What's in the F0 of Mandarin Speech --Tones, Intonation and beyond

Chiu-yu Tseng
Institute of Linguistics, Academia Sinica
Taipei
cytling@gate.sinica.edu.tw

Zhao-yu Su
Institute of Linguistics, Academia Sinica
Taipei
morison@gate.sinica.edu.tw

***Abstract*—We analyzed F0 contours of fluent Mandarin speech using a modified command-response model. Adopting the multiple-phrase speech paragraph as a discourse prosodic unit, we investigated the composition of F0 contours to see whether additional prosodic information beyond tones and intonation exists. Testing F0 contributions with a previously constructed prosody hierarchy the HPG (Hierarchy of Prosodic Phrase Grouping), results showed that tone identities only make up 40-45% of output F0 while other higher layers of information contributes to the rest. Final F0 output is cumulative of all layers combined. The results thus provide an account of why prosodic context consists of both adjacent and cross-over associations and how global prosodic context is reflected in the formation of output F0. We believe these results shed new lights on speech technology development.**

Keywords HPG, tones, intonation, higher-level contributions, prosody context, F0 contour, cross-over, adjacency.

I. Introduction

Mandarin tones by syllable have been considered the most significant prosodic feature of Mandarin speech. Most Mandarin ASR efforts aim mainly at tone identification. However, we have studied the prosody of fluent Mandarin speech extensively and found even tones and intonation combined is insufficient to characterize fluent Mandarin prosody, and prosody units require re-definition. We have constructed a prosody framework the HPG (Hierarchy of Prosodic Phrase Group) by adopting the multiple-phrase speech paragraph as a prosodic unit, and demonstrated how each prosodic unit from prosodic word (PW), prosodic phrase (PPh), and phrase groups (PG) contributes to output prosody. [1] [2] [3] Most importantly, the HPG framework specifies layer-dependent prosodic contributions, thus accommodating more source of overall prosody information from different size of unit. In particular, how at the phrase level, a three-way patterned specification indicating the initiation, continuation and termination of a paragraph forms the obligatory prosodic context in output prosody that contains both adjacent and cross-over associations of within-paragraph phrases, as well as paragraph association that forms the larger discourse. It was evident that prosodic context goes beyond tone and intonation concatenation and smoothing. Breaking down fluent speech into units of tones and intonation only makes it impossible for higher level prosodic context to surface.

In the following study, we will present corpus analysis the F0 contours by the HPG prosodic units the Syllable (SYL), PW, PPh and PG to illustrate F0 patterns could be derived at each and every unit and why no single unit accounts for output prosody. In particular, how quantitatively these units contribute to output F0. Finally, we will discuss the implication of better understanding of F0 composition to speech technology development.

II. Speech material

Two types of text were used: (1.) Plain text of 26 random discourse pieces (CNA, approximately 6700 syllables), and (2.) three rhyme formats of Chinese Classics (CL approximately 1600 syllables). One male and one female read each text type at sound proof chambers into microphones. A total of 4 sets of speech corpora were obtained, namely, M051 and F051 for CNA and M056 and F054 for CL. Pre-analysis included automatic annotation of segmental labeling by the HTK toolkit were first obtained using the SAMPA-T notations [4], then spot-checked for segmental alignments by trained transcribers. Manual tagging of perceived prosodic units and boundary breaks was performed using the Sinica COSPRO Toolkit [5]. The mean syllable duration is 199ms and 189 ms for F051 and M052 (data type CNA); and 265ms and 202ms for F054 and M056 (data type CL), indicating positive cross-speaker correlation of duration style and format may be inherently related to speaking rate regulation.

III. F0 Analysis

A. Analyze F0 contour using the command-response model

The physiology based command-response model, commonly known as the Fujisaki model, was used to analyze F0 contours. [6] The model can be decomposed into three components and major corresponding parameters including base frequency, a Phrase command Ap indicating the magnitude of global contour of a phrase and Accent command Aa indicating local humps of smaller domain. Aa is superimposed onto Ap to derive the ultimate output F0 contour. The model inherently assumes that F0 is generated from more than one component differing in size and scale, as defined as below.

$$\ln F_0(t) = \ln F_b + \sum_{i=1}^{I} Ap_i G_p(t - T_{0i}) + \sum_{j=1}^{J} Aa_j [G_a(t - T_{1j}) - G_a(t - T_{2j})]$$

i, j =Index of phrase command, Index of accent command
F_b =Base frequency
Ap_i =Phrase command magnitude
Aa_j =Accent command magnitude
$G_p(t - T_{0i})$ = Phrase command response function
$[G_a(t - T_{1j}) - G_a(t - T_{2j})]$ = Accent command response function

