuated text may also need to accommodate individual variation. In other words, an ideal framework would allow for the possibility that more than one prediction may be derived. Specifications of the range of variation would also be desirable.

2.2 Experiment 2

The aim of Experiment 2 was to investigate (1.) whether speakers exhibit similar chunking and

grouping behavior to break up speech flow, (2.) whether speakers demonstrate an overall prosodic pattern for phrase grouping, and (3.) whether individual variations in chunking and grouping also exist in speech read aloud from paragraphs of text.

2.2.1 Methodology

Speech data of 4 native speakers of Mandarin Chinese (2 males and 2 females) were used in Experiment 2. Each speaker read a total of 599 paragraphs of text composed of the highest-frequency words from the CKIP database (**http://godel.iis.sinica.edu.tw/CKIP/**). These paragraphs were hand tailored to resemble spoken style more closely, and were balanced for phonetic and tonal distributions. The paragraphs ranged from simple 2-character sentences up to 181-character complex sentences, with a focus on longer, complex sentences. Table 3 summarizes the speech data used in Experiment 2. A total of 24,803 syllables (490 minutes or 1,101MB) of speech data were labeled and analyzed.

Table 3: 599 paragraphs of text, with a total of 24,803 characters (which is also 24,803 syllables). This table also shows the total duration of speech data and the corresponding size of digitized and labeled speech

Speaker	Total Characters/Syllables	Analogue Data	Labeled Digitized Data
F01	24,803	176m	322MB
F03		151m	277MB
M01		151m	276MB
M02		123m	226MB

Waveforms and Fundamental frequency contours were obtained for the speech data. Three sets of manual transcriptions were obtained. The data were labeled by 3 independent transcribers for perceived boundaries and breaks (pauses), using a 5-step break labeling system developed in the spirit of TOBI (Tseng & Chou 1999a, 1999b). The rationale of the labeling system emphasized perceived boundaries in speech flow, marked by breaks defined in terms of the duration of their silent portions and their respective preceding chunks, as well as in terms of units larger than the individual phrase. In this system, break labels included the following: B1 indicates a syllable boundary; B2 indicates a break after a prosodic word (PW); B3 indicates a prosodic phrase boundary (PPh), B4 comes directly after a complete intake of breath or a breath group (Lieberman, 1976) that is preceded by an utterance (UTR), and B5 occurs at the end of a prosodic group (PG). Prosodic units included syllable, PW, PPh, UTR and BG and PG, where PG indicates the end of a

complete speaking unit in speech flow. The results of perceptual labeling were consistent within and across transcribers.

2.2.2 Results

Analysis of these perceptual results showed that grouping of phrases into identifiable larger prosodic units was consistent across listeners; perception of smaller prosodic units and boundaries exhibited the same consistency. Moreover, the boundaries and the breaks that followed those small units also exhibited a systematic consistency in duration, apart from some individual variation. Table 4 shows the range and average of perceived breaks for each speaker, as well as statistical analyses between breaks. Figure 1 shows the distribution of duration for perceived breaks across the 4 sets of speech data.

Table 4: Range and mean of labeled breaks in msec for 4 speakers are shown in the upper panel. The lower panel shows statistical analyses of break durations.

(msec)	**B2 (F03/F01/M02/M01)**	**B3 (F03/F01/M02/M01)**	**B4 (F03/F01/M02/M01)**	**B5 (F03/F01/M02/M01)**
MAX	**477 / 521 / 890 / 840**	**1228/1906 /1860 /1820**	**1272/2033/2315/1865**	**# / 2866 / 590 / 1685**
MIN	**0 / 0 / 0 / 0**	**0 / 0 / 0 / 0**	**354 / 0 / 0 / 0**	**# / 451 / 1690 / 545**
AVG	**13 / 11 / 10 / 7**	**274 / 336 / 311 / 445**	**648 / 539 / 583 / 636**	**# / 799 / 948 / 1161**

	F01			F03			M01			M02		
Bi	B3	B4	B5	B3	B4	B5	B3	B4	B5	B3	B4	B5
N	3553	140	30	4180	132	#	2912	140	21	4227	195	41
avg	342	720	799	275	648	#	451	990	995	314	737	948
sd	245	213	433	193	136	#	343	304	498	264	231	241
t	-20.407	-0.969		-30.5	#		-20.3	-0.044		-24.8	-5.1	
df	153	32		148.2	#		156.4	22.3		218	56.4	
SIG.	Yes	No		Yes	#		Yes	No		Yes	No	

