Mandarin Chinese Prosodic Phrase Grouping and Modeling—Method and Implications

Chiu-yu Tseng & ShaoHuang Pin

Institute of Linguistics, Academia Sinica
Taipei, Taiwan 115
cytling@sinica.edu.tw
brian@phslab.ihp.sinica.edu.tw

Abstract

One major feature of the prosody of Mandarin Chinese speech flow is prosodic phrase grouping [1, 2, and 3]. Phrasal and sentential intonations are governed by a prosody framework that structurally group phrases into a larger/longer and identifiable unit. An overall prosody pattern of such phrase grouping with prosodic specifications is superimposed on phrase group. In turn, individual phrasal intonation under prosody grouping has to adjust in accordance with structural specification from the prosody framework. The output is then seen as derived outcome. The aim of the present paper is to experiment how to simulate prosodic phrase grouping using the Fujisaki intonation model that originally specifies only phrasal or sentential intonations, and how such an intonation model can be further enhanced by incorporating prosodic specifications such as boundaries and breaks, prosody levels/layers and phrase positions under the notion of phrase grouping. The experiments began with aligning the phrase command of the intonation model to boundaries of breaks in the speech flow, then examining prosodic characteristics such as relative position of the target phrase within a prosodic phrase group. Finally, using a linear regression model to predict prosody output from prosodic words upward, predictions of an overall pattern for prosodic phrase grouping was derived. The pattern matched with a prosody base form aimed at prosodic phrase grouping; it also accounted for how and why phrasal intonations were modified in relation to prosody organization. Hence, phrasal intonation is seen as components of prosodic phrase grouping.

1. Introduction

The central issue for prosody of speech flow is most notably what constitutes an overall pattern of natural connected speech. One most pronounced feature of Mandarin Chinese speech flow is paragraphing, or, the chunking of phrases and grouping of them into larger prosodic units. We believe that an organization of paragraphing and its corresponding prosodic characteristics are key to understanding how speech flow is structured instead of treating speech flow as concatenation of unrelated phrasal intonations. Such paragraphing involves grouping of phrases that are not only prosodically structural and representative, but also perceptually identifiable. Within such a prosodic group (PG) [1, 2 and 3] of multiple phrases, each phrase could then be further specified in relation to their PG related positions, and a canonical base form of prosody organization was proposed [4]. In short, a PG can be seen as the highest node of a prosodic hierarchy that branches into prosodic levels or layers. That is, a PG branches into UTR's followed by B4's (or BG's), UTR into prosodic phrases (PPh's) followed B3's, PPh's into PW's separated by B2's; PW's into syllables that correspond to individual characters in the Chinese orthography. For F0 contour patterns, a PG is characterized by two resets and F0 peaks (PG initial and PG final), a terminal trailing off and F0 fall (PG final) and units separated by breaks. Or, from the perspective of prosodic units, a PW is followed by B2, PPh by B3, UTR by B4 and PG by B5. Corresponding temporal allocation and distribution are also systematic [4, 5]. A combination of F0 modification and rate allocation should constitute a better framework of speech prosody other than simple concatenation of phrasal intonations into strings. The focus of the present study is to show how to organize phrasal intonations into prosodic grouping that reflects paragraphing using the Fujisaki intonation model, while corresponding studies of speech rate is discussed elsewhere [4, 5].

2. Speech Material:

The speech material used in the study consists of two parts. Part 1 included 133 prosodic groups in 25 high quality speech files or read speech recorded in professional studio by a female Mandarin broadcast speaker at mean syllable duration of 195ms. The minimum number of syllable in a paragraph is 5, whereas the maximum 229. The mean length of these perceived prosodic groups (PG) is 85 characters/syllables; the mean number of prosodic phrases within a PG is 11.7. Part 2 included 319 prosodic groups in 319 selected files from our 599 paragraph database [1], again in high quality speech files of read speech recorded in our lab's sound proof chamber by a different female Mandarin speaker at mean syllable duration of 209ms. Each file is tagged using the same notations as files used in Part 1. The minimum number of syllable in a paragraph is 6, whereas the maximum 178. The mean length of these perceived prosodic groups (PG) is 48 characters/syllables; the mean number of prosodic phrases within a PG is 9. All of the speech files were first labeled by the HTK automatic alignment tool for phonetic transcription and then tagged for boundary prosodic information manually. All the results of HTK alignment tags were also adjusted manually. We noted that the speech data from the second speaker was perceived as slower than data from the first speaker due to more and longer pauses in the speech flow.

