CHAPTER 2C

PROSODY ANALYSIS

Chiu-yu Tseng

Institute of Linguistics, Academia Sinica, Taipei

E-mail: cytling@sinica.edu.tw

This chapter discusses why Mandarin speech prosody is not simply about tones and intonation, and how additional but crucial prosodic information could be analyzed. We present arguments with quantitative evidences to demonstrate that fluent speech prosody contains higher-level discourse information apart from segmental, tonal and intonation information. Discourse information is reflected through relative cross-phrase prosodic associations, and should be included and accounted for in prosody analysis. A hierarchical framework of Prosodic Phrase Grouping (PG) is used to explain how in order to convey higher-level association individual phrases are adjusted to form coherent multiple-phrase speech paragraphs. Only three PG relative positions (PG-initial, -medial and -final) are required to constrain phrase intonations to generate the prosodic association necessary to global output prosody which independent phrase intonations could not produce. The discussion focuses on why the internal structuring of PG forms prosodic associations, how global prosody can be accounted for hierarchically, how the key feature to speech prosody is cross-phrase associative prosodic templates instead of unrelated linear strings of phrase intonations; and how speech data type, speech unit selection, and methods of analysis affect the outcome of prosody analysis. Implications are significant to both phonetic investigations as well as technology development.

1. Introduction

By definition prosody is an inherent supra-segmental feature of human speech that carries stress, intonation patterns and timing structures of continuous speech (as described elsewhere in this volume). From the supra-segmental perspective, given that Mandarin tones are lexical, then by definition tone is lexical prosody. Given that intonation are syntactic, defining simple phrase and/or sentence types as an intonation unit (IU), by definition intonation is syntactic prosody. Much discussion in the literature has been devoted to tones and intonation as well as their interaction. However, in the following discussion we will argue why fluent speech prosody is not simply about tones and intonation and how fluent continuous speech prosody is in fact discourse prosody that exists in addition to and above tones and intonation. Our aim it to show and how to capture and account for discourse information in addition to tones and intonation when analyzing prosody and the implication of discourse prosody.

What is the role of discourse in speech prosody? Our earlier investigations of Mandarin Chinese fluent spontaneous speech revealed that only 36% of syllables possess one-to-one phonological-to-phonetic correlations,[25] that is, with identifiable tone contours. The results suggest that (1) tonal specifications are not always realized in connected speech, and (2) lexical prosody makes up less than half of the output F0 contours. The same study also compared phrase intonation of identical simple declarative sentences first extracted from spontaneous conversational speech, then produced in isolated read form later. It was found that in spontaneous conversation only 20% of the declarative sentences possess declination contour patterns, 45% of them with terminal fall only, and the remaining 35% with unidentifiable contour patterns. When read as isolated single sentences, 50% of these declaratives were produced with declination contours, 27.5% with terminal falls and 22.5% with unidentifiable contour patterns.[25] Results further suggest that syntactic specifications are not always realized in connected speech, either. Why are both tones and intonations so distorted in continuous speech? Rather than treating the above results as tonal and intonation variations, we argue alternatively from a top-down perspective that higher-level discourse information is involved in continuous speech and also contributes to final output prosody. In other words, instead of treating intonation units (IU) as the ultimate prosodic unit and looking for variation patterns of tones and intonations themselves, we argue that tone information (lexical prosody) and intonation patterns (syntactic prosody) combined are insufficient to account for fluent speech prosody. The question then is: what does discourse information signifies, how does it contribute to output prosody and how to analyze and account for it?

Consider first what conditions would call for fluent continuous speech production. Typically, it involves expressions, narrations and/or discussions that require more than one single sentence to convey. The phenomenon is identified as intonation group in the literature of discourse analysis.

Nevertheless, the key feature of intonation group is often not discussed. That is, intonation group is not simply unrelated intonations connected into strings, but a coherent multiple-phrase speech paragraph. It can be either a small discourse by itself or part of a larger discourse. What connects these sentences/phrases has to reflect their coherence; the relative between-phrase semantic association that cannot be expressed by unrelated single sentences must somehow be expressed. Therefore, some additional devices must be available in speech production for speakers to express this semantic association in order to form the coherence that connects between and among sentences. That same device is also used by the listeners to process, derive and recover intended coherence. This is essentially what speech communication is about apart from lexical and syntactic information. Therefore, we argue that fluent speech prosody is basically about between-phrase coherence and association aside from tone and intonation. Higher-level discourse information is the governing constraint of speech prosody above lexical specifications of tones and syntactic specifications of individual phrases. Additional global semantic association is expressed not through each and every phrase intonation, but through cross-phrase global associations. Therefore, issues to be discussed are higher level discourse information, semantic coherence and cross-phrase relative associations.

However, methodological caution must be exercised to analyze prosody. Note that elicited single phrases produced in isolation (one at a time with full stop at each phrase's end) would always yield nothing more than tones and canonical intonations. This is because such phrases contain no discourse information and bear no associative relationship with other phrases. Similarly, single phrases lifted out of continuous speech and studied as independent IU only complicate the matter because they contain fragments of overall discourse prosody that canonical intonations could not accommodate. By analogy, a jigsaw puzzle could never be fully reconstructed unless both relatively large and small scales of reference are used. Likewise, fluent speech prosody is clearly NOT merely strings of independent tones and intonation, but how tones and intonations are systematically structured and modified into coherent speech paragraphs. From this more holistic and top-down perspective, we now need to the following three problems: (1.) identify where additional prosodic information is located in the speech signals, (2.) separate discourse prosody from tones and intonation in prosody analysis, and (3.) account for it through quantitative analysis.

