From One Base Form to Multiple Output Styles-- Predicting Stylistic Dynamics of Discourse Prosody

Chiu-yu Tseng and Zhao-yu Su

Phonetics Lab, Institute of Linguistics, Academia Sinica, Taipei, Taiwan
cytling@sinica.edu.tw

Abstract

We hypothesize that various prosody output styles can be predicted and simulated from one default base form by accounting for contributions from higher level information to cross-phrase prosodic relationship. Speech materials of four prosody styles were selected: (1.) Han and Tang poetry, (2.) Tang Ballads and Song poetry, (3.) Qin, Tang and Song classic prose and (4.) contemporary TV weather forecast. F0 contours were analyzed using the Fujisaki model, while quantitative analyses of predictions from layered-and-cumulative contribution specified by the HPG (Hierarchical Prosodic phrase Grouping) framework [Tseng et al, 2004; 2005; 2006] were performed across styles and speakers. Results confirmed that higher level contribution is significant across style; contribution distribution patterns and style specific; more regular prosodic formats require more contribution from higher level; stylistic dynamics are predictable; and the HPG base form is indeed default.

Index Terms: Hierarchical Prosody Group, HPG, discourse prosody, linear regression model, higher level contribution, prosody stylistic dynamics

1. Introduction

We have established previously [1, 2 ,3] from quantitative corpus analyses of Mandarin Chinese that fluent speech prosody contains higher level discourse information above intonation unit (IU). We further stated that higher information is the semantics that associates phrases and sentences into coherent speech paragraphs beyond syntax government, delivered through cross-phrase prosodic context, most notably as intonation variations. Our Hierarchical Prosodic phrase Grouping (HPG, formerly termed PG) framework, specifies how higher level discourse information constrains and triggers individual phrase intonations to adapt systematically in order to yield multi-phrase paragraph prosody; how layered and contributions cumulatively make up output prosody; how contributions can be accounted for quantitatively; and why output intonation variations are systematic and predictable. (See details in [2, 3].) We specify discourse-defined roles by phrase units in three relative HPG-positions: HPG-initial, -medial and –final. Cross-phrase dynamic but systematic templates for F0 contours, syllable duration adjustment, intensity distribution and boundary break patterns were quantitatively derived. Correlating modular acoustic simulation models were also constructed [3]. Figure 1 shows the 6-layer tree diagram of the HPG framework in prosodic units that accounts for multi-phrase output prosody. From bottom up, the layered nodes are syllables (SYL), prosodic words (PW), prosodic phrase (PPh), breath groups (BG), prosodic phrase groups (PG) and Discourse. The upper prosodic layers/levels above PPh can also collapse to accommodate discourse of various lengths.

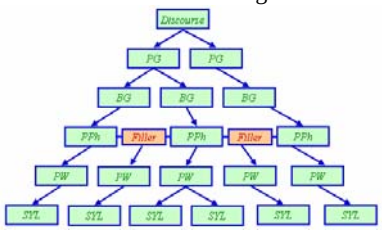

Figure1. A schematic tree diagram of phrase-grouping discourse organization in prosodic levels and units, including between-phrase fillers and markers.

Our current hypothesis is as follows (1.) the multi-phrase base from of the HPG framework can function as a default base form from which various output prosody styles could be predicted. (2.) Prosody styles of regular formats and templates (such as poetry and ballads) can be attributed to more contribution from higher level information above PPh. (3.) Dynamic distribution of layered contribution from each prosodic level is systematically patterned by prosody style and across speakers.