3. Analysis

The analysis procedure consists of two parts: (1.) the extraction on model parameters and (2.) the statistical characteristics on these parameters.

Part (1) involved an optimization process to extract parameters needed by the model. Automatic parameter extraction using the Fujisaki model has been reported in many studies; one of the algorithms is based on filtering the f0 contours and has been implemented in German, Vietnamese, Thai, and Mandarin Chinese [6-11]. However, our prosody

framework specifies intonation in relation to their roles and positions within a PG [4]. However, instead of the interpolation and detection procedures described in [6], we located phrase commands by detecting resets and pauses of pitch contours, and accent commands by giving each syllable an initial accent command. For phrase commands, the onset time is thus related to the prosodic boundaries; for accent commands, the onset time T1 and the offset time T2 are assigned to the syllable's voiced onset time and offset time. Based on this criterion, an initial seed model is assigned to each paragraph. An optimization procedure based on [6] was then implemented to adjust the modeled contour closer to the original sample from the speech data.

The Fujisaki model parameters used in optimization are phrase and accent command magnitude. The position of our phrase command stands for a pitch reset event, and the magnitude of it represents the degree of the reset; the position of our accent command stands for the portion of the syllable's voiced part, and its magnitude (both positive and negative are allowed) represents the combination of tone identity, accentuation, or local variation on the overall intonation.

An example of modeled pitch contour after optimization is shown in Figure 1. Since our purpose here is to demonstrate only the global overall contour patterns across phrases, detailed local variations were not specified. We chose to use the phrase commands to represent prosodic phrase grouping that reflects scope and units of speech planning and target shooting towards a terminal fall of intonation, as phrase command is the result derived from physiological considerations [9].

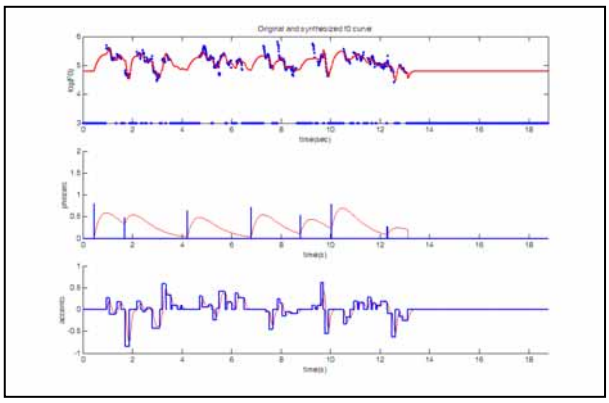

Figure 1*: An example of extracted Fujisaki models, and modeled pitch contour.*

Part (2) aimed to derive statistical characteristics after optimization. The first step was to align each optimized phrase command to a labeled boundary in the speech flow before looking for reset phenomena in the context of boundaries, and the distribution of phrase command magnitude in each boundary. Three sets of statistical analyses were performed. The first analysis derived the means and distribution of phrase commands magnitude. The second was a simple ANOVA looked for PG position related patterns. A PG is defined into three relative positions, namely, PG-initial, PG-middle and PG-final, to see if position bears any prosodic characteristics. The first prosodic phrase into a PG, separated by the first B3, is denoted by I, the final prosodic phrase separated by the last B3 is denoted by F, while all other prosodic phrases separated by other B3's are classified into M. The third was a linear regression analyses that looked for phrase-grouping effects in terms of PG-related positions and magnitude.