Our previous corpus studies of read discourses have demonstrated that intonation groups in continuous speech are actually structured into three relative discourse positions to yield higher-level information, namely how and where speech paragraphs begin, continue and ends. Through a multiple-phrase prosody hierarchy called Prosody Phrase Grouping (PG),[20,27] whereby PG stands for the prosodic organization that specified phrases it groups through three PG-related positions PG-initial, -medial and –final, corresponding statistical analysis of speech corpora revealed how layered contributions cumulatively accounted for output prosody. These quantitative evidences confirm the existence of cross-phrase prosodic associations in fluent continuous speech, and explain how higher-level discourse information is realized in cross-phrase associations. Evidences of cross-phrase templates for syllable duration patterns, intensity distribution patterns, and boundary breaks as well as systematic account of layered contributions have been reported elsewhere[20, 27]. Hence in the following discussion we will only present analysis of F0 contour patterns to illustrate discourse prosody. Fluent speech prosody, continuous speech prosody and discourse prosody are used interchangeably. The term *prosody*, italicized, will be used as an abbreviation to refer to all three prosody-types.

2. Phrase Grouping: Organization and Framework of Speech Paragraph

The following are prerequisites for an investigation on *prosody*. (1) Only fluent continuous speech should be used for *prosody* analysis so that the associative relationships between and among units within each grouping are available in the speech data. (2) Corpus-based approaches are preferred in order to better accommodate speech variations and facilitate quantitative analyzed. (3) Top-down rather than bottom-up perspective of segmenting speech data is preferred in order for coherent multiple-phrase speech paragraphs to emerge and better reflect the necessary prosodic associations. (4) Speech units above IU should be available in the analysis so that analysis would not focus on individual phrase behavior. (5) Finally, methods of quantitative analysis and predictions should accommodate associative relationship and layered contributions. In other words, speech data type, speech data quantity, segmentation perspective, speech domain type, prosodic units, as well as the quantitative approach would all affect the results of prosody analysis.

The concept of phrase grouping is not just specific to Mandarin. It has been well accepted that utterances are phrased into larger constituents; together they (utterances and larger constituents) are hierarchically organized into various domains at different levels of prosodic organization.[10-12] Unfortunately this hierarchical organization is often ignored, as the necessary distinction between syntactic prosody (intonation) and discourse prosody (*prosody*) often goes un-clarified. In particular, how the phrases PG groups within a hierarchy are associated and what roles IU and intonation are in *prosody* have not received due attention.

Our other corpus studies demonstrated clearly that by adopting a top-down perspective to dissect spoken discourse, it was more than significant to take clearly audible and identifiable multiple-phrase speech paragraphs as prosodic units and work from there, instead of taking one IU at a time. By postulating speech paragraph as a higher-order node of IU, quantitative evidences of layered contributions could be found whereby corresponding cross-phrase acoustic templates could also be derived. [17,20] Our PG hierarchy specifies lower-level units are subject to higher-level constraints while both local (phrase/sentence) and higher (discourse/global) levels of supra-segmental information contribute cumulatively to output *prosody*. *Prosody* is therefore a package of globally associated multiple phrases rather than unrelated strings of IUs. Our simple prosody framework states explicitly that by adding a higher PG level/node above phrases/IU,[17] the respective prosodic roles of phrases PG groups can be defined by simply three PG positions, namely, PG-initial, -medial and -final. These positions implicitly indicate the way a multiple-phrase begins, continues and ends. Compared to other attempts of automatic prosodic segmentation for continuous speech that proposed the classification of phrases into eight phrase types,[8, 9] the PG framework may appear somewhat simplistic on the surface. However, the major difference lies in the sufficiency of only three PG relative positions to capture and explain cross-phrase associations in relation to higher-level discourse information; whereas the eight types remain arbitrary numbers that still assume phrases as independent, unrelated prosodic units without any relationship to each other.

Our PG framework not only specifies phrase as immediate subordinate units, but also by default specifies phrases at the same layer as subjacent sister constituents. By the same logic, PGs can further be extended as immediate constituents of a yet higher node discourse. Figure 1 is a schematic illustration of the framework that also includes the node *Discourse* above PGs.

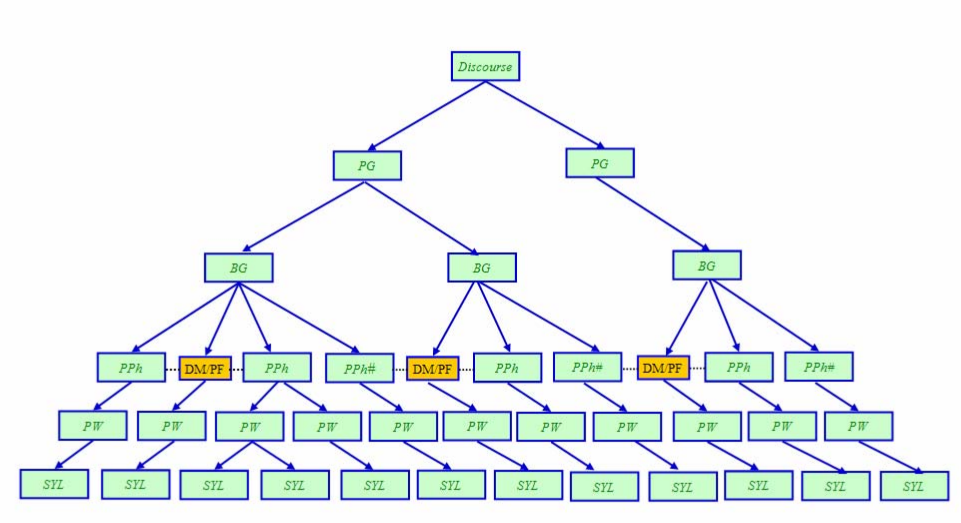


Figure 1: A schematic representation of how PGs form spoken discourse and where DM (Discourse Marker) and PF (Prosodic Filler) are additional associative linkers.