2. Speech Materials and Data

Speech data consisted of a total of 4 prosody styles. Read speech of three different literary styles from fixed rhyme templates (Han and Tang poetry), semi-fixed rhyme templates (Tang Ballads from Music Bureau and Song lyrics to free style (Qin, Tang and Song classic prose), as well as a fourth type of reading text pieces from TV weather broadcast are collected. While the poetry and ballads contain built-in regular to semi-regular prosodic formats including rhyming, free style classic prose is the finished product of refined rhetoric editing without rhyming templates. These three styles all bear prosodic characteristics that can be attributed to distinct stylistic variations. The fourth choice of text from TV weather broadcast contains frame sentences without rhyme patterns. 26 pieces (approximately 3,400 syllables) were selected to cover the first thee styles whereas 34 pieces (approximately 7,703 syllables) of TV were selected. Table 1 summarizes the speech materials. A total of 3 speakers produced microphone speech in sound proof chambers. One female speaker read all the materials; one male read the first three styles; one other male read the fourth style. Table 2 summarizes annotated speech data of prosody formats 1 to 3 by speaker in number of PPh, discourse and speaking rates. Table 3 summarizes annotated TV weather forecast in the same categories as in Table 2. Note

that there are approximately the same number of PPh from prosody styles 1 to 3 (710 and 711) and prosody style 4 (720 and 747).

Table1. Summary of 4 types of materials by prosody format and style. Styles number indicates: 1. Han & Tang poetr 古詩; 2. Tang Ballads 規則樂府; 3. Tang Ballads 不規則樂府; 4. descriptive prose interspersed with verse 賦; 5. Song lyrics 宋詞); 6. ballad 民歌; 7. Qin,Tang and Song classic prose 古文; 8. weather broadcast 氣象播報

prosody format	1. regular		2. semi-regular				3. irregular	total	4. TV weather
style	1	2	3	4	5	6	7		8
range/syllable	40-300	150-262	75-176	341	47-104	330	107-202	40-330	80-434
piece	6	4	2	1	8	1	4	26	34
total # of syl	610	692	151	341	631	330	646	3407	7703
# of piece	10		12				4	26	34

Table2. Summary of speech materials from styles 1 to 3 in number of syllable, PPh, Discourse, speaking rate by speaker.

Poems, ballads and classics	# of Syl	# of PPh	# of Discourse	speech rate (ms)/Syl
female f054	3502	710	26	271
male m056	3510	711	26	202

Table3. Summary of speech data of TV weather forecast in number of syllables, PPh, Discourse and speaking rate by speaker

Weather forecast	# of Syl	# of PPh	# of Discourse	speech rate (ms)/Syl
female f054	7054	720	34	193
male m054	7096	747	34	165

The speech data were manually annotated by trained transcribers for perceived PPh, BG and PG by the HPG framework (see Figure1 and [1, 2, 3]). Note also that the speaking rates of prosody styles 1 to 3 (271 and 202 ms/Syl) is slower than those of prosody style 4 (193 and 165 ms/Syl), indicating even non-professionals reading text of weather forecast adopted a faster speaking rate.

3. Method of Analysis

Central to the present study is analyzing F0 contour patterns from the PPh and BG, PG levels in the HPG discourse hierarchy, and accounting for layered and cumulative contributions quantitatively. The Fujisaki model [4] was adopted for F0 analysis while a linear regression model [3, 5, 6] was adopted to account for contribution distributions. By definition, the Fujisaki model specifies two basic layered parameters: a phrase component Ap at the phrase level for global contour; and an accent component Aa for local focus. The Aa command was later adopted at the syllable level to simulate syllabic tone patterns [7]. This adaptation readily renders the obligatory interactive relationship between syllable tone and phrase intonation hierarchical because the syllable tone is from smaller and lower level by lexical definition, while phrase intonation from larger and higher level by syntactic definition. The two parameters were extracted automatically [7, 8] before applying the adopted linear regression model [1] to predict contribution from each level specified by the HPG framework.

3.1. Extracting the Fujisaki parameters

Filter-based automatic extraction of the Fujisaki parameters Aa and Ap [7, 8] is to separate high frequency contour (HFC) and low frequency contour (LFC) from the speech signals. However, note though the original Fujisaki model is phrase based, reported studies on Mandarin have not defined units for extraction and simulation. Mandarin units varying a great deal by syllable numbers and/or syntactic structure were taken one at a time for simulation at the phrase level only without any discourse context. Our approach differs from reported studies most characteristically in that we assume higher level contribution from the discourse hierarchy is what constrains and triggers to phrase intonations to modify systematically, thus attempted to include prosodic information above phrase. As a result, when extracting the Fujisaki parameters automatically, we use only annotated PPh's by HPG specifications instead of random units, thereby in subsequent tests make it possible to justify the hierarchical hypothesis quantitatively.

