Modeling Prosody of Mandarin Chinese Fluent Speech via Phrase Grouping

Chiu-yu Tseng & ShaoHuang Pin
Phonetics Lab, Institute of Linguistics
Academia Sinica, Taipei, Taiwan 115
cytling@sinica.eud.tw

Abstract

We have proposed that Prosodic Phrase Grouping (PG) best characterize the prosody of Mandarin Chinese fluent speech [1]. PGs reflected a higher semantic and cognitive unit of speech planning in discourses. Corresponding prosodic characteristics were obtained, demonstrating how prosody of fluent speech was organized. PG-related global intonation and duration patterns [1, 2, and 3] indicated that phrasal and sentential intonations were subordinate prosodic units and modifications are required. We will show how we model and simulate the global PG intonation on top of the Fujisaki model [4], a physiologically based phrasal intonation model. We believe capturing the PG effect helps understand prosody of fluent speech, and simulation of this kind could directly improve output naturalness of unlimited TTS. Our methods included quantifying PG related F0 characteristics from speech corpus with commands from the Fujisaki model, and subsequently utilizing these features as variables to predict prosody of Mandarin fluent speech.

1. Introduction

In our previous studies, we performed extensive acoustic analyses of spoken discourses and showed that phrase grouping was essential to characterize the prosody for Mandarin fluent speech [1]. Evidence of prosodic phrase grouping (PG) was found both in adjustments of F0 contours [2] and temporal allocations within and across phrases [3]. A canonical base form of PG, specifying trajectories of a series of F0 contours, was postulated to account for the F0 patterns of related multiple phrases in fluent speech [1, 2, 3]. Figure 1 is a schematic representation of multiple-phrase intonation of a PG. The most important F0 features included an initial rise, non-terminal fall for intermediate phrases, and a secondary rise into the last phrase, followed by a terminal trailing off to the lowest. Under this framework, phrases are no longer unrelated intonation units, while degrees of boundary breaks (shown in Figure 1 as empty intervals between contours) also require specification.

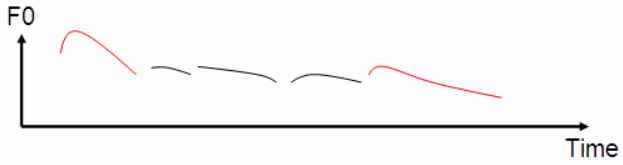

Figure 1. A schematic illustration of F0 contours within a PG base form. The PG-initial Prosodic Phrases (PPh) and PG-final PPh are in red; while PG-medial PPhs are in black.

Our PG framework can also be viewed a tree-branching hierarchical organization that groups phrases as shown in Figure 2.

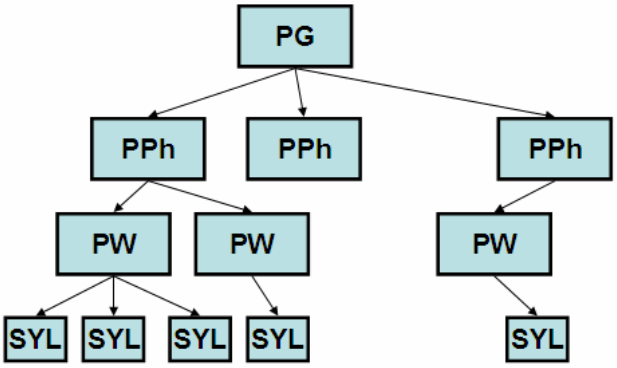

Figure 2. A hierarchical organization of phrase grouping, where PG denotes a Prosodic Group, PPh as a Prosodic Phrase, PW as a prosodic word and SYL as syllable.

Note that the proposed prosody hierarchy is layered, assuming a governing relationship from the higher nodes. Higher levels of prosody information are superimposed onto lower units, while corresponding boundaries and breaks [5, 6] also occur. Most importantly, the highest layer, namely PG, assumes a top-down window or projection of a speech planning unit above phrases.

Our current goal was to see (1) if PG could be added to an existing layered phrasal intonation model, (2) if PG could be implemented into building a layered model to predict global prosody, and (3) if prosody of Mandarin Chinese fluent speech could be predicted using the finding and patterns obtained.