The 6-layer framework is from Tseng 2005[22] and based on the perceived units located within different levels of boundary breaks across the speech flow. The same framework is also used for tone modeling elsewhere in this volume. The units used were perceived prosodic entities. The boundaries (not shown in Figure 1 to keep the illustration less complicated), annotated using a ToBI-based self-designed labeling system,[13] marked small to large boundaries with a set of 5 break indices (BI), B1 to B5, purposely making no reference to either lexical or syntactic properties in order to be able to study possible gaps between these different linguistic levels and units. Phrase-grouping related evidences were found both in adjustments of perceived pitch contours, and boundary breaks within and across phrases, with subsequent analyses of temporal allocations and intensity distribution.[14-16]

Looking at Figure 1 from bottom up, the layered nodes are syllables (SYL), prosodic words (PW), prosodic phrases (PPh) or utterances, breath groups (BG), prosodic phrase groups (PG) and Discourse. Optional discourse markers (DM) and prosodic fillers (PF) between phrases are linkers and transitions within and across PGs, whereby DMs function as attention callers and PFs as parenthetical speech units. These constituents are, respectively, associated with break indices B1 to B5. B1's denote syllable boundaries and may not correspond to silent pauses; B2's, perceived minor breaks between PWs; B3's, breaks between PPhs; B4’s, points when the speaker takes in a full breath upon running out of breath, and also breaks at the BG layer; and B5’s, perceived trailing-to-a-final-ends that occur followed by the longest break. In the framework, an IU is usually a PPh. When a speech paragraph is relatively shorter and does not exceed the speaker’s breathing cycle, the top two layers BG and PG collapse into the PG layer. Both BGs and/or PGs can be immediate subjacent units of a discourse.

The most significant features of the PG framework are how it explains and accounts for variations in intonation across the speech flow and higher-level contributions to *prosody*. The multi-layer framework presented assumes an independent higher level that reflects the scope and unit of online discourse planning and processing. Put simply, the PG framework accounts for why *prosody* denotes *global package* prosody and how that it is formed. Hence it is feasible to assume corresponding canonical and default global templates contribute to the planning within and across units before and during speech production, very similar to cadence templates in music pieces. They also entail that the scope of cross-phrase planning and anticipation is far-reaching and does affect physiologically conditioned articulatory maneuvers at the segmental, tonal, and intonation levels. The additive and trading relationships between tones and sentence intonation were described over half a century ago as “…small ripples riding on large waves”[1] and have been well-known to the Chinese linguistic community. Our framework simply assumes that larger and higher layer(s) exist and may further be superimposed over intonation and tones as tides over both waves and ripples; the reason is to supply more and higher levels of information and discourse association.

In addition, by considering higher-level discourse information with regard to global package prosody, we are able to explain how and why global cross-phrase prosody involves an internal structuring that treats the IUs within as subjacent sister constituents, thus global prosody is therefore systematic and predictable. The framework also implies how the most significant features of discourse information dwell not in individual IUs, but in cross-phrase associations between and among them. Thus either treating *prosody* as strings of unrelated intonations or deliberating on IU behaviors regardless of their relative prosodic context would result in missing the picture of *prosody* completely.

2. 1. Speech Melody: Global F0 Patterns of PG

A cadence template of perceived *prosody* melody is presented in Figure 2, the trajectory denotes a 5-phrase PG, preceded and followed by B4 or B where phrases within are separated by B3's. Note that a PG is featured by how it begins, holds and ends[4]. The unit of the template is a PPh, or an IU.

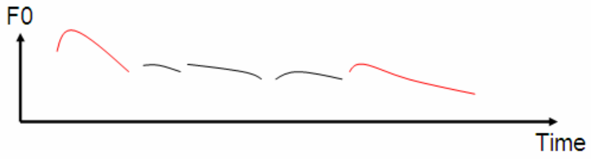


Figure 2. Schematic illustration of the global trajectory of perceived F0 contours of a 5-PPh PG preceded and followed by B4 or B5. Within-PG units are PPh's and separated by boundary breaks B3's.

The following experiments illustrate how to analyze *prosody* from speech data.

2. 1. 1. Speech data

Mandarin Chinese speech data from Sinica COSPRO 08 were used[23]. A 30-syllable, 3-phrase complex sentence representing a short PG was constructed as a carrier paragraph with target single syllables embedded in three PG positions, namely, "△是一個常見的字，一般人常把△字掛在嘴邊，講話時動不動就會提到△. (Translation: △ is a frequently used word, people often use the word △ in their speech, and make mention of △ from time to time quite frequently.)". △ denotes the target syllable. The PG-initial, -medial and -final phrases in the carrier PG consisted of 8, 11 and 11 syllables, respectively. Note that (1) the speech paragraph is designed to remove as much lexical and semantic focus as possible and renders canonical global PG patterns, (2) the target syllables were embedded into the 1st, 6th and last syllable of the first, second and third phrases, thereby occurring at the initial, medial and final locations of the PG-initial, -medial and -final positions respectively; and (3) in spoken discourse, a multiple-phrase PG usually exceeds three phrases indicating the PG-medial phrases are often more than one. Furthermore, when compared to the reading of text passages, although such a 3-phrase complex sentence contains relatively minimal *prosody*, we believe it would still contain discourse information and at the same time offer repetitions of syllables in a uniform context for tone-*prosody* investigations. Speech data from a male

(M054C) and a female (F054C) native speaker of Mandarin Chinese spoken in Taiwan were recorded in sound proof chambers. Both were instructed to read 1,300 speech paragraphs at their normal speaking rate with natural focus into microphones. The speaking rates are 289 and 308 ms/syllable for M054C and F054C, respectively. 60 files from F054C with target syllables of tone 1 were analyzed to illustrate PG effects. Analyses and predictions of F0 values were performed via parameters of the Fujisaki model (*Ap, Aa*) .[2-4, 28-29]

2.1.2. Speech Data Annotation

The speech data were manually labeled by independent transcribers for perceived boundaries and breaks (pauses), using a 5-step break labeling system corresponding to Figure 1. Pair-wise consistency was obtained from the transcribers.