When the model is used to analyze tone languages, it has been commonplace that the Ap command is applied to phrase units to represent the intonation contour pattern, and the Aa command to the syllable to represent tones. We modified methods by Mixdorff [7] [8] to auto extract the Ap commands, fitting one Aa to the syllable only while the scale selected for Ap is one PPh each time [9]. In the

978-1-4244-2942-4/08/$25.00 (c) 2008 IEEE

following sections, we calculated these two parameters to each of the HPG specified layer from the syllables and above.

B. Calculating layered contributions by the HPG

Figure 1 shows a schematic representation of the HPG framework and corresponding regression processes used for contribution calculation. The hierarchy specifies how sister units at the same layer bears adjacent relationships, how higher layers above bears governing constrains to lower layer(s) to provides cross-over associations, and how each of the units and layers contribute systematically to global output paragraph prosody. The layered HPG prosodic units from bottom upward the hierarchy are the syllable (SYL), the prosodic word (PW), the prosodic phrase (PPh), the breath group (BG) and the multiple phrase group (PG). The hierarchical relationships among these nodes are SYL<PW<PPh<BG<PG. Not shown in the scheme due to space limitation are respective discourse boundaries B1, B2, B3 B4 and B5 that are HPG prosodic units as well and bears correlative HPG associations. [1] [2] [3]

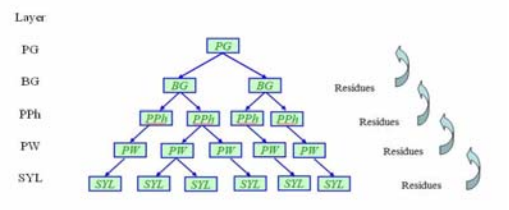

Figure 1 A schematic representation of linear regression predictions from the Syl level u*pward;* whereby each level contributes to final output independently and cumulatively.

Using a step-wise linear regression technique [10][11] adopted for the HPG framework, a linear model is developed to predict F0 contour with the Fujisaki parameters for each layer. Because the Fujisaki model defines Ap and Aa to separate contour patterns from global and local, the procedure is executed in two separate parts, namely, a higher part and a lower part. The higher part is larger and the predictions relatively global; the lower part smaller and the predictions relatively local. Residuals between predictions and original values are regard as contribution from the immediate higher layer instead of error. For example, when deriving PW predictions no linear association between SYL is assumed for PW prediction. The same predictions are repeated at each HPG layer from the SYL upward to PG, defined as below.

Syl

Aa=f(FollowingTone, PrecedingTone, CurrentTone) +Delta1

PW

Delta1=f(PW Boundary Infor, PWSequence)+Delta2

Boundary effect above PPh

Delta2=f(PPh Boundary Infor, PG Boundary Infor) +Delta3

PPh

Ap=f(FollowingPPh_Length, PrecedingPPh_Length, CurrentPPh_Length)+Delta4

BG

Delta1=f(BGSequence)+Delta5

PG

Delta2=f(PGSequence)+Delta6

At the lowest SYL layer, an Aa is predicted by the current tone and context of its preceding and following tone while the residuals are predicted by the next higher layers PW. An Ap is predicted from the PPh layer using the same rationale, i.e., by context of respective lengths of the preceding and following PPh, while the residuals are predicted by the next higher layers BG and PG, respectively. Finer tuning includes calculating boundary context and properties by PW, PPh and above. [12] [13] Ultimate prediction is then derived by adding up layered contributions, thus accounting for layered-and-cumulative contributions.

IV. Results

A. Ap & Aa predictions by the HPG

Table 1 shows Aa predictions from SYL, PW and boundary effects above PPh, i.e., contributions from the lower HPG layers. The final cumulative accuracy of Aa prediction ranges from 56.25% to 73.80%

Table1. Cumulative accuracy of Aa prediction from SYL, PW and Boundary effect above PPh

Corpus	Speaker	Boundary effect above PPh		Contribution of boundary effect
		PPh Info	PG Info	
CL	F054	72.98%	73.80%	7.19%
	M056	64.13%	66.89%	5.43%
CNA	F051	54.41%	56.25%	4.98%
	M051	57.43%	59.32%	4.79%