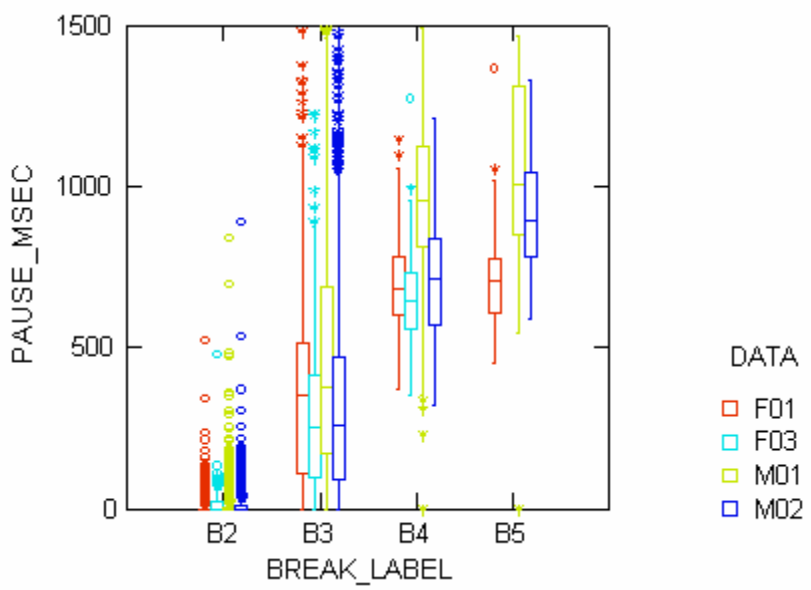


Figure1: Distribution of break duration (in msec) plotted in box-plot by speaker.

Although significant differences exist for each speaker between B2 and B3, and B3 and B4, there are none between B4 and B5, due to the small number of B5 tokens available, since paragraph- or discourse- final breaks without an utterance to follow could not be measured. However, a significant difference does exist between B3 and B5. Moreover, it is evident from the distribution shown in Figure 1 that differences in duration do exist for each level of perceived break in speech flow. Table 5 shows the average duration of PG for each speaker (in seconds) as well as average syllable duration (in ms). Note that the average numbers of PG did not vary much across speakers, even though the 4 speakers did exhibited differences in speech rate.

Table 5: Average numbers of syllable per PG and PG's relative durations for each speaker.

	M01	M02	F01	F03
# of Syls	38	37	37	39
avg length(sec) of each PG	10.389	8.399	10.438	8.019
avg syllable duration(msec)	266	221	273	201

Furthermore, an overall prosodic pattern for phrase grouping has been observed. Figure 2 presents a base form of the F_0 contours of a PG that characterizes phrase grouping in Mandarin Chinese

speech flow. This base form characterizes phrase grouping and the acoustic features related to grouping, including boundaries, breaks and respective F_0 contours. A prosodic group (PG) may consist of anywhere from 3 to 5+ phrases. Figure 2 includes 5 phrases to better illustrate the grouping effect. In terms of F_0 contour patterns, a PG can be characterized by two resets and two F_0 peaks (PG initial and PG final), a terminal trailing off and F_0 fall (PG final) and distinct units separated by breaks. The first reset and highest F_0 peak occur during the initial PPh of a PG; the second reset and the second highest F_0 occur during the last PPh of the PG. A sharp fall of the F_0 contour follows the highest peak; whereas a final lowering of the F_0 contour plus final lengthening follows the second highest peak at the PPh. These two F_0 patterns characterize the beginning and end of a PG, whereas the F_0 contour patterns of phrases that occur between the initial and final PPh's do not necessarily possess identifiable intonation patterns, nor do their respective F_0 contours dip lower than that of the PG-final PPh. Basically, the role of each phrase in relation to its respective position within a PG is quite clear, and a terminal fall is only exhibited at the end of the last phrase. In addition, specifications of breaks between prosodic units are necessary. The longest break (B5) always occurs between PG's, or rather, PG's are separated by B5. A B5 is always preceded by a final lowering and tapering-off of F_0 contour, accompanied by final lengthening. Together, these two cues signal the end of a prosodic group. At the same time, a B5 is always followed by another PG, so an F_0 reset plus peak-and-sharp-fall will signal the beginning of a new, upcoming prosodic group. From a top-down perspective, a PG can also been seen as the highest node of the prosodic hierarchy, which branches into prosodic levels or layers. A PG branches into UTR's, which are followed by B4's (or BG's); UTR's branch into prosodic phrases (PPh's), which are followed by B3's; PPh's branch into PW's, which are separated by B2's; PW's branch into syllables, which correspond to individual characters in the Chinese orthography.