2.1.3. Higher-level Discourse Information in Prosody Analysis

The goal of the following two experiments is to look for phrase components and accent components that also contain additional higher-level information from the PG hierarchy. The Fujisaki model operates on IU to derive F0 curve tendency of both the syllables and the phrase.[2-4, 28-29] Therefore, the three phrases are first analyzed independently then compared in relation to their relative PG positions. Accent components (Aa) and phrase components (Ap) are first separated by a lowpass-filter[5-7] then calculated independently, whereby (Aa) predicts more drastic local F0 variations over time and (Ap) predicts smoother global F0 variations over time. The steps involved are first, analyzing these two components at the PPh level, that is, F0 curve tendency of individual phrases. Next, the same two components are analyzed in relation to higher-level PG information, that is, PPh's are classified by the three PG positions and analyzed respectively. Following that, a comparison of whether differences exist among the three PG positions is made. Lastly, we add contributions from the PPh level and the PG level to derive cumulative predictions and these predictions are then compared with speech data to test the validity.

2.1.3.1 Experiment 1

The aim of this experiment is to investigate (1) whether patterns of Ap could be derived from speech data, (2) whether there is evidence of interaction between Ap predictions from the PPh level and Ap predictions from higher-level PG positions. and (3) whether the evidence found could predict pitch allocation in the speech flow. Two levels of the PG framework are examined. According to the definition of PG hierarchy, all three PPh's at the PPh level are subjacent subordinate constituents of PG which are sister constituents to each other; each PPh is still an independent IU without any higher-level PG information. At the immediate upper PG level, each PPh is then assigned a PG role in relation to the three PG positions. Thus, at the PPh level, each of the three phrases is assumed as an independent prosodic unit. The magnitude of Ap's is generalized and assigned to predict the Ap within, while ignoring higher-level PG information. Next, at the PG level, the PG effects are considered where different values of Ap are assigned to predict phrase components according to where each of the three PPh's is located in PG-positions. Finally, prediction accuracy between PPh's with and without PG effects are compared with the original speech data for validity.

First, speech data are analyzed to provide prediction references. Ap values are extracted from the speech data and their characteristics examined. The respective range and distribution of extracted Ap values in each PG-position from the speech data are illustrated in Figure 2 and Table 1. Next, the characteristics of distribution in each PG-position are generalized and used for subsequent Ap predictions. Using a step-wise regression technique, a linear model is developed and modified for Mandarin Chinese to predict Ap. The hierarchical PG organization of prosody levels (the aforementioned system of boundaries and units) is used to classify Ap at the levels of the framework. Moving from the PPh level upwards to the PG level, we examine how much was contributed by the PG level. All of the data are analyzed using DataDesk™ from Data Description, Inc. Two benchmark values are used to evaluate how close predicted values are when compared with values derived from original speech data. The first benchmark is percentage of sum-squared errors at the lower PPh layer. The PG framework assumes that errors at a lower level are due to lack of information from higher levels. Therefore, residual errors (RE), defined as the percentage of sum-squared residues (the difference between prediction and original value) over sum-squared values of original speech data, are then included into the immediate higher-level for further predictions. If predictions improve from a lower level upward, the difference between two subjacent levels are considered as contributions from the immediate higher level.

Table 1. Range of values of Ap from phrases produced by female speaker F054C in three PG-related positions are presented.

PG Position	Ap range
-Initial	0.959~0.499
-Medial	0.615~0.04
-Final	0.678~0.093

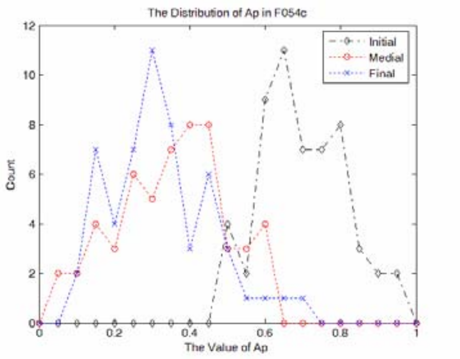


Fig. 3. A schematic representation of the distribution of the Ap's of speaker F054C where the horizontal axis represents values of Ap and the vertical axis represents number of Ap occurrence.

- Results

Table 2 illustrates the coefficients of Aps from PPhs in a PG. At the PPh level, when each PPh is treated as independent prosodic unit, the expected cell mean is at 0.4595. However, at the PG level, where the PPh's were classified by the three PG positions, namely, PG-initial, -medial and -final, the expected cell mean with PG effects are 0.6984, 0.3536 and 0.3265, respectively. In contrast to PG-initial PPh, the Ap of PG-final PPh is shortened. The coefficients reflect a clear distinction between PG-initial and PG-final prosodic phrases.

Table 2. The expected cell mean of predictions with and without the PG effect. The top row shows the expected cell mean value when PG effects are ignored. The bottom row displays the expected cell mean values when PG effect is considered.