Table 2 shows the Ap prediction from PPh, BG and PG, namely the higher layer contributions of HPG. The final accuracy of Ap prediction ranges from 73.66%% to 88.20%

Corpus	Speaker	SYL Contribution		PW Contribution	
		Tone	Tone Context	PW Boundary Info	PW Position Sequence
CL	F054	46.21%	54.74%	60.54%	66.61%
	M056	39.12%	47.86%	57.68%	61.45%
CAN	F051	38.40%	45.00%	48.43%	51.27%
	M051	41.61%	47.96%	51.33%	54.53%

Table2. Cumulative accuracy of Ap prediction for PPh, BG and PG

Corpus	Speaker	PPh	BG	PG
CL	f054	58.79%	63.58%	76.66%
	m056	37.89%	48.99%	73.66%
CNA	F051	80.17%	81.46%	87.71%
	m051	81.53%	82.72%	88.20%

Average of Aa and Ap predictions were used as the final accuracy of total F0 contour prediction. The results are presented in Table 3

Table3. Cumulative accuracy of Ap prediction for PPh, BG and PG

Corpus	Speaker	Aa	Ap	Total
CL	f054	76.66%	73.80%	75.23%
	m056	73.66%	66.89%	70.28%
CNA	F051	87.71%	56.25%	71.98%
	m051	88.20%	59.32%	73.76%

B. Tone model of Aa

Figure 2 shows that at the SYL level, tone identities are distinct. The Aa patterns of each tone are similar across the 4 speakers, as shown in previous studies using the Fujisaki model. [9]. However, correct prediction of Aa by tone identities is about 40~45%.

978-1-4244-2942-4/08/$25.00 (c) 2008 IEEE

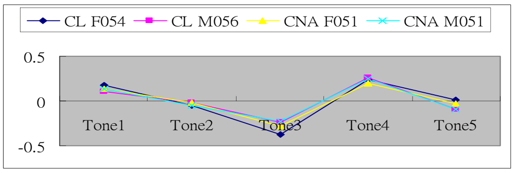

Figure2. Tone model of Aa. The horizontal and vertical-axis indicate the tone index and average Aa value, respectively.

C. PW model of Aa

We classified the PW by three PG positions –Initial, -Medial and -Final to observe the PW model of Aa irrespective of tone effects. Figure 3 shows PW-final syllables exhibited pre-boundary F0 lowering while no obvious reset was found in other positions. Similar pre-boundary declination is found in all three PG-initial positions across speakers except for speaker f051 though the degree of declination differs by PG positions. However, the second syllable into PW at –initial and –final positions may vary, perhaps due to stress and/or other prosody effects that merit further investigation in the figure.

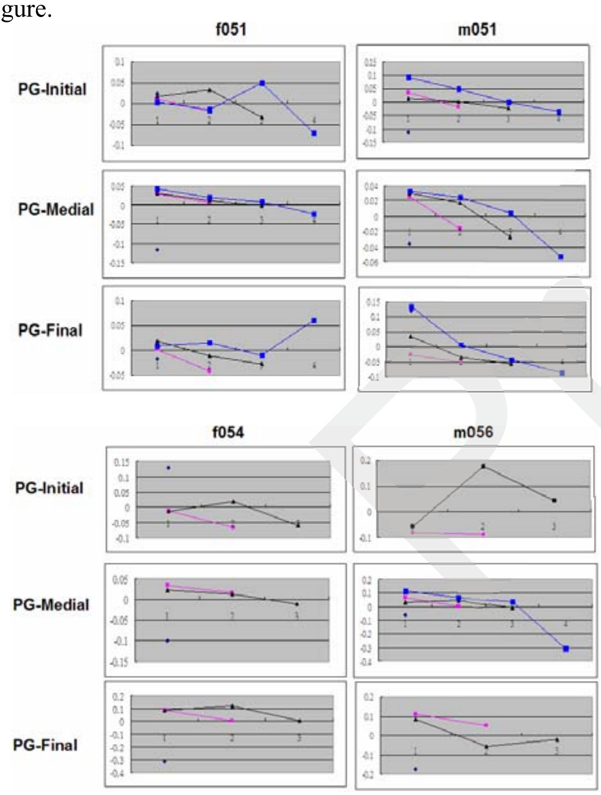

Figure3. PW model of Aa by *P*G positions. Each trajectory denotes PW model for specific PW length. The horizontal and vertical-axis indicate the Syl sequence index in PW and average Aa value, respectively.