4. Results and Discussion

4.1. Alignment of prosodic boundaries:

A hierarchical prosodic organization on the basis of analyses of speech data have been proposed earlier [3, 4]. The highest node of the organization PG consists of grouping of prosodic units and their respective characteristics. PG includes groups of UTR's; UTR groups of PPh's; PPh groups of PW's; and PW groups of syllables. All of these prosodic units are separated by different boundary index and following pauses/breaks although boundaries at the lower levels such as syllables and PW's may not be followed by pauses. However, higher prosodic boundaries require following breaks for physiological reasons. B5 is the longest and PG-final boundary, B4 the breath-group boundary, B3 the prosodic phrase boundary, B2 the prosodic word boundary. We then built an alignment procedure between phrase commands and these boundaries. The alignment was to find the nearest phrase command before the prosodic unit started, as shown in table 1. The distribution of magnitude of phrase command (AP) by speaker is shown in Figure 2. Both speakers exhibited similar patterns of boundary distribution, i.e., positive correlation between boundaries and phrase commands.

Table1*: The relationship of each prosodic boundaries and extracted phrase command in data-F03, F051.*

F03	B2	B3	B4	B5
Phrase cmd.	369	1983	112	319
Total break	3806	3288	119	319
Rate (%)	9.7	60.3	94.1	100.0
Avg. Ap	0.47	0.62	0.81	0.88
F051				
Phrase cmd.	266	433	77	70
Total break	1913	776	95	92
Rate (%)	13.9	55.8	81.0	76.1
Avg. Ap	0.69	0.85	0.98	1.00

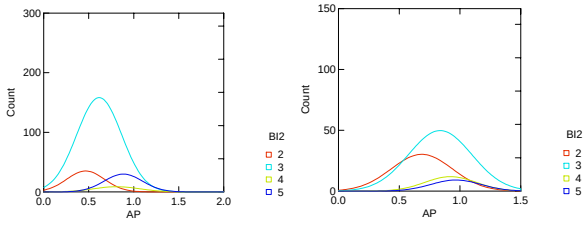

Figure 2*: Histogram of Ap distribution with respect to t boundaries, the left panel shows statistical results from speaker F03, the right panel F051.*

The results showed that most of the higher-node boundaries were aligned to phrase commands; greater boundaries tended to align with greater phrase commands where clusters of Ap's were also formed. The positive correlation between phrase commands and boundaries indicated clearly that the phrases under consideration bore some kind of structural relationship with each other. Clustering of Ap's also indicated possible

correlations among the phrase commands. By incorporating boundaries into the intonation model and looking for Ap related patterns, structural relationships were found. These relationships could be seen as evidence of a higher prosody organization in operation. In other words, no such patterns would emerge if phases were considered independently. Thus it became clear that phrases within a PG bore prosodic structural relationships to boundaries rather than independent units of intonation separated by pauses.

However, individual variations were found between the two speakers. As in both sets of speech data, the mean value of Ap is in the descending order from B5 to B2, and the value in F051 shows much flatness in the intonation than in the F03. The ratio of aligned phrases and B5 is 1 for speaker F03, as shown in Table 1. However, the same ratio is down to 0.76 for speaker F051. This is largely due to the fact that in F03's speech data, clear pauses in each boundary above B3 existed; whereas in F051's speech data, we did not find such clear pauses. The latter indicated that boundaries may very well be an integrated outcome of acoustic modifications and following pauses, and we would need to investigate more on pre-boundary related acoustic characteristics of the speech signals themselves. We expect to find more speaker dependent variations that may bear significance to different speaking habits. For the time being, we included both of these materials here as a mutual reference. In the case of B3 and B2, the ratio for both data are in 0.56~0.60, and 0.9~0.14 respectively.

4.2. Grouping of phrases by PG positions:

In addition to boundary features, we also analyzed the phrase commands in relation to PG positions to see if position bore prosodic correlation to phrases that were governed under a PG. That is, whether phrase commands exhibited different patterns accordance to their respective positions within a PG. To do so, we first defined a PG by three relative positions, namely, PG-initial, PG-middle and PG-final, to see if position bore any prosodic characteristics. By PG-initial (I), we meant the first PPh into a PG separated by the first B3; by PG-final (F) the last PPh before the end of a PG preceded by the last B3. All other PPh's between were termed as PG-middle (M). ANOVA was performed on phrase commands with respect to PG positions to see if significant differences were found. Table 2 summarizes the results.