Expected Cell Mean at the PPh level without PG effects:	0.4595		
Expected Cell Mean at the PG level with PG effect:	PG Initial	PG Medial	PG Final
	0.6984	0.3536	0.3265

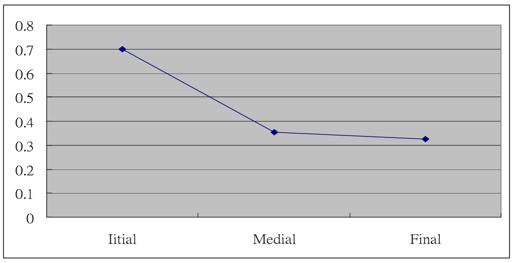
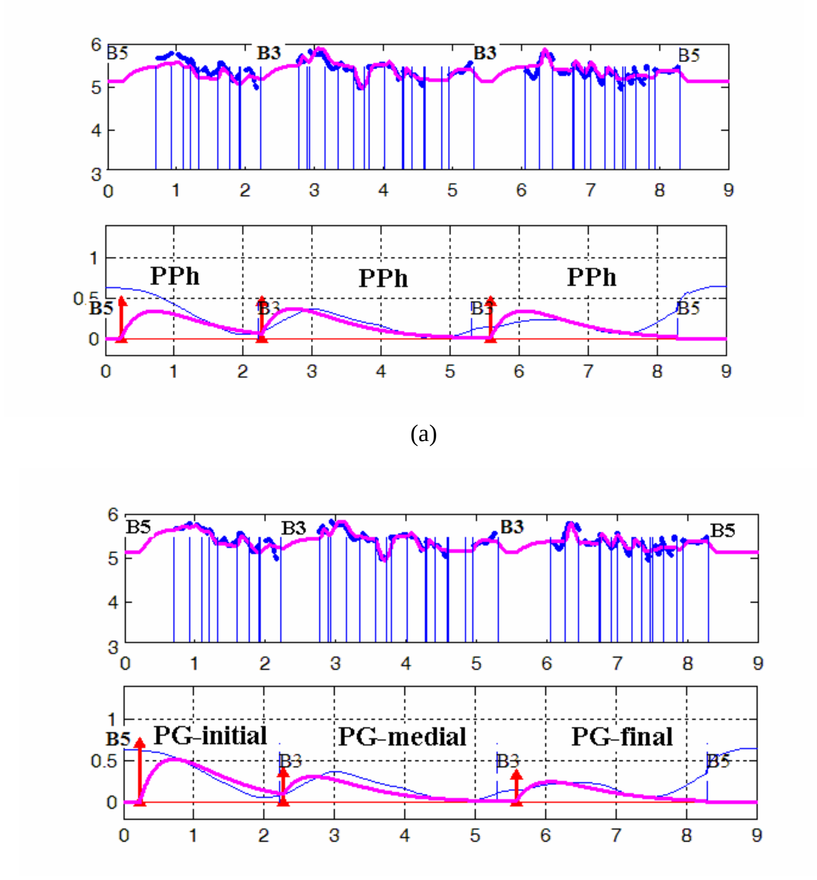


Fig. 4. is a schematic representation of the patterns of phrases after PG effect is taken into consideration. Note how the PG-initial and PG-final groups possess the sharpest distinction.

When each IU (PPh in our framework) is analyzed independently, results revealed that correct predictions were only 40.15% and 59.85% were errors. After considering PG effects one level upward of the prosodic hierarchy, predictions were improved by 24.84%. Cumulative perdition accuracy was 65%. Ap adjustments with respect to PG positions provide further evidence of how prosodic units and layers function as constraints on the Ap in the speech flow and how higher-level prosodic units may be constrained by factors that differ from those constraining lower-level units. If higher-level information is ignored, inputs of prediction would be insufficient.

Finally, by adding up the predictions of the PG layer, we are able to derive a prediction of F0 curve allocation for all three phrases. Comparisons between predictions with and without PG effects are then made with the original speech data. Figures 4 and 5 show these comparisons. The final cumulative predictions indicate that patterns of F0 allocation in Mandarin speech flow cannot be adjusted by the PPh level alone. Input from the PG level must be included. Moreover, these results are also evidence demonstrating that the PPh is constrained and governed by higher-level information (PG). As illustrated in

Figure 5, the distinction between PG-initial and PG-final is most obvious. If PG effect is neglected, the accuracy will diminish.



(a)

(b)

Figure 5. Comparisons of Ap predictions without PG-effect (a) and with PG-effects (b) to the original speech data. The darker line in the upper panels shows F0 plotting of 3 phrases, while the lighter line indicates 3 predicted F0 curves; vertical lines denote syllable boundaries. In the lower panels, the thin line shows comparisons of lowpassed F0 curve, while the thicker line indicates predicted phrase components. Each arrow on the lower panels denote an Ap.; their heights represent Ap values. In each panel, the vertical axis represents logarithm value of F0 curve; while the horizontal axis represents temporal code.

In summary, the PPh layer only constitutes around 40% of the prosody output while higher-level discourse information at the PG layer puts in an additional 25%. Together, the PPh and PG layer make up a total of 65% of prosody output. Note however, that since the PG layer is higher in the prosody hierarchy and commands all phrases under it, its effect is not to be ignored. Without it, there would be no discourse prosody. By definition of the PG hierarchy, the remaining 35% of contributions should come from the lower syllabic (tonal) and word (both lexical and prosodic) levels. Working upwards in the prosody hierarchy, tonal information certainly is not the most significant contributor of fluent speech prosody.