Figure 4 shows PW model of Aa by PPh-position. Declination of Aa is found in PW-final of PPh-medial positions while PPh-medial positions exhibited a more general pattern across speakers. In addition, the sharpest declination slope is found at PPh-final positions across the board, indicating pre-boundary declination is systematic by prosodic unit.

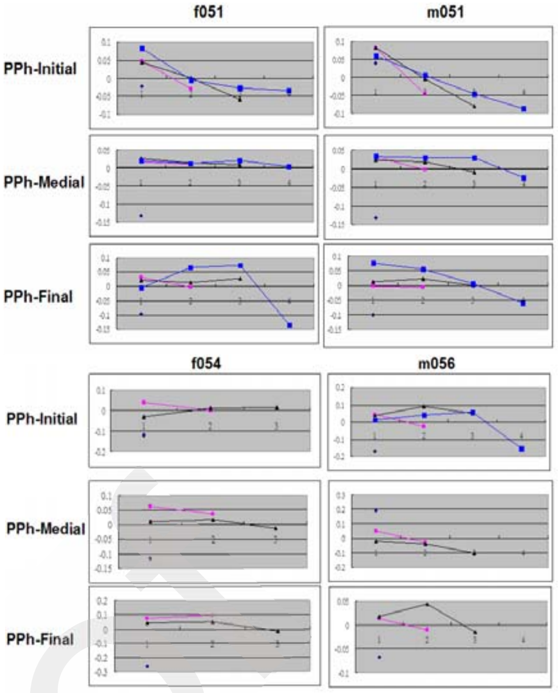

Figure4. PW model of Aa by PPh position. Each trajectory denotes PW model for specific PW length. The horizontal and vertical-axis indicate the SYL sequence index in PW and average Aa value, respectively.

D. BG & PG model of Ap

Figures 5 and 6 shows higher level Ap models above the PPh, namely, the PG & BG models. Both models show similar tendency across the 4 speakers. Figure 5 shows the within-PG adjacent as well as cross-over association of phrases by PG-position, namely, adjacency from PG-initial to –medial to –final and cross-over between -initial and –final. Figure 6 shows between-paragraph association and the –final to –initial contrasts in pitch.

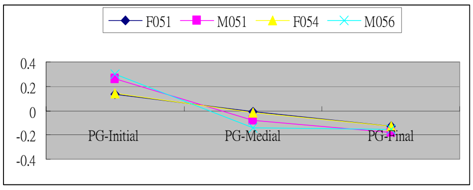

Figure5. PG model of A*p* by PG position. The horizontal and vertical-axis indicate the PG-position index and average Ap value, respectively.

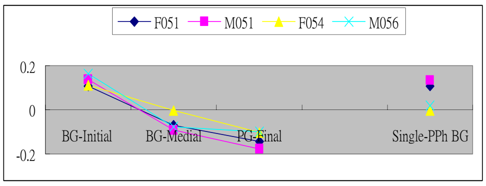

Figure6. BG model of A*p* by BG position. The horizontal and vertical-axis indicate the BG-position index and average Ap value, respectively.

Statistical analyses were performed on all Ap values to see if higher level contributions are significant, as shown in Table 4. All Ap values are classified by BG and PG index for ANOVA irrespective of PPh length effects. The number of category is 12 and the df is (1,7). Significant differences are found only among the BG and PG categories. The results imply that to output F0, contributions from higher-level larger-unit BG- and PG- positions are more significant than those from lower-level smaller-unit SYL and PW.

Table 4 ANOVA for Ap in different PG & BG position

Speaker	F051	M051	F054	M056
F-ratio	25.633	62.061	13.048	35.103
Prob	<=0.0001	<=0.0001	<=0.0001	<=0.0001