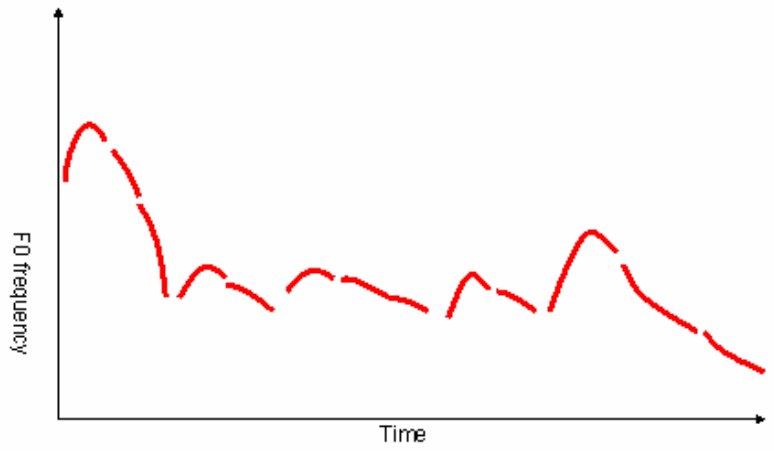


Figure 2: Base form of a Mandarin Chinese PG

The base form characteristic of phrase grouping has the following implications: (1.) The start of a PG is always marked by F_0 reset and a sharp fall, (2.) A second and lower reset occurs at the left edge of the terminal phrase. Only the terminal phrase exhibits characteristics of F_0 terminal fall and other acoustic phenomena correlated with final position, such as final lengthening, weakening of amplitude and the final and lowest drop in F_0. (3.) The domain of prosodic planning has a look-ahead window of 3+ prosodic phrases. In other words, in the absence of narrow focus or emphasis, the base form reflects speech planning and target shooting, the ultimate target being the PG-terminal fall with its set of finality features. As a result, phrase intonations within the PG have to be modified so that they reflect this initial high, overall target shooting across phrases toward the ultimate target.

Specifications of break levels are also necessary. However, results of break-labeling tasks showed that when reading punctuated text, speakers showed non-overlap of grouping even in cases of identical punctuation, indicating speakers' relatively free interpretation of punctuation marks within a text. Table 6 shows the results of the occurrence of aligned breaks across 4 speakers in actual numbers.

Table 6: Overlap of perceived breaks across 4 speakers.

	Number of Cross-Speaker Overlap
PW/B2	1643
PPh/B3	1164
UTR/B4	2
PG/B5	220

Table 7 shows the percentage of overlap in each speaker's speech data. Under each speaker, the black number in the upper panel gives the actual occurrence of labeled breaks; the purple number in the lower panel gives the percentage of overlapped breaks.

Table 7: Total count of each level of labeled break for each speaker and percentage of cross-speaker overlap relative to total number of occurrences.

	F01		F03		M01		M02	
	Occurrence	Ovlp %	Occurrence	Ovlp %	Occurrence	Ovlp %	Occurrence	Ovlp %
PW/B2	4555	36	4221	39	5365	31	4997	33
PPh/B3	3526	33	3656	32	2560	45	4267	27
UTR/B4	182	1	110	2	192	1	247	1
PG/B5	460	48	516	43	447	49	543	40

The low overlap across speakers shows that there is considerable variation in phrase grouping, indicating the possibility of individual differences in speech planning, particularly with respect to constraints on units, boundaries and groupings before and during speech production. Again, the reasons for the overlaps and non-overlaps may not be purely the result of syntactic and semantic constraints, but rather a reflection of windows of targeting and planning for different levels of constraints during speech production.

2.3 Experiment 3

The aim of Experiment 3 was to investigate (1.) whether patterns of syllable duration adjustment could be derived from speech data, (2.) whether there is evidence of interaction between syllable duration adjustments and prosodic-level units and (3.) whether the evidence found could predict temporal allocation and rhythmic structures in speech flow. Our hypothesis is that duration adjustment constitutes the tempo or rhythm of a language. We believe that duration adjustment is language-specific, systematic and predictable, and that its patterns should be incorporated into a framework of prosodic organization.

2.3.1 Methodology

Mandarin speech data representing 2 different speech rates, slower *vs.* faster speech, were used for Experiment 3. The slower speech was recorded from 1 male untrained subject (hence SMS for Slower Male Speech) reading 595 paragraphs ranging from 2 to 180 syllables; the faster speech from 1 female radio announcer's relatively faster reading (hence FFS for Faster Female Speech) of 26 long paragraphs ranging from 85 to 981 syllables. 90% of the two sets of text overlap. A total of 22350 syllables of SMS and 11592 syllables of FFS were analyzed. Average syllable duration was

304.7ms for SMS and 199.75ms for FFS. Both sets of speech data were first labeled automatically for segments using the HTK toolkit and SAMPA-T notations (Tseng and Chou, 1999), then hand labeled for perceived prosodic boundaries by 3 trained transcribers using the same system as Experiment 2. 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. A layered, hierarchical organization of prosody levels (the aforementioned system of boundaries and units) was used to classify prosodic units at levels of the syllable, PW, PPh, UTR and PG, with PG being the highest node of the hierarchy. Moving from the syllable layer upward to each of the higher prosodic units and levels, we examined each higher layer independently to see if it could account for residuals from one of the lower layers, and if so, how much was contributed by each level. All of the data were analyzed using DataDesk™ from Data Description, INC. Two benchmark values were used in this study to evaluate the closeness of the predicted value to that of the original speech data: residual error (R.E.) and correlation coefficient (r). Residual error was defined as the percentage of the sum-squared residue (the difference between prediction and original value) over the sum-squared original value.

2.3.2 Results

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

A Syllable-Layer Model was subsequently postulated as follows:

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

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

Table 8 shows benchmark values of the Syllable-Layer Model found in the two different speech rates. The residual error was 48.9% in SMS and 40.1% in FFS. In other words, the Model was able to account for 51.1% of syllable duration of the SMS and 59.6% of the FFS at the syllable layer. The residue that could not be accounted for at this layer was termed as Delta 1 and was dealt with at higher layers.