2.1.3.2. Experiment 2

We assume that accent components (Aa, in the Fujisaki model) are also governed by the PG hierarchy as specified by our PG framework. Hence, the SYL, PW and PPh levels in the PG framework should all contribute to output prosody, respectively. The aim of this second experiment is to investigate the contributions of the SYL to PPh prosodic levels from an analysis of Aa. A similar regression technique is used to calculate contributions from each prosodic level to the final output in terms of magnitude of Aa from the SYL, PW and PPh levels.

At the syllable layer, the method adopted is to approach the F0 curve of each syllable by one accent component. In other words, each syllable is connected to one Aa, which makes us unable to extract SYL Aa accurately at the current stage. Nevertheless, the SYL, PW and PPh level models are postulated as follows:

The SYL Layer Model:

$$\mathrm{Aa} = \mathrm{constant} + \mathrm{SYL} + Delta1 \qquad (1)$$

SYL in the above represents syllable type. Factors considered include 23 syllable categories (excluding target syllables), and 5 tones (4 lexical tones and 1 neutral tone).

The PW Layer Model:

$$Delta1 = f(PWLength, PWSequence) + Delta2 \qquad (2)$$

Each syllable is 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 is then calculated using linear regression techniques identical to those of the preceding layer.

The PPh Layer Model:

$$Delta2 = f(PPhLength, PPhSequence) + Delta3 \qquad (3)$$

Each syllable was labeled with a set of vector values; for example, (8, 4) denotes that the unit under consideration is the fourth syllable in an 8-syllable

PPh. The coefficient of each entry is calculated using linear regression techniques identical to those of the preceding layer.

- *Results*

Table 3 shows contributions and cumulative prediction accuracy at each prosodic level from Aa analyses.

Table 3. Cumulative accuracy of Aa predictions from SYL, PW and PPh levels.

Prosodic level	Contribution	Cumulative accuracy
SYL	19.89%	19.89%
PW	1.1%	20.99%
PPh	5.07%	25.16%

If the factors considered include only 5 tones without syllable categories, the accuracy of Aa prediction is about 12.5%. When syllable categories are included, the cumulative accuracy is improved to a cumulative 19.89%. From the SYL layer upwards to the PW level, cumulative prediction is improved to 20.99%. Finally at the PPh level, the cumulative accuracy of Aa prediction is 25.16%.

2.2. Speech Rhythm, Intensity Distribution and Boundary Breaks.

In addition to analysis of speech melody presented in F0 analyses in Section 2.1, we have reported elsewhere similar systematic and layered contributions from the PG hierarchy to global *prosody* in speech rhythm, intensity distribution and boundary break investigations with quantitative evidences[16-21]. In the chapter of tone modeling in this volume, syllable duration patterns are based on patterns derived at both the SYL and PW levels from our PG framework. Detailed discussion of cross-phrase syllable duration templates exhibiting similar pre-boundary lengthening-shortening pattern of the last pentameter at the PPh and PG levels is available[16, 17, 20, 21]. The pentameters are also PG-position conditioned, consistent across Taiwan Mandarin and Beijing Putonghua,[26] where cumulative contributions were accounted for using the same modified linear regression analyses.[18-21] That is, quantitative evidences provided a global cross-phrase rhythmic pattern with three templates for

PG-initial, -medial and -final PPh's, respectively; the patterns interact with syllable durations at PPh/IU, PW and SYL levels and cumulatively add up to output speech rhythm. Relative intensity distribution patterns were found significant only from the PPh level and above, i.e., the longer an IU/PPh is the more energy it requires initially. Significant difference in intensity distribution patterns were found between PG-initial and -final PPh.[18-21] Finally, we found similar PG effects across boundary breaks as well,[18-21] thus proving why pauses across speech flow are PG-conditioned and are therefore constitute systematic and significant prosody information. Without discourse context, pauses are at best concomitant syntactic components of major and minor phrases. With higher-level discourse constraints, at least three degrees of pauses and boundary effects are necessary. Discourse boundary breaks are systematic and predictable as well[16].

3. Discussion

The most important features of the PG framework are how it (1.) captures cross-phrase prosodic associations, and (2.) explains why tones and independent intonation contours are insufficient to account for *prosody*, and (3.) accounts for why discourse information is crucial. Through the PG hierarchy and only three relative PG positions – the PG-initial, medial and –final – the hierarchy specifies subjacent individual PPhs their respective but relative prosodic roles to generate the necessary coherence in multiple-phrase speech paragraphs. We note from the experiments presented above that once a PPh becomes a PG constituent, it is no longer an independent IU but is required to adjust its intonation contour pattern to convey discourse association. The PG-initial and –final positions specify two respective PPhs to retain intonation contours differing in relative starting point, slope, with boundary effects and boundary breaks. However, though both the PG-initial and -final PPh's may exhibit declination, the degree and slope of relative declination differs and final-lengthening-and-weakening only occurs at the PG-final PPh; the PG-medial position specifies all other PPhs between to flatten intonation to signal continuation. The relative positions are dependent on each other and the specifications a package. Hence the phrases in continuous speech must be considered collectively in relation to one another instead of individually one at a time. The PG framework also explains systematically why intonation variations in fluent continuous speech are not random at all but predictable in addition to lower-level syntactic specifications, why speech melody can not be one intonation declination followed by another and why pair-wise contrast between the PG-initial and -final phrases is significant. The global melodic pattern is only present when all PPh's under a grouping are present in the specified linear order; reverting the intonation of PG-initial and –final intonation would not render acceptable *Prosody*. Note also in the experiments presented above, the selected 3-PPh complex sentence represents a relative unmarked representation of a canonical prosodic group but still provides a default PG prototype on which a multiple-phrase paragraph of up to 12 PPh's (as in COSPRO) could be extended. In other words, between a PG-initial and -final PPh, depending the speaking rate, up to 10 PPh's could be accommodated with relatively flatter intonation to signal continuation. This explains the why only 20% to 50% of intonation contours could be identified from read and spontaneous speech reported in Section 1[25]. The PG framework also presents a canonical form for multiple phrase speech paragraphs while stress, focus, and emphasis could all be treated as subsequent add-ons. Without PG specifications, independent individual intonations from continuous speech are "distorted" to almost unlimited variations, even data driven classifications could be arbitrary by nature.[8, 9]