V. Discussion

We have shown from analyzing the F0 contours by the HPG layers across speakers speaking rate how contributions in addition to tones and intonation could be accounted for. In the Ap & Aa predictions presented above, layered contributions also support both adjacent and cross-over formation of global prosodic context. These results mirror our findings of previous HPG investigations in other supra-segmental acoustic correlates, namely, patterns of duration, amplitude and boundary pauses across fluent Mandarin speech. [1] [2] [3] The present study of F0 shows how at the syllables layer that correct prediction of Aa by tone identities amounts to only 40~45%; while the contribution from PW is 15~20%. In other words, it means only less than half of syllable tones can be predicted correctly. Or, tone-dependent information only contributes to less than half of correct prediction. Furthermore, by including contextual information of the next higher layer PW which specifies syllable neighborhood interaction, cumulative prediction accuracy at its best (65%) is still less than satisfactory. It is therefore clear how tone adjacency is insufficient to account for the F0 contours in fluent Mandarin speech. However, the contribution from higher layers BG and PG is about 7~35%, thus demonstrating the existence of additional F0 information from higher paragraph layers. We argue that these contributions from global prosodic information should not be overlooked. We noted also that the F0 patterns of speakers are more similar at the higher levels while variations are resulted at the lower ones, indicating global prosodic patterns are more uniform across speakers and speaking rate. In fact, individual variation occurs mostly by PW variations at both the PPh-initial vs. –final and PG-initial vs. –final positions. The above results suggest that higher-level effects are more stable across speakers than lower level ones [14], reflecting how large-scale global planning of semantic cohesion is a more general strategy used for prosody production along with lower-level individual planning. Therefore, it is feasible to assume that both small- and large-scale templates are employed by listeners during on-line speech processing. Large unit global look-ahead for top-down processing may require little information well ahead of production time.

VI. Conclusion

We have shown that in the F0 of Mandarin fluent speech, in addition to tones and intonation and global prosodic information and patterns also exist. In this light prosodic context should include both adjacency and cross-over associations of various sizes of prosodic units. Methodologically, we have also shown that when layer-by-layer analysis is adopted and lower level effects eliminated, higher level prosodic information could be accounted for, supporting our argument that tones and intonations are not sufficient to characterize the prosodic context of fluent Mandarin prosody. Most interesting of all is the relationships between PG-initial and –final, because within-paragraph –initial vs. –final signifies cross-over prosodic context while adjacent –final vs. –initial signifies between-paragraph discourse association. In summary, speakers do not plan or produce spoken discourse in discrete unrelated tones and intonations; bottom-up tone-and-intonation planning must interact with top-down discourse planning; the signals contain more than local information; and linear smoothing is simply insufficient. Therefore, the results and argument combined offer alternative thinking towards Mandarin ASR.

Future works include extending our present paradigm to spontaneous speech to further test how more understanding of discourse organization, higher-level planning and top-down processing could be applied to natural language processing and speech technology development.

Reference

[1] Tseng, C. "Prosody Analysis", Advances in Chinese Spoken Language Processing, World Scientific Publishing, Singapore, pp. 57-76, 2006.

[2] Tseng, C., Pin, S., 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, pp. 284-309, 2005.

[3] Tseng, C. and Lee, Y. "Speech rate and prosody units: Evidence of interaction from Mandarin Chinese", Proceedings of the International Conference on Speech Prosody 2004, pp. 251-254, 2004.

[4] Tseng, C and Chou, F. "Machine readable phonetic transcription system for Chinese dialects spoken in Taiwan" The Journal of the Acoustical Society of Japan (E) 20(3), 215-223, 1999

[5] Tseng, C, Cheng, Y. and Chang, C. Sinica COSPRO and Toolkit—Corpora and Platform of Mandarin Chinese Fluent Speech, *Oriental COCOSDA 2005*, (Dec. 6-8, 2005), Jakarata, Indonesia, 2005.

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

[7] Mixdorff, H. "A Novel Approach to the Fully Automatic Extraction of Fujisaki Model Parameters", Proceedings of ICASSP 2000, vol. 3, pp.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, pp. 873-876, 2003.

[9] Wang, C., Fujisaki, H., Ohno, S. and Kodama, Tomohiro. "Analysis and synthesis of the four tones in connected speech of the standard Chinese based on a command-response model", Proceedings of EUROSPEECH'99, pp. 1655-1658, 1999.

[10] Keller, E., and Zellner, K.,. "A Timing model for Fast French", York Papers in Linguistics, 17, University of York, pp.53-75, 1996.

[11] Zellner, K., and Keller, E., "Representing Speech Rhythm" Improvements in Speech Synthesis. Chichester: John Wiley, pp. 154-164, 2001.

[12] Tseng, C., Su, Z., "Boundary and Lengthening—On Relative Phonetic Information."

[13] The 8th Phonetics Conference of China and the International Symposium on Phonetic Frontiers, Beijing, China., 2008

[14] Tseng, C., Su, Z., "Discourse Prosody Context--Global F0 and Tempo Modulations", INTERSPEECH 2008, Brisbane, Australia, 2008.