Table 8: Evaluation of duration predictions at the Syllable Layer

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

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

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

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

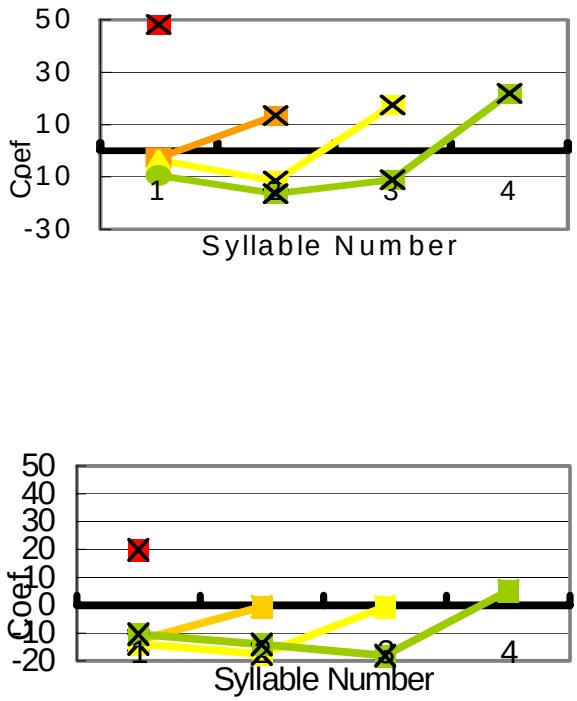


Figure 3: Coefficients of syllable durations obtained for both speech rates using the PW model. The horizontal axis represents the position of each syllable within a PW; the vertical axis represents the coefficient values. The upper panel shows coefficients of FFS; the lower panel shows those of SMS.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

Several interesting phenomena were observed: (1.) both speakers exhibit a pattern of PW-final syllable lengthening relative to the other syllables considered; (2.) the longer the PW, the longer the duration of the final syllable and (3.) different speech rates contribute to different degrees of variation in syllable duration. At the PW Layer, SMS showed within-layer syllable shortening but final-syllable lengthening in comparison with lengthening predictions made at the Syllable Layer. However, FFS showed the opposite: even when syllables of a PW were shortened, the final syllable maintained the duration predicted by the Syllable Layer. These results could be used to characterize speaker-independent beat and tempo, and could be a major feature used to describe and characterize individual speaking style. Table 9 shows benchmark values of the PW Model.

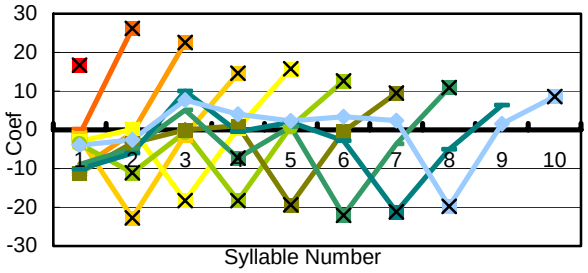
Table 9: Evaluation of duration predictions at the PW Layer

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

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

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

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



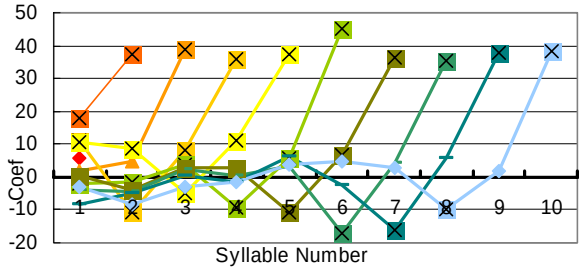


Figure 4: Coefficients of syllable durations obtained for both speech rates from the PPh model. The horizontal axis represents the position of each syllable within a PPh; the vertical axis represents the coefficient values. The upper panel shows the coefficients of FFS; the lower panel shows those of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

Figure 4 shows the following results: (1.) A clear cadence-like phenomenon in PPh. (2.) That there is not only lengthening of the PPh-final syllable, 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 10 shows the evaluation of predictions at the PPh Layer.

Table 10: Evaluation of duration prediction at the PPh Layer.

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

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

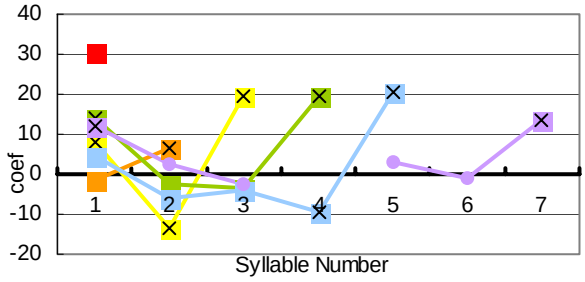
In order to investigate the influence of syllable duration on breath-group effect (the longer pause created by breathing which follows a UTR), we studied the residue from the PPh Layer (Delta 3) at the UTR Layer. Duration differences were found to occur more often at the initial and the final portions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

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

Table 11: Evaluation of duration predictions at the UTR Layer.