2. Building intonation models on the Fujisaki model

The aim at this stage was to first build data-driven prosody models that group phrases together from data analyses. The corpus used was female read speech data of 26 long paragraphs or discourses in text, or a total of 11592 syllables (or Chinese characters). These speech data were first semi-automatically aligned with initial and final phones using the HTK tool-kit, and then manually labeled by trained transcribers for perceived prosodic boundaries and breaks/pauses [1]. A total of 136 PGs and 1253 prosodic phrases (PPh) were identified and labeled. The mean number of PPhs by PG is 10, indicating that an average of 10 prosodic phrases made up a prosodic phrase group. The intonation patterns were built in two steps. The first step was to extract parameters that characterize intonations with the Fujisaki model; the second step was to build statistical models of intonation predict intonations.

Figure 3 displays the block diagram of the Fujisaki model [4]. The model is able to produce F0 contours of phrases with reasonably good approximations to original speech data from two kinds of commands, namely, phrase and accent commands, with the base frequency Fb. The phrase commands are impulses represented by magnitude parameter Ap, timing parameter T0 and time constant α The accent commands are stepwise functions represented by magnitude parameter Aa, timing parameters T1 and T2 and time constant β. The base frequency is a speaker-dependent asymptotic value of F0 in the absence of accent commands.

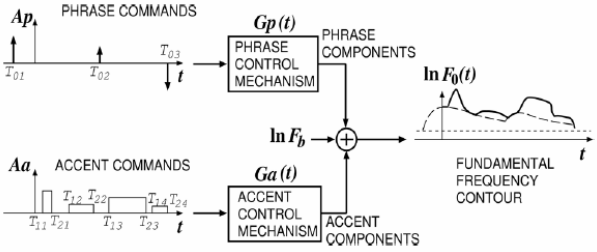

Figure 3. Generation process of the Fujisaki model (from Fujisaki, 1982)

The model connects the movements of cricoid's cartilage to the measurements of F0 and is hence based on constraints of human physiology. Therefore, it is reasonable to assume that the model could accommodate F0 output of different languages. In fact, Fujisaki and Mixdorff have already tested the model with many languages successfully [7, 8]. In the case of Mandarin Chinese, phrase commands were used to produce intonation at the phrase level while accent commands were used to predict lexical tones at the syllable level [9]. Phrasal intonations are superimposed on sequences of lexical tones. Therefore, interactions between the two layers cause modifications of F0 to produce the final output. The superimposing of a higher level onto a lower level leaves room for even higher level(s) of F0 specification to be built. Thus, we decided to test 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 F0 patterns of phrase grouping could be achieved.

2.1. Parameter extraction

F0 contours of speech data were first smoothed, then quadratic spline interpolated, and finally separated into two parts [10]. Using information from labeled boundaries as reference of prosodic phrases, we first assigned phrase and accent commands, then optimized them according to the interpolated and separated f0 contours of each phrase and accent. Figure 4 shows an example of the automatic extraction results.

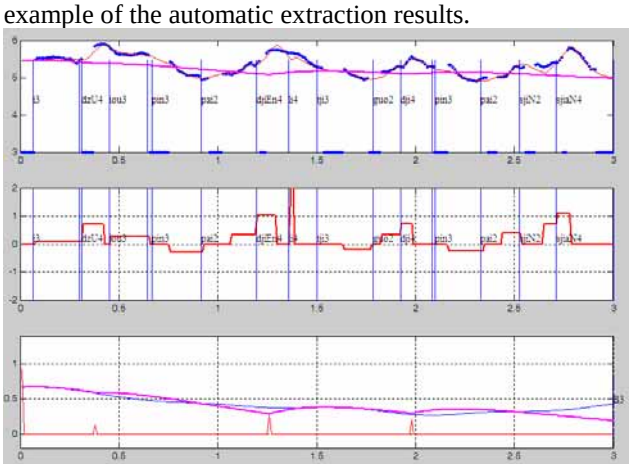

Figure 4. F0 commands extraction and boundary information. (Top panel: segmental boundary information, observed f0 (blue dot), extracted phrase component (magenta line), modeled f0 (red line), middle panel: accent commands, bottom panel: phrase commands)

The extracted commands were then converted in relation to syllable timing. Table 1 shows the conversion table from the Fujisaki model to syllable time.