Table 2: *ANOVA of phrase commands in three relative positions. I (PG-initial), M(PG- middle) and, F(PG-final) by speaker.*

		I	M	F	F-ratio
F03	Ap	0.87	0.62	0.40	257.5
	# of Ap	331	2231	221	
F051	Ap	0.90	0.84	0.61	43.9
	# of Ap	102	658	86	

Tables 2 showed that PG-initial phrases possessed large magnitude than other positions, and PG-final phrases smaller magnitude of phrase command than PG-initial ones. The F-ratio in the table shows the degree of significant difference among the three positions. The results showed that phrase commands under each PG position were significantly different from each other, and the difference was more evident in speaker F03, as shown in the upper panel of Table 2. More difference was found between I and F, further demonstrated that PG positions were more significantly different between I and F. Further studies of the acoustic properties of the speech data with respect to these findings should provide more evidence from the speech data. By incorporating PG positions into the intonation model and looking for Ap related patterns, more structural relationships were found. These relationships were also seen as supporting evidence of a higher prosody organization in operation that phrase groups within a PG abide by.

However, the value of the F-ratio was greater in F03 than F051, which we again attribute to more and longer pauses. In other words, we see pauses as a necessary feature to prosodic boundaries, and possibly in compliment with specification of pre-boundary acoustic features.

4.3. Clustering of Phrase Commands by Magnitude

In the Fujisaki model, each component of the intonation is a superposition of the responses of all phrase commands. The canonical intonation form specifies a decrease of the magnitude of the phrase commands along the time domain as the phrase approaches its end, where the intonation contour would also decline unless otherwise specified. But this kind of specification could not accommodate why sometimes the contour did not decline to a terminal fall, or rather, how the magnitude of the phrase commands varied, as our data demonstrated. Our prosody framework specifies a higher governing constraint that groups phases into PG, hence we were interested to see if evidence of the higher node could be found through analyses of magnitude. Using a simple general linear model, we tried to tease apart possible contributions from the higher node from those of the lower level phrases through the residues and coefficients. Our approach was to cluster phrase commands by magnitude to remove contributions from the lower level, then further analyzed the residues with respect to PG positions of the PPh's to see if positive correlations could be obtained.. Table 3 summarizes the predictions of the clusters, while Figure 3 shows prediction from higher level information, in particular with respect to PG's of 7, 8 and 10 phrases.

Table 3: *Coefficients from a linear regression model on clusters of phrase commands (Ap) by magnitude.*

Phrase order	Data	1	2	3	4
Coefficients	F03	0.80	0.51	0.42	0.38
	F051	0.989	0.692	0.607	0.519
# of Ap	F03	1360	974	341	92
	F051	422	292	96	28

.

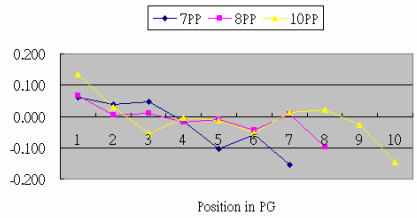

Figure 3*: PG patterns of which contain 7, 8, and 10 number of prosodic phrases (F03)*

Figure 3 can be seen as a representation of the overall pattern of a multiple-phrase PG. Note how the pattern can also be characterized in terms of positions. Like an intonation pattern, this overall pattern can be viewed as a superimposed frame on phrases groups, thereby causing phrasal intonations within to

modify accordingly. The pattern shown in Figure 3 also coincided with the basic canonical form we proposed [4] as shown in figure 5.

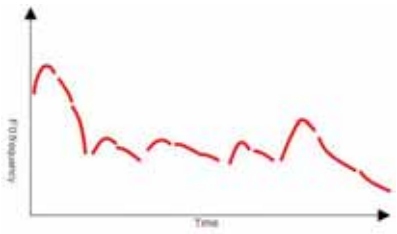

Figure 4: *Base form F0 contour patters of PG (Tseng, 2004)*

Since each such cluster also represented a F0 reset, we were further interested to see whether some kind of correlation could be found between the clusters and boundaries between, using also earlier results obtained from alignment of prosodic boundaries (See 4.1). Table 4 summarizes the percentage of overlap between results of pause/break analyses in relation to F0 reset.