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

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



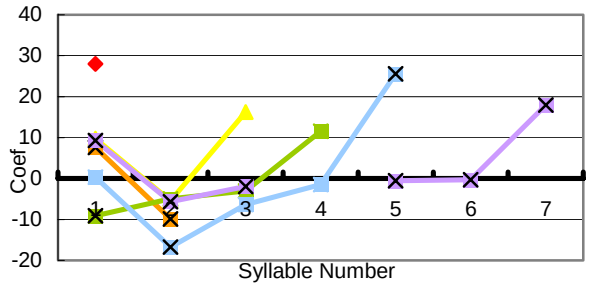


Figure 5: Illustration of the coefficients of the initial PPh's at the UTR Layer. The upper panel shows coefficients of syllable durations of SMS; the lower panel shows coefficients of FFS.

Figure 6 shows effects of the UTR layer on UTR-medial PPhs. The first syllable is shortened while the final one is lengthened for the UTR-medial PPhs considered, although this influence is more pronounced in FFS than in SMS. However, duration adjustments for UTR-medial PPh's are not as pronounced as UTR-initial ones.

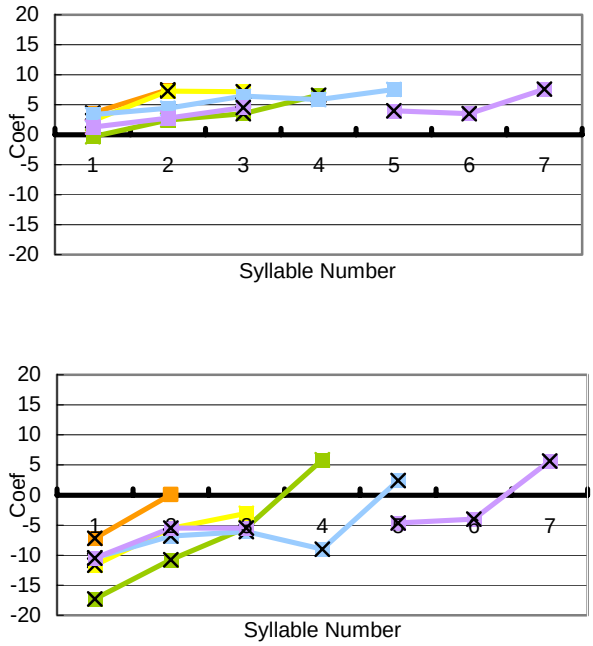


Figure 6: Illustration of the coefficients of medial PPh's at the UTR Layer. The upper panel shows coefficients of syllable durations of SMS; the lower panel shows coefficients of FFS.

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

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

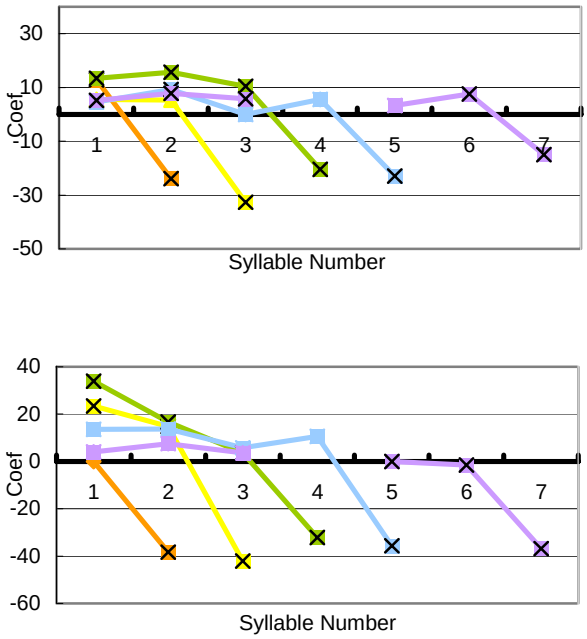


Figure 7: Illustration of the coefficients of final PPh's at the UTR Layer. The upper panel shows coefficients of syllable durations of SMS; the lower panel shows coefficients of FFS.