Table 1. Conversion from the Fujisaki timing parameters to relative syllable timing parameters.

symbol	meaning
Rel_T0	Syl_{on} – T0
Rel_T1	T1 – Syl_{on}
Rel_T2	T2 - Syl_{on}

2.2. Statistical model building

The phrase command in the Fujisaki model is specific for modeling intonation of a single phrase, rendering a gradual asymptotic trajectory for a relatively short time frame. Our goal is to build a prosody model for a succession of related phrases with specifications from PG. Therefore, more than one phrase command is needed across phrases to achieve the grouping effect, especially the features described in Figure 1. We utilized the labeled breaks in our corpora to denote a possible phrase boundary and repeated application of the phrase commands until the designated last phrase. Since the length of our PPhs varied considerably, more than one application of the phrase commands were also necessary for some PPhs.

Figure 5 shows three different Ap values and their respective responses when α is 2 /s. The three values illustrated that the succession of phrase commands within phrases could not be placed too far from each other on the time domain, and more than one phrase commands may be needed to model the actual contour. In addition, in order to model larger unit than phrase, we chose a heuristic value 1 second for a maximum distance between phrase commands to model the non-terminal intonations between the PG-initial PPh and PG-final PPh.

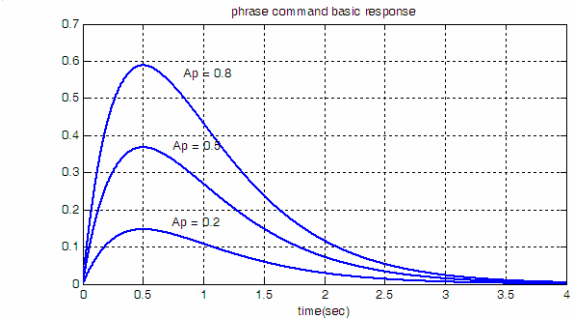

Figure 5. Response of Phrase commands from three different Ap values. Ap 0.5 is the mean value in the speech data while the other two are applied deviations.

Following our PG framework, statistic models were established at two separate layers to predict intonation contours at two different levels. At the first and lower layer, individual PPh intonations corresponding to phrasal intonations were modeled using all of the available local information, namely, pauses, position in PPh, previous phrase commands etc. At the second and higher layer, PG intonation corresponding to global contour patterns was modeled using the residuals from the PPh layer. Thus, when a string of phrases were specified as a PG, boundary positions and the aforementioned heuristic rule were first used to decide the position of phrase commands. Then the first/lower layer of statistical model was applied to make a decision for all the magnitudes of phrase commands. Sections 3 reports statistical models and experiments of the predictions from the models to the original speech data, and our explanations of the findings.

3. Testing of Proposed Models

The following statistical models for individual local intonations and global phrase-group intonations were tested to see if predictions could be made satisfactorily.

3.1. The Phrasal intonation model

A general linear model of phrasal intonation corresponding to the PPh layer was built using independent variables shown in Table 2.

Table 2. Variables used in PPh layer

meaning	symbol
pause before phrase command	pause
previous phrase command	pre_phr
fujisaki parameter : base f0	f0min
index of syllables in PPh	iSyl_PPh

The model would be used to model individual phrasal intonations. Modeling outcomes were used to represent phrasal intonations in each PPh, seen as local and micro intonations in our PG framework.

3.2. The PG intonation model

Another linear model needed to be built to account for the global PG intonation that governed the phrases it grouped. However, unlike the phrasal intonation model, two alternative assumptions were considered and tested. One assumed that each PPh interacts with the higher level PG independently; each underwent significantly different modifications from the other. Under this assumption, each PPh represented an index as input information to build PG intonation, and all indices were used. The other assumption was that information of PG intonation affects the phrasal intonations by positions, and only three relative positions were needed, namely, PG-initial, PG-medial and PG-final. Note that PG-initial and PG-final indicate the initial and final PPh of a PG whereas more than one phrase may occur in the PG-medial position. In this case, only three indices were necessary to predict the global PG intonation contour. These two possible models for the PG intonation were built and tested.