Following rationale mentioned above, the extracted LFC represents goal of phrase components where PPh boundary is used as decision position to insert phrase commands. Figure 2 show an example a short 3-phrase paragraph and patterns of extracted Aa and Ap. The top panel displays original F0 contour patterns in red and simulated F0 contours in blue; the middle panel extracted accent command Aa corresponding to tone components; the lower panel extracted phrase command Ap corresponding to phrase components. Magnitudes of extracted phrase components are then subject to subsequent quantitative analysis via linear regression.

Figure2. An example of a 3-phrase paragraph and auto-extracted Fujisaki parameters. The top panel displays original F0 contour patterns in red and simulated F0 contours in blue. The middle panel displays extracted accent command Aa corresponding to tone components. The lower panel displays extracted phrase command Ap corresponding to phrase components.

3.2. Accounting for higher Level contributions and distributions above PPh

Using a step-wise linear regression technique adopted for the HPG framework [3], a linear model with 3 layers is developed to predict speakers' F0 behavior over time with the Fujisaki parameters from the PPh layer upward. Prediction begins at the PPh layer where we predict the F0 patterns by each PPh independently while residual between prediction and original values is regarded as contribution from the immediate higher

level instead of error. In other words, within each layer, we assume no linear association between phrases. The same prediction is then repeated at the BG layer by including residuals from the immediate lower PPh layer, and regards residuals at the BG layer as contribution from the next higher layer PG. The same prediction is repeated one last time at the PG layer. Ultimate prediction is derived by adding up layered contributions, and thus accounts for layered-and-cumulative contributions. Figure 3 is a schematic representation of the regression processes. Note that in the present study on F0 contour patterns we only considered contributions from the PPh level and above.

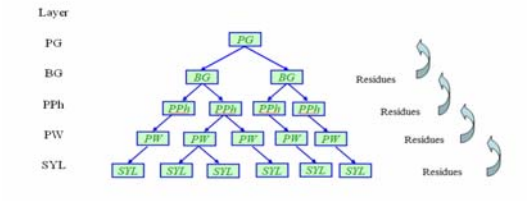

Figure3. A schematic representation of linear regression predictions from the Syl level upward; whereby each level contributes to final output independently and cumulatively.

At the PPh layer, the magnitude of phrase components (Ap) are categorized by the length (in syllable numbers) of the preceding PPh and current PPh. Then the residuals of Ap regarded as effects from the immediate higher level BG are predicted by its respective within-BG positions at the next level up. For example, "BGSequence =1" indicates the Ap is located in the first PPh of a specified BG. The prediction process is as follows. Delta represents the residual of Ap in each prosodic layer.

PPh Layer :

$$\text{Ap} = f(precedingPPhLength, currentPPhLength) + Delta1$$

BG Layer :

$$Delta1 = f(BGSequence) + Delta2$$

PG Layer :