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

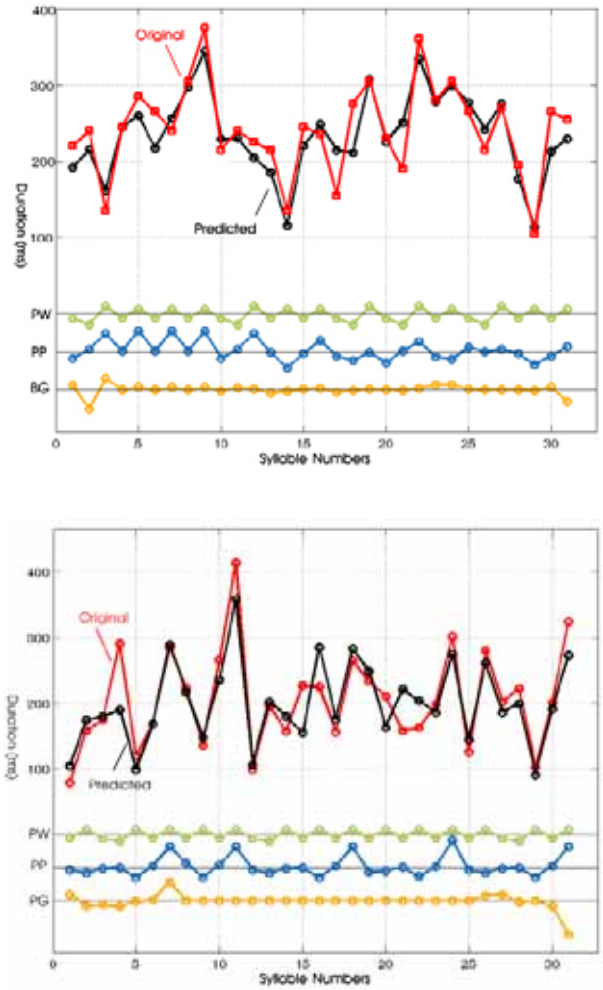


Figure 8: The upper portion shows a comparison of derived predictions from all prosodic layers combined (in black) to the original speech data (in red). The lower portion shows the prediction generated at each prosodic layer. The upper panel shows coefficients of syllable durations of SMS; the lower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

3. Discussion

In the context of our investigations, we will discuss the prosody of fluent speech with respect to the following three issues: 1. Collecting appropriate speech data for use in investigations of prosody; 2. Developing working models of prosody, and sets of constraints for those models; and 3. Exploring the implications of the experiments we have presented to inform future research directions and to improve computational modeling of speech..

Over two decades of speech synthesis, research on Mandarin Chinese has made significant progress; nevertheless, some major bottlenecks remain. At the prosodic level, two issues that merit proper attention remain unresolved. The first issue, best illustrated by previous collective efforts to develop unlimited TTS, is the need for tools capable of predicting prosody from text for cases in which prosody runs counter to punctuation. The second issue, best exemplified by persistent problems in development of speech synthesis, is finding a way to improve the overall naturalness of fluent speech. These two issues compel us to face the inadequacy of an approach that models intonation on the basis of syntactic sentence types only, treating each phrase as an independent simple sentence. In other words, gathering evidence about Mandarin Chinese declarative sentences and yes-no questions exhibiting intonation patterns similar to those of their English counterparts (Lin, 2002) without further specification above the sentence level moves us no closer to understanding the factors contributing to prosody of Mandarin fluent speech.

The results of Experiment 1 showed that the simple sentence was not the preferred size for a text-processing unit. Instead, the units subjects chose to mark with a period usually consisted of at least three phrases, which were set off by commas. Therefore, it would be natural to assume that such groupings are manifested in speech using prosodic forms that consist of something more complex than identical intonation patterns strung together. Moreover, these groupings are subject to individual differences. Thus, a model of Mandarin prosody for fluent speech should have the capacity to characterize such groupings and grouping-related characteristics, to predict how such groupings will be formed and to formulate constraints on the range of individual differences. Related prosodic manifestations and the potential for overall prosody patterns for such groupings also deserve careful investigation through analysis of speech data.

The second issue is that of inter-speaker variation. The relative flexibility of the domain size, which may relate to a speaker's planning strategy and/or speaking rate, suggests that more than one prediction for any given text should be expected and accommodated. In short, the results of Experiment 1 suggest that: (1.) In Mandarin Chinese, punctuation serves as only a loose reference to syntactic structures as well as semantic domains, which indicates that chunking and paragraphing strategies in connected speech may differ from those represented by punctuation. (2.) Inter-speaker variation has been found in punctuation of written texts, indicating that individual speakers' intention and planning scope may vary. (3.) Simple and short sentences, which are often seen in intonation studies, are not represented in text processing. Note that the rather flexible use of punctuation in Chinese, which allows 4+ commas before a period in running text, may also contribute to the long groupings in text. All observations collectively suggest that the canonical forms of phrase and sentence intonations determined by sentence types in isolation are insufficient to predict the prosody of connected speech output.

The speech data used in Experiment 2 were collected to investigate grouping phenomena, and our hypothesis was confirmed. It is important to note that paragraphs of up to 181 characters,

(corresponding to 181 syllables) with only one final period were deliberately selected to avoid production of short phrases or sentences in isolation. These materials elicited prosodic manifestations on levels higher than the individual phrase or sentence, thereby highlighting that the kind of speech data collected may also influence the perspective of studies conducted. As previously stated, the speech data used were transcribed and hand labeled for perceived chunks (units) and the breaks that followed them, all using cross-transcriber consistency controls. A labeling system that allows tagging of units and the breaks that follow them above the phrase/sentence level is instrumental to studying the prosody of speech flow. The labeled results were also indications that a model of phrase grouping must take the existence of distinct levels of boundaries and pauses into account. (This has been observed in many previous studies, such as Lehiste & Wang, 1977; Lehiste, 1979; Thorsen, 1985 to name a few). In order to study the prosody of speech flow, we need to collect fluent speech data, rather than short phrases in isolation, and we must examine the data from a broader perspective. This broader perspective calls for specifications of domain (or scope), units and boundaries, and boundaries require additional specifications for the strength of breaks between units. The results of Experiment 2 also proved that chunking in running speech is perceived systematically. Thus, we propose a hierarchical framework called PG (Prosodic Group), which reflects speakers' target shooting as the underlying canonical organization for Mandarin Chinese speech prosody. PG will consist of prosodic units and varying degrees of breaks, and its overall F_0 contour pattern was given in Figure 2. The overall pattern is characterized by reset, F_0, peaks, followed by another F_0 reset, followed final lengthening, weakening of amplitude and F_0 lowering. For obvious physiological reasons (see Lieberman, 1976), higher prosodic units will have longer breaks following them.