3.3. Testing the phrasal intonation model

Our PG framework assumes that the phrasal linear model would account for the intonation of each PPh superimposed on lower units (PWs) before it interacts with the higher level PG intonation. We used Fujisaki parameters extracted from the speech material as independent variables presented in table 2 to model PPh intonations as follows:

$$\textit{Phrase command Ap} = \textit{constant} + \textit{coeff1} \times \textit{pause} + \textit{coeff2} \times \textit{pre\_phr} + \textit{coeff3} \times \textit{f0min} + \textit{iSyl\_PPh}$$

Here *coeff1* to *coeff3* represents the coefficients derived from our data and used in the general linear model. The residuals derived meant the portion of the data that could not be accounted for at the PPh level, or, by local phrasal intonation modeling. Under our framework, these residuals may represent influence of information from the higher level. Therefore, the residuals were subsequently moved up to the next higher layer, namely, the PG layer, to test if they could account for the modeling of PG intonation.

3.4. Testing the PG intonation model

Our PG framework assumes that the PG linear model accounts for the global and higher level intonation superimposed on phrases. The question was whether PG effect could be found on each and every PPh, or on three position-related PPhs only, namely, the PG-initial PPh, PG-medial PPh(s) and PG-final PPh. Therefore, two linear models were tested. The first model used all the PPhs as indices whereas the second model used only three independent PG-position variables. ANOVA was performed on training results from the PG linear models.

4. Results and analysis of model testing

At the phrasal intonation level, a correlation of 0.79 was obtained between the modeled results and the data, meaning 62.4% of data variations were accounted for by using these independent variables at the PPh layer. Table 3 shows results of ANOVA of the phrasal model where significant effect on predicting phrase commands was found.

Table 3. ANOVA for phrase model

	Mean-Square	F-ratio	P
Model	0.815	22.696	0.000
Error	0.036	-	-

The results from the correspondence between our prediction and the actual speech data indicated that individual phrasal intonations can only account for 62.4% of the final output of intonation strings in fluent speech. This implies that the prosody of fluent speech is more than concatenating individual intonations into strings.

At the PG level, Tables 4 shows the statistical results of the PG model that used all PPhs as indices.

Table 4. ANOVA of PG model by PPhs

	Mean-Square	F-ratio	P
Model	0.035	1.068	0.159
Error	0.033	-	-

The p-value in Table 4 (0.159) means no significant difference was found by using index of each prosodic phrase in PG.

Statistical analyses were performed to test whether each initial-final PPh pairs in PG were significantly different from each other. Table 5 shows the results.

Table 5. ANOVA for PG model by positions (PG-initial vs. PG-final)

	Mean-Square	F-ratio	P
Model	0.572	12.127	0.001
Error	0.047	-	-

Significant difference was found between initial PPhs and final PPhs. The significance can be derived to a set of coefficients and added as PG effects to the initial and final PPh in the following form. A global PG intonation could thus be built accordingly.

Phrase command Ap = constant + coeff1 ⨯ *pause*
+ coeff2 ⨯ *pre_phr + coeff3* ⨯ *f0min*
+ iSyl_PPh (syl position in PPh)
+ PG effect coefficients (initial, final PPhs)

The results from the above two PG intonation models imply that the PG model specifying PG related positions should be used to predict and later simulate the global PG intonation. Note that the model also characterizes how phrases grouped under PG became related as sister phrases, and where individual PPh intonation within a PG is required to modify accordingly. In other words, the sisterhood among PPhs under a PG could be defined by their respective positions, and concatenation of phrasal intonations is thus structured by specification from PG positions.

The above statistical models were then used to predict intonation contours of PGs. In addition to the same limitation of heuristic rules and procedures of Fujisaki model extraction, the prediction process also needed information of prosodic boundaries as references to the positions of phrase commands, as well as timing information of each syllable. The prediction involved first applying accent command to each syllable for tones [9], then the phrasal intonation model for each PPh, and finally the PG intonation model for the final output contour. Figure 6 shows an example of simulation output in comparison with original speech data. The results indicated that successful prediction of PG intonation patterns can be achieved.

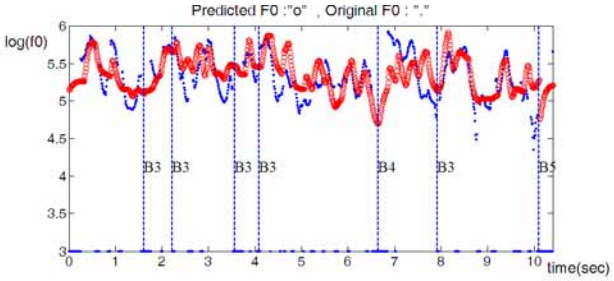

Figure 6. Simulation result of global intonation modeling of a PG. The red line represents simulated global contours; the blue represents contours of the original speech data.