The above results demonstrate that in fluent speech, higher-level information is involved in the planning of speech production; speech units are no longer discrete intonation units. Larger multi-phrase prosodic units reflecting higher-level discourse organization are in operation during the production of fluent speech. Hence methodologically, an IU produced without discourse context does not provide global prosody information. Removing fragments from fluent continuous speech and analyzing microscopic phonetic or acoustic details across segments and/or syllables would not yield systematic accounts towards the structures involved in the semantic coherence as a package, either. To test the validity of the PG framework, we have also constructed a mathematical modular acoustic model that could be used directly in text-to-speech development .[18, 21]

Furthermore, we argue that though global melody and rhythm may differ from one language to another, higher level discourse prosody is not language-specific. Any attempt at prosody organization and modeling should incorporate language-specific patterns of duration allocation and intensity distribution in addition to F0 contours, but maintain the discourse coherence and association.

4. Conclusion

From the evidence presented, we argue that Mandarin speech prosody is not simply about tones and intonation. Any prosody organization of fluent connected speech should go beyond intonation strings and instead accommodate higher-level discourse information above both lexical and syntactic prosody to account for the relative cross-phrase relationship of speech paragraphs. All three acoustic correlates, namely, F0, duration and amplitude, should be accounted for with respect to phrase grouping, along with at least 3 degrees of boundary breaks. Global F0 contour patterns alone are NOT sufficient to represent and characterize the features of *prosody*. Rather, the roles of syllable duration adjustment with respect to temporal allocation over time as well as boundary effects, and intensity distribution with reference to overall cross-phrase relationships should also be included in prosody analysis. Boundary breaks across speech flow are also linguistically-significant components in spoken discourse and deserve a legitimate place in any *prosody* framework. We believe that together with cross-phrase F0 associations, syllable duration patterns, intensity distribution patterns and boundary breaks, a major part of speech melody, rhythm, loudness distribution as well as various degrees of lengthening and pauses collectively reflect the domain, unit and to quite an extent, the strategies of how speech is planned and processed. In short, systematic template and boundary breaks are used by the speaker for planning in the speech production process, and as processing apparatuses by the listener as well. What speakers deliver through prosody by maneuvering available acoustic vehicles are also what listeners utilize to process and predict incoming speech signals. Furthermore, we suggest that both global cross-phrase template fitting and filtering as well as partial local unit recognition should be integrated to facilitate recognition of fluent speech. [20, 27]

To summarize, the most significant features of the PG framework are the following: (1) the framework specifies how three PG relative positions PG-initial and -final subjacent individual PPh assume their respective place under PG and adjust both lexical and syntactic specifications in order to generate global *prosody*. (2) The framework provides a crucial explanation as to why intonation variations in fluent continuous speech are not random, and to what extent PPh's may or may not preserve their phrase intonations. (3) PG effects are evidenced via quantitative analyses at all acoustic correlates, namely, the F0 contour, syllable duration and intensity distribution. (4) Boundary breaks are also PG-governed, systematic and predictable, and are therefore legitimate units of discourse prosody as well. (4) Finally, each layer of the PG hierarchy contributes to output prosody and cumulatively adds up to the final prosody output.[20, 27] Last but not least, the presented analysis and mathematical models[18-20] could also be applied to enhance technological and computational applications, in particular, speech synthesis and unlimited text-to-speech systems.

Future directions include on-going research and preliminary evidence regarding how in addition to semantic coherence speech paragraphs form a discourse through the associations of between-paragraph units.[22,24] These between-units include discourse markers (DM) and prosodic fillers (PF), as shown in the schematic representation of discourse prosody in Figure 1. In other words, an even higher node exists that conveys discourse information between and among paragraphs; the paragraphs in turn become subordinate subjacent discourse units.

Acknowledgement

This work was supported by the Academia Sinica Theme Project “New Directions for Mandarin Speech Synthesis: From Prosodic Organization to More Natural Output” (2003-2005). Experiments and statistical analyses were carried out by Zhao-Yu Su. Formatting was assisted by Yun-Ching Cheng.

References

1. Chao, Y. R., 1968. *A Grammar of Spoken Chinese*. University of California Press, Berkeley and Los Angeles, California.
2. Fujisaki, H., Ohno, S., Tomita, O., 1996. Automatic parameter extraction of fundamental frequency contours of speech based on a generative model, Proceedings of 1996 International Conference on Signal Processing, vol. 1, pp. 729-732.
3. Fujisaki, H., Ohno, S., Wang, C., 1998. A command-response model for F_0 contour generation in multilingual speech synthesis, Proceedings of the 3rd ESCA/COCOSDA International Workshop on Speech Synthesis, pp. 299-304.
4. Fujisaki, H., Ohno, S., Gu, W., 2004. Physiological and physical mechanisms for fundamental frequency control in some tone languages and a command-response model for generation of the F_0 contour, Proceedings of International Symposium on Tonal Aspects of Languages with Emphasis on Tone Language, pp. 61-64.
5. Mixdorff, H., Fujisaki, H., 1997. Automated Quantitative Analysis of F_0 Contours of Utterances from a German ToBI-Labeled Speech Database. In: Proceedings of the '97 Eurospeech, vol.1, pp. 187-190.
6. Mixdorff, H., 2000. A Novel Approach to the Fully Automatic Extraction of Fujisaki Model Parameters. Proceedings of ICASSP 2000, vol. 3, pp. 1281-1284.