Successful prediction of prosody from text will additionally require a greater understanding of the interrelation among the different levels of constraints: syntax, semantics and the speaker's intention, under the assumption that these constraints are reflected in the speaker's choice of planning and processing units. It is no easy task by any means. However, characterizing and implementing PG-related characteristics using our current findings has proved to be relatively easier. Our recent attempt (Tseng, 2003, 2004) showed that a sentence intonation model tested on many languages (Fujisaki et al, 1996; 1998; Fujisaki, 2002; Mixdorff, 2000; 2001; Jokisch et al, 2000; Wang et al, 2000) could quite easily be extended to accommodate PG-related phenomena. By subsuming the immediate prosodic unit PPh under a PG node and further specifying its position, (i.e. PG-initial, PG-medial or PG-final), individual intonations could be successfully modeled. (Tseng, 2003, 2004). By further imposing the framework of boundaries and groupings, and categorizing intonations accordingly, the PG effect could be simulated. The results also showed that a prosody framework for Mandarin would require more detailed specification of the grouping effect than other languages. Phrasal intonation in Mandarin Chinese was found to be a component of a larger prosody unit (PG); as such, it is only significant in the context of the overall organization of connected speech prosody.

Speech technology applications can also achieve a better speech flow output by adopting the PG

framework. For speech synthesis, this framework can better specify how phrasal intonation can be adjusted and modified as a component within a larger prosody unit, an issue which is directly related to output naturalness. For speech recognition, the perceptual relationship between prosodic organization, semantics and speaker's intentions is an issue directly related to processing and parsing, which can be dealt with much more successfully using PG-based chunking.

Accepting and incorporating a PG-oriented framework is only the first step. The mismatches in Experiment 1 and the non-overlap in Experiment 2 both raise issues of variation in individual speakers' planning and intention. It is obvious that individual speakers tend to plan speech output somewhat differently. Speech technology development would be improved by a deeper understanding of this issue. In other words, inter-speaker variations deserve more attention than they have been getting so far, and may be better integrated using principles of general cognition. Our preliminary results already indicate that there is little overlap among speakers in their assignment of boundaries in actual production, although each speaker operated with the PG framework. It is quite obvious that these variations are significant contributors to the overall rhythm of speech output. For computational and technological applications, the canonical form of the proposed PG might serve as a base form for prosody generation. In future research, it would be desirable to pinpoint which issues are planning and intention related. The contribution of other factors, such as tone of voice, should also be investigated.

Experiment 3 shows the underlying patterns of temporal allocation and the ways in which knowledge of these patterns has furthered our understanding of temporal arrangement across speech flow. What the earliest ToBI system saw as rate of speech proved to be much more complex. We also need to more precisely characterize duration adjustment across speech flow to provide a better account of both tempo and rhythm. Being able to specify temporal allocation across speech flow would certainly benefit any framework of prosodic organization. The results of Experiment 3 showed that the constraints on different levels of prosody, tested in layers, did interact with syllable durations, as did the position of prosodic units within a PG. In short, the Mandarin Chinese data showed that temporal arrangement is a derived outcome of syllables interacting with each level of prosodic layering, rather than being solely determined at the level of lexical stress. Duration adjustments that could not be accounted for at lower prosody layers could be accounted for at higher layers, offering support for the following points: (1.) Since Mandarin Chinese is a syllable-timed language, temporal distribution of syllable duration should be with calculated with respect to the level of prosodic organization that groups individual syllables into prosodic units, instead of being classified in terms of other units, such as words, phrases and sentences. (2.) Different speech rates may have an influence on the interaction effect between syllables and higher-level prosodic units. (3.) Trade-off effects were found between prosodic levels and syllables, which accounted for the final output durations. Figure 7 provides an example of the influence of the highest prosodic level on final shortening of syllable duration. However, since the cumulative lengthening from the lower layers was greater than the shortening at the highest layer, the PG-final prosodic phrase maintained

its final lengthening. (4.) A hierarchical organization functions as a set of constraints on speech production, indicating that a possible optimization schema may underlie speech production planning.

4. Conclusion

Up to this point, research seeking to describe and predict Mandarin prosody has focused most of its attention on the intonation of phrases or sentences in isolation. For example, Lin (2002) found that isolated yes-no questions were produced in an overall higher register than their isolated declarative counterparts, a finding which has yet to be tested on fluent speech data. Studies of tonal coarticulation, such as Xu (2002), described anticipatory effects found in tone concatenation at the single-sentence level, but it remains to be seen how these effects interact with higher prosodic levels in fluent speech materials. These studies have yielded detailed information about intonation in sentences of 10 syllables or less, which were produced in isolation, under the tacit assumption that fluent speech would be a concatenated version of such sentences.