5. Discussion

Phrase group is a well known phenomenon in Chinese and have been quite well researched in semantic stylistic and rhetoric studies [11], sometimes termed as speech paragraphs. By taking a top-down approach to analyze speech data of discourses instead of approaching sentences or phrases individually, speech paragraphs of multiple phrases could be identified consistently via prosodic cues. We believe that these speech paragraphs represent speakers' planning and intentions before and during speech production, and are therefore also likely reflections of cognitive constraints. Understanding and modeling phrase groups are more than necessary to generate prosody of fluent speech for Mandarin Chinese. The present study also shows how PG functions like a superimposed window that governs the overall intonation contours of phrases it groups together.

We believe PG assumes a governing relationship to intonations under it and constrains the global prosody output. Intonation patterns exist at each prosodic layer, with the higher level superimposing onto the immediate lower prosodic units. In turn, phrasal intonations are seen as constituents in a phrase group and thus are required to adjust in fluent connected speech [1]. Different layers of intonations also interact with each other Furthermore, we believe that in fluent speech, phrasal intonations are only significant after their roles within PG are defined [12, 13]. The PG framework not only accounted for prosody of fluent speech, but also explains the inadequacy of taking phrasal intonation as independent prosodic entities.

6. Conclusion

In this paper, we have shown initial attempts to implement a layered prosodic organization that characterizes Mandarin fluent speech. Two linear models representing two upper layers of the prosodic framework were constructed and tested. By providing a higher prosodic layer, concatenating phrasal intonations could be further specified systematically for a global contour, and the prosody of fluent speech is thus better accounted for. This framework should not be limited to Mandarin Chinese only and could also apply to intonations of complex sentences in non-tonal languages such as English. Future directions include building PG-constrained patterns of temporal allocation [2] and intensity distribution into the framework as well. We believe the model and prediction can be applied directly to unlimited TTS to improve prosody output for connected speech as well.

7. Reference

[1] Tseng, C, S. Pin and Y Lee, "Speech Prosody: Issues, Approaches and Implications", in Fant, G., H. Fujisaki, J. Cao and Y. Xu Eds. *From Traditional Phonology to Mandarin Speech Processing*, Foreign Language Teaching and Research Process, Beijing, China, 2004, pp. 417-438.

[2] Tseng, C. and Y. Lee, "Speech Rate and Prosody Units: Evidence of Interaction from Mandarin Chinese", *Proceedings of Speech Prosody 2004*, Nara, Japan, March 23-26, 2004, pp. 251-254.

[3] Tseng, C. and S, Pin, "Mandarin Chinese Prosodic Phrase Grouping and Modeling--Method and Implications", *Proceedings of International Symposium on Tonal Aspects of Languages—with Emphasis on Tonal Languages (TAL 2004)*, Beijing, China, March 28-30, 2004, pp. 193-19.

[4] Fujisaki, H. and Hirose, K. "Analysis of voice fundamental frequency contours for declarative sentences of Japanese". *Journal of the Acoustical Society of Japan (E)*, 5(4): pp. 233-241, 1984.

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

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

[7] Fujisaki, H. "Modeling in the Study of Tonal Feature of Speech with Application to Multilingual Speech Synthesis", *SNLP-O-COCOSDA 2002*, Prachuapkirikhan, Thailand, May 9-11, 2002.

[8] Mixdorff, H., "Quantative Tone and Intonation Modeling across Languages", *International Symposium on Tonal Aspects of Languages With Emphasis on Tone Languages*, Beijing, China, March 28-30 2004, pp. 137-142.

[9] Mixdorff, H., Hu, Y. and Chen, G., "Towards the Automatic Extraction of Fujisaki Model Parameters for Mandarin", *Proceedings of Eurospeech 2003*, Geneva, Switzerland.

[10] 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.

[11] Wu, W. and X. Tian, *Hanyu Juqun*, Shangwu Press, 2000, Beijing, China.

[12] Tseng, C., "Prosodic Group: Suprasegmental Characteristics of Mandarin Connected Speech from a Speech Data Base", ICCL-6, 1997, Leiden, the Netherlands.

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