$$Delta2 = f(PGSequence) + Delta3$$

4. Results

4.1. Contribution from higher level information above PPh and distribution patterns across style

Figure 4 shows results of respective dynamic contribution patterns and distributions within and across 4 prosody styles regular (R), semi-regular (SMR), irregular (IR) and weather broadcast (WIR); and by 3 HPG prosodic layers, PPh, BG and PG. Table 4 summarizes cross-speaker comparison of prediction percentage by the same parameters. Contributions from three prosody layers PPh, BG and PG are accounted for. Note that (1.) result confirm that higher level information contributes to output prosody across prosody styles and speakers whereby most significant contributions come from the BG layer. Note that BG layer contributions by speaker account for 28.1% and 51.57%, respectively, of output prosody R; 20.39% and 33.67% of output SMR; 8.7% and 16.52% of output IR; and 5.063% and 7.36% of output WIR. Contributions from the PPh layer are in complimentary distribution of the BG layer while contributions from the PG layer insignificant. (2.) Patterns of contribution distribution are prosody-style dependent by HPG-layer. Note how each prosody style possesses distinct distribution patterns in the PPh and BG layers. (3.) More regular prosodic style shows more contribution from the BG layer while BG contribution reduces by prosody style from R, SMR, IR to WIR, as shown in Figure 4 and summarized in Table 4. The gradation from R to WIR is also systematic by prosody style. In short, distribution patterns are both style-dependent and style-specific. (4.) Contributions from both the PPh and BG layers are obligatory across prosody style, together they comprise output prosody. (5.) The results also confirmed the hypothesis of the HPG base as default. Dynamic variations of prosody by style can now be accounted for on an R-to-IR or even –WIR continuum by varying proportional contributions from the PPh to BG layers, rather than attributing each prosody style independently.

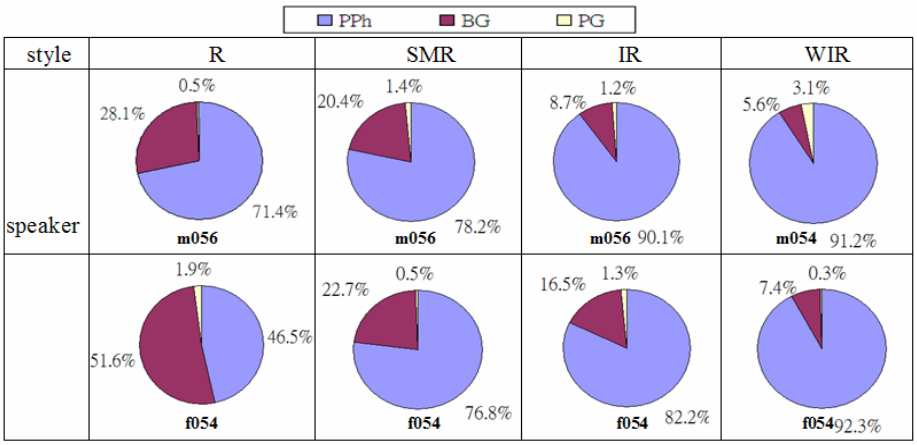

Figure4. Cross-speaker (m056; m054 and f054) comparison of respective contribution distributions within and across 4 prosody styles regular (R), semi-regular (SMR), irregular (IR) and weather broadcast (WIR); and by 3 HPG prosodic layers, PPh, BG and PG.

Table4. Cross-speaker (m056; m054 and f054) comparison of prediction percentage by 3 prosodic layers PPh, BG and PG; by 4 styles within and across 4 prosody styles regular (R), semi-regular (SMR), irregular (IR) and weather broadcast (WIR); and by 3 HPG prosodic layers, PPh, BG and PG.

speaker	style	PPh	BG	PG
m056	R	71.44%	28.10%	0.46%
	SMR	78.22%	20.39%	1.39%
	IR	90.08%	8.70%	1.22%
m054	WIR	91.22%	5.63%	3.14%
f054	R	46.50%	51.57%	1.93%
	SMR	76.82%	22.67%	0.50%
	IR	82.16%	16.52%	1.32%
	WIR	92.30%	7.36%	0.34%

4.2. Comparison of Speaker Behavior by Style

We noted a relatively sharp between-speaker difference of BG-layer contribution for prosody style R, and further compared speaker variations by prosodic layer and across style in order to see if the general cross-style patterns remain similar in spite of contribution variation. Figure 5 shows the accuracy of Ap at the PPh layer by speaker and style. Note how the trajectory of Ap accuracy for both speakers rises as the prosody format becomes more irregular from R to SMR to IR. Figure 6 shows the same comparison at the BG layer where the accuracy patterns of Ap are opposite from the PPh layer but similar across speakers.