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

The present study demonstrates that: (1.) Larger units for overall speech output planning must be taken into consideration, (2.) Phrasal intonations in tone languages are not as significant as they are in intonation languages, unless their positions are specified within a given PG. (3.) Patterns of temporal allocation, speech rate and rhythm are also related to prosodic organization, and some of their characteristics can be predicted by examining higher-level prosodic patterns. Information on temporal allocation and duration adjustment is also fundamental to understanding how to best represent Mandarin fluent speech.

Moreover, the PG model offers a viable framework for formulating theories of prosodic organization in other syllable-timed languages. Future studies should also include investigations of intensity in relation to prosodic organization and global planning. As for technological and computational applications, we predict that implementation of a modified version of this framework into current speech synthesis models would result in a better quality, more natural speech output.

5. References

Bratman, M. “Intention and Personal Policies” CSLI Report (No. CSLI-88-118), 1988

Fujisaki, H. “Modeling in the study of tonal feature of speech with application to multilingual speech synthesis.” *Joint International Conference of SNLP-Oriental COCOSDA 2002*, pp.D1-D9, Prachuapkirikhan, Thailand, 2002 (Invited papers)

Fujisaki, H., S. Ohno and O. Tomita “On the levels of accentuation in spoken Japanese” *Proceedings of 1996 International Conference on Spoken Language Processing*, vol. 2, pp. 634-637, Philadephia, USA 1996

Fujisaki, H., S. Ohno and C. Wang “A command-response model for F0 contour generation in multilingual speech synthesis”, *Proceedings of the 3rd ESCA/COCOSDA International Workshop on Speech Synthesis*, pp. 299-304, Blue Mountain, Australia, 1998

Jokisch, O., H. Mixdorff and U. Kordon “Learning the parameters of quantitative prosody models.” *Proceedings of 2000 International Conference on Spoken Language Processing,*, vol. 1, pp. 645-648. Beijing, China, 2000

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

Keller. E., and B. Zellner Keller, “How much prosody can you learn in twenty utterances?” *Linguistik Online*, 17, 5/03, 2003

Lieberman, P. “Phonetic features and physiology: a reappraisal”, Journal of Phonetics, 4, pp. 91-112, 1976

Lieberman, P. *Intonation, Perception and Language*, MIT Press, Cambridge, MA

Lin, M-C. “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, 2002

Jokisch, O., H. Mixdorff and U. Kordon “Learning the parameters of quantitative prosody models” *Proceedings of 2000 International Conference on Spoken Language Processing,*, vol. 1, pp. 645-648. Beijing, China, 2000

Lehiste, I., and W. S-Y Wang. “Perception of sentence boundaries with and without semantic information”, in *Phonologica* 1976, *Innsbrucker Beitrage zur Sprachwissenschaft*, edited by W. U. Dressler and O. E. Pfeiffer (Institut fur Sprachwissenschaft der Universitat Innsbruck, Austria), pp. 277-283, 1977

Lehiste, I. “Perception of sentence and paragraph boundaries”, in *Frontiers in Speech Perception*, edited by B. Lindblom and S. Ohman (Academic, London), pp. 191-201, 1979

Mixdorff, H., “A novel approach to the fully automatic extraction of Fujisaki model parameters.” *Proceedings of ICASSP 2000*, vol. 3, pp.1281-1284, Istanbul, Turkey, 2000

Mixdorff, H. “MFGI, a linguistically motivated quantitative model of German prosody.” *Improvements in Speech Synthesis*, E. Keller, G. Bailly, A. Monaghan, J. Terken and M. Huckvale (Ed.), Wiley Publishers, pp.134-143, 2001

Thorsen, N. “Intonation and text in Standard Danish”, Acoust. Soc. Am, 77(3), pp. 1205-1216, 1985

Tseng, C. and F. Chou, “Machine readable phonetic transcription system for Chinese dialects spoken in Taiwan”, J*ournal of the Acoustical Society of Japan*, (E), 20, 3, pp. 215-223, 1999

Tseng, C. and F. Chou, “A prosodic labeling system for Mandarin speech database”, *Proceedings of the XIV International Congress of Phonetic Science*, Aug.1-9, 1999, San Francisco, USA, pp2379-2382

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

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

Tseng, C. and Y. Lee, “Speech rate and prosody units: Evidence of interaction from Mandarin Chinese”, Speech Prosody 2004, 23-36 March, Nara, Japan

Tseng, C. and S. Pin, “Mandarin Chinese prosodic phrase grouping and modeling—Method and implications”, International Symposium on Tonal Aspects of Languages—with Emphasis on Tone Languages (TAL 2004), 28-30 March, 2004, Beijing, China

van Santen, C., C. Shih and B. Mobius “Intonation” *Multilingual Text-to-Speech Synthesis: The Bell Labs Apporach*, R. Sporat Ed. Kluwer Academic Publishers, Ch. 6, pp.141-1901998

Wang, C., H. Fujisaki, R. Tomana and S. Ohno “Analysis of fundamental frequency contours of standard Chinese in terms of the command-response model and its application to synthesis by rule of intonation”, *Proceedings of 2000 International Conference on Spoken Language Processing*, vol. 3, pp. 326-329. Beijing, China, 2000

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

Zellner Keller B and E. Keller, “Representing Speech Rhythm” Improvements in Speech Synthesis. pp. 154-164. Chichester: John Wiley. 2001