Table 4*: Percentage of overlap between F0 reset by clusters of phrase commands and boundaries.*

	F03				F051			
	B2	B3	B4	B5	B2	B3	B4	B5
%	22	44	85	99	28	52	81	86

Results from Table 4 of the higher overlap between clusters and B4, B5 show that the bigger the break/pause is, the higher the overlap is found, indicating a greater pause in the speech flow is more likely to be followed by higher F0 reset. Again, explanations could be found from the physiological aspects in relation to pulmonary air flow and energy needed for speech production. In short, the higher the F0 reset is, the more energy it requires, and hence the more time before both for breathing and for articulation maneuvers to take place.

5. Conclusion

Prosodic phrase grouping constitutes the most important feature of Mandarin Chinese speech flow. A prosody framework should accommodate and account for the phenomena. Our prosody framework states that (1.) larger units consisted of multiple phrases that imply overall planning of speech output could be identified. (2.) Phrasal intonations should be further specified in relation to prosody organization. (3.) PG related characteristics could be derived with respect to overall pattern, position, reset, temporal adjustment and intensity. As a consequence, how to best simulate such grouping merits experimentation. We presented initial analyses and examination of the prosodic-phrase-grouping phenomena using the Fujisaki intonation model to show how such grouping not only exists, but also possesses a canonical form. The canonical form can be viewed as a planning unit for speech flow that governs and constrains modification of phrasal intonations within. Evidence of modifications of phrasal intonations in accordance with PG positions was obtained. Statistical analyses of clusters by magnitude of phrases commands also showed how prosody levels and layers interacting with phrasal intonations and an independent base form of PG could be derived. At the same time, we also showed how an intonation model could be adapted to incorporate prosodic phrase grouping. In so doing, we believe we have learned more about prosody structure and organization in general, and Mandarin related phenomena in particular. Future attempts should focus on at least two aspects, namely, (1.) how to modify accent command in the model to accommodate prosody related characteristics in addition to intonation related ones, and (2.) how to apply such modeling to speech synthesis and TTS.

6. Reference

[1] Tseng, C; Chou, F., 1999. A prosodic labeling system for Mandarin speech database. In *Proceedings of ICPhS 2003*, San Francisco, USA, 2379-2382.

[2] Tseng, C., 2002. The prosodic status of breaks in running speech: Examination and evaluation. *Speech Prosody 2002*, Aix-en-Provence, France, 667-670.

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

[4] Tseng, C.; Pin, S.; Li, Y., 2004. Mandarin Speech Prosody: Issues, Approaches and Implications. (to appear in *Festschrift celebrating the 90th birthday of Professor Wu Zontji, 2004*)

[5] Tseng, C.; Li, Y., Speech Rate and Prosody Units: Evidence of Interaction from Mandarin Chinese. *Speech Prosody 2004*, Nara, Japan.

[6] Mixdorff, H., 2000. A novel approach to the fully automatic extraction of Fujisaki model parameters. In *Proceedings of ICASSP 2000*, Istanbul, Turkey, 1281-1284.

[7] Mixdorff, H.; Fujisaki, H., 1999. The Influence of Focal Condition, Sentence Mode and Phrase Boundary Location on Syllable Duration and the F0 Contour in German. In *Proceedings of the ICPhS*, San Francisco, USA, 1537-1540.

[8] Fujisaki, H., 2002. Modeling in the study of tonal features of speech with application to multilingual speech synthesis. *Joint International Conference of SNLP-Oriental COCOSDA 2002*, Hua Hin, Prachuapkirikhan, Thailand.

[9] Mixdorff, H.; Fujisaki, H., 2000. Symbolic versus quantitative descriptions of F0 contours in German: Quantitative modelling can provide both. In *Proceedings of Prosody 2000*. Kraków, Polen.

[10] Mixdorff, H.; Fujisaki, H.; Chen, G.; Hu, Y., 2003. Towards the automatic extraction of fujisaki model parameters for mandarin. In *EUROSPEECH 2003* Geneva, Switzerland, 873-876.

[11] Hirose, K.; Furuyama, Y.; Narusawa, S.; Minematsu, N.; Fujisaki, H., 2003. Use of linguistic information for automatic extraction of F0 contour generation process model parameters. In *EUROSPEECH 2003*, Geneva, Switzerland, 141-144.

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

[13] Tseng, C., 2002. The prosodic status of breaks in running speech: examination and evaluation, *Speech Prosody 2002*, Aix-en-Provence, France, 667-670.