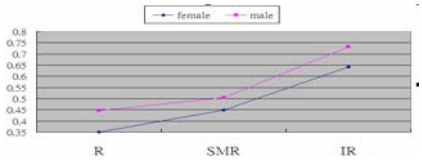

Figure5. Cross-style comparisons of Ap accuracy by speaker at the PPh layer. The horizontal axis denotes prosodic styles from R, SMR to IR; the vertical axis accuracy of Ap.

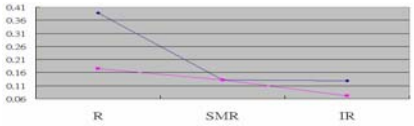

Figure6. Cross-style comparisons of Ap accuracy by speaker at the BG layer. The horizontal axis denotes prosodic styles from R, SMR to IR; the vertical axis accuracy of Ap.

Cumulative Ap accuracies from the PPh and BG layers are shown in Figure 7. Note how the additive outcome from two inverse patterns compensates each other and results the final trajectories to become flat.

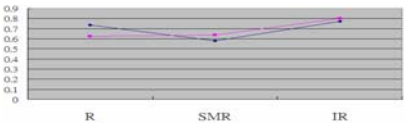

Figure7. Cumulative cross-style comparisons of Ap accuracy by speaker. The horizontal axis denotes prosodic styles from R, SMR to IR; the vertical axis accuracy of Ap.

5. Discussion and Conclusion

From the above results of F0 analyses, we confirmed our hypothesis that by accounting for contributions from the immediate prosodic layer BG above prosodic phrases PPh, various output styles can be predicted and simulated from one base form. Higher level information contributes to output prosody across prosody styles and speakers while the significance of contributions from the BG layer can NOT be ignored. The results also serve as further evidence that individual PPh (as intonation unit IU) in fluent speech are not independent unrelated prosody units, but rather subordinate subjacent sister units associated by semantic cohesion governed by information above phrases; and delivered through prosodic context. Cross-style comparisons also revealed systematic style-specific layer-dependent patterns of contribution distribution from the PPh and BG layers, respectively. In short, the more regular the prosodic format is; the more contribution comes from upper layer BG; and vice versa. Output prosody are cumulative outcome from layers involved; while the HPG framework quantitatively accounts for the contribution patterns by prosodic layer and prosody style. Frame sentences and/or paragraphs used as in TV weather forecast also function as a default prosody base form. In addition, the female speech showed larger contribution from higher level BG than from PPh for style R, thus further supports higher level information in prosody formation and speaker style. Thus in conclusion, we establish that one default base form by the HPG framework can systematically account for different levels of contribution; while dynamic output prosody styles can be generate by altering contribution distributions only. We believe the results are also significant to discourse comprehension, spoken language processing as well as technological implementations.

6. References

[1] Tseng, Chiu-yu. "Prosody Analysis", *Advances in Chinese Spoken Language Processing*, World Scientific Publishing, Singapore: 57-76, 2006.

[2] Tseng, C. Pin, S. and Lee, Y., Wang, H. and Chen, Y. "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, 2005.

[3] Tseng, Chiu-yu and Lee Yeh-lin (2004). “Speech rate and prosody units: Evidence of interaction from Mandarin Chinese”, *Proceedings of the International Conference on Speech Prosody 2004*, 251-254.

[4] Fujisaki H, Hirose K. "Analysis of voice fundamental frequency contours for declarative sentences of Japanese", *J.Acoust, J.Acoust. Soc. Jpn.(E)*, 1984; 5(4): 233-242,1984.

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

[6] Zellner Keller B, Keller E., “Representing Speech Rhythm” *Improvements in Speech Synthesis*. Chichester: John Wiley, 154-164,2001

[7] Mixdorff, H., "A Novel Approach to the Fully Automatic Extraction of Fujisaki Model Parameters", *Proceedings of ICASSP 2000*, vol. 3: 1281-1284, 2000.

[8] Mixdorff, H., Hu, Y. and Chen, G. "Towards the Automatic Extraction of Fujisaki Model Parameters for Mandarin", *Proceedings of Eurospeech* 2003; 873-876.