7. Mixdorff, H., Hu, Y. and Chen, G., 2003. Towards the Automatic Extraction of Fujisaki Model Parameters for Mandarin. In Proceedings of Eurospeech 2003.
8. Singer., H. and Nakai., M. 1993. Accent Phrase Segmentation Using Transition Probabilities Between Pitch Pattern Templates, *EUROSPEECH'93*, pp. 1767-1770
9. Nakai., M., Singer., H., Sagisaka., Y. and Shimodaira., H. 1995. Automatic prosodic segmentation by *F0* clustering using superpositional modeling, *ICASSP95*, 624–627.
10. Shattuck-Hufnagel, S., Turk, A., 1996. A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguist Research*, 25(2): 193.
11. Gussenhoven, C. 1997. Types of Focus in English? In Daniel Buring, Matthew Gordon and Chungming Lee (eds.) *Topic and Focus: Intonation and Meaning: Theoretical and Crosslinguistic Perspectives*. Dordrecht: Kluwer.
12. Selkirk, E. 2000. The interaction of constraints on prosodic phrasing. In Merle Horne (ed.) *Prosody: Theory and Experiment*, Dordrecht: Kluwer. 231-262.
13. Tseng, C. and Chou, F. 1999. "A prosodic labeling system for Mandarin speech database" *Proceedings of the 14th International Congress of Phonetic Science*, (Aug. 1-7, 1999), San Francisco, California, 2379-2382.
14. Tseng, C. 2002. "The prosodic status of breaks in running speech: Examination and Evaluation" *Proceedings of the 1st International Conference on Speech Prosody 2002*, (Apr. 11-13, 2002), Aix-en-Provence, France, 667-670.
15. Tseng, C. 2003. "Towards the organization of Mandarin speech prosody: Units, boundaries and their characteristics" *Proceedings of the 15th International Congress of Phonetic Science* (ICPhS-2003), (Aug. 3-9, 2003), Barcelona, Spain, 599-602.
16. Tseng, C. and Lee, Y. 2004. "Speech rate and prosody units: Evidence of interaction from Mandarin Chinese" *Proceedings of the International Conference on Speech Prosody 2004*, (Mar. 23-26, 2004), Nara, Japan, 251-254.
17. Tseng, C., Pin, S., Lee, Y., 2004. 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*, 417-438.
18. Pin, S., Lee, Y., Chen, Y., Wang, H. and Tseng, C. 2004. "Mandarin TTS system with an integrated prosody model," *Proceedings of the 4th International Symposium on Chinese Spoken Language Processing*, (Dec. 15-18, 2004), Hong Kong , 169-172
19. Tseng, C. and Lee, Y. (2004). "Intensity in relation to prosody organization ," *Proceedings of the 4th International Symposium on Chinese Spoken Language Processing*, (Dec. 2004), Hong Kong , 217-220
20. Tseng, C. Pin, S. and Lee, Y., Wang, H. and Chen, Y. 2005. "Fluent Speech Prosody: Framework and Modeling", *Speech Communication (Special Issue on Quantitative Prosody Modeling for Natural Speech Description and Generation)*, Vol. 46:3-4, 284-309.
21. Tseng, C. and Fu. B. 2005. "Duration, Intensity and Pause Predictions in Relation to Prosody Organization," *Proceedings of Interspeech 2005* ,(September 4-8 ,2005) ,Lisbon ,Portugal,1405-1408
22. Tseng,C., Chang, C. and Su, Zh. 2005. "Investigation F0 Reset and Range in relation to Fluent Speech Prosody Hierarchy", *Technical Acoustics,* Vol. 24, 279-284.

23. Tseng, C., Cheng, Y. and Chang, C. 2005. “Sinica COSPRO and Toolkit—Corpora and Platform of Mandarin Chinese Fluent Speech” *Proceedings of Oriental COCOSDA 2005*,(Dec. 6-8, 2005), Jakarata, Indonesia, 23-28
24. Tseng, C., Su, Zh., Chang, C. and Tai, C. 2006. Prosodic filers and discourse markers—Discourse prosody and text prediction. *TAL 2006 (The Second International Symposium on Tonal Aspects of Languages)* April 27-29, 2006, La Rochelle, France.
25. Tseng, C. 2006. An Acoustic Phonetic Study on Tones in Mandarin Chinese. Institute of Linguistics, Academia Sinica, Taipei, Taiwan. (2nd ed. CD-rom)
26. 鄭秋豫、李岳凌、蔡蓮紅、鄭雲卿 （排印中）“兩岸口語語流韻律初探—以音強及音節時程分佈為例” *首屆海峽兩岸現代漢語問題學術研討會論文集*. 上海商務印書館
27. Tseng, C. 2006. ”Higher Level Organization and Discourse Prosody“, Invited keynote paper, *TAL 2006 (The Second International Symposium on Tonal Aspects of Languages),* April 27-29, 2006, La Rochelle, France. 23-34
28. Wang. C., Fujisaki, H., Ohno, S., T, Kodama., 1999. Analysis and synthesis of the four tones in connected speech of the Standard Chinese based on a command-response model, Proceedings of the 6th European Conference on Speech Communication and Technology, vol. 4, pp. 1655-1658.
29. Wang. C., Fujisaki, H., Tomana, R., Ohno, S., 2000. 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 the 6th International Conference on Spoken Language Processing, vol. 3, pp. 326-329.