Investigating F0 Reset and Range in Relation to Fluent Speech Prosody Hierarchy

Chiu-yu Tseng, Chun-Hsiang Chang and Zhao-yu Su

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

Abstract

We studied F0 reset and F0 range change in relation to a hierarchical multiple-phrase prosody framework for fluent speech that accounts for cumulative fluent speech output. Both F0 reset F0 range modifications are analyzed with respect to levels of boundary breaks within and across phrases. We found patterns in speaking strategy and gender difference. The findings are instrumental to enhance both our prosody framework and application to synthesis output.

1. Introduction

In a prosody framework for fluent speech postulated by our research team [1, 2], we stressed the significance of establishing cross-phrase characteristics and relationships across speech flow in narratives and/or spoken discourses. The hierarchical framework is based on perceived units located inside different levels of boundary breaks across speech flow, and specifies how different levels of prosodic units and boundary breaks cumulatively form multiple-phrase speech paragraphs. The hierarchical governing and constraining functions of Prosodic Phrase Grouping (PG) over phrases are illustrated schematically in Figure 1, in which the framework can also be viewed as a tree-branching organization for multi-phrase prosody. Units postulated were perceived prosodic entities. From bottom up, the layered nodes are syllables (SYL), prosodic words (PW), prosodic phrases (PPh) or utterances, breath group (BG) and prosodic phrase groups (PG). These constituents are, respectively, associated with break indices B1 to B5 (not shown in Figure 1 to keep the illustration less complicated.). However, B1 denotes syllable boundary at the SYL layer where usually no perceived pauses exist; B2 a perceived minor break at the PW layer; B3 a perceived major break at the PPhs layer; B4 when the speaker is out of breath and takes a full breath and breaks at the BG layer; and B5 when a perceived trailing-to-a-final-end occurs and the longest break follows. A modular acoustic model was subsequently constructed. [2]

Note that both the framework and prosodic units used, including boundary breaks, purposely made no reference to either lexical or syntactic properties so that it was possible to study possible gaps between these different linguistic levels and units. In the framework, the unit where intonation pattern applies is usually a PPh.

When a speech paragraph is relatively short and does not exceed the speaker's breathing cycle, the top two layers BG and PG collapse into the PG layer. Viewed from bottom upward, the framework also accounts for how PG groups PPhs and other lower nodes. The proposed hierarchical prosodic framework not only takes into consideration physiological constraint of breathing as well as cognitive constraint of speech planning, but also accounts for layered contributions from each prosodic level that cumulatively derive the overall prosodic output of multiple-phrase speech paragraphs.

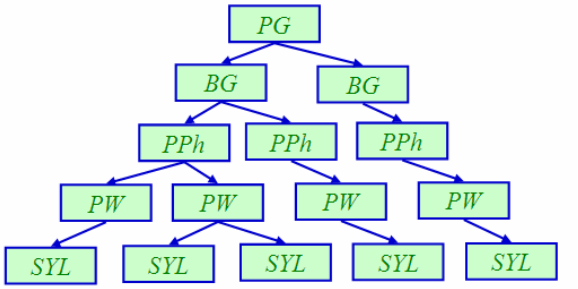


Figure 1*: A schematic representation of the hierarchical organization of multiple-phrase grouping on perceived units and boundaries.*

We have since further investigated the acoustic domains of prosodic characteristics within and across PGs in spoken discourses under the prosody framework described above. In the present study, we studied F0 reset within and across PGs in relation to pre- and post-reset pause durations [3], and further studied F0 range changes in relation to resets. Figure 2 shows how the afore-mentioned hierarchical prosodic framework could be elaborated to further accommodate build-up of a spoken discourse upward. In other word, how speech paragraphs represented as PGs in our framework could structurally build up spoken discourse or narrative.

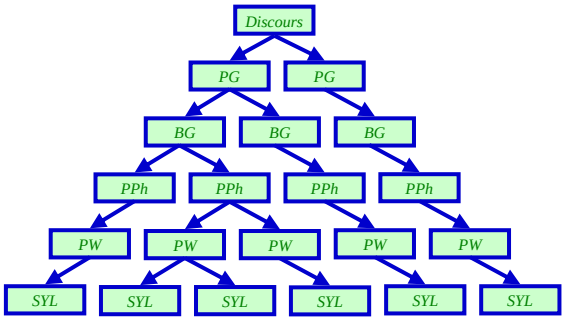


Figure 2*: A schematic representation of multiple-phrase grouping that form a spoken discourse.*

Each of such multiple-phrase prosodic phrase group is most notably marked by the prosodic characteristics of the first (PG-initial) and last (PG-final) prosodic phrases (PPh) that signal the beginning and ending of a speech paragraph, respectively. One of the major acoustic features associated with PG specified

positions is in F0, thereby making F0 reset and possible modifications of F0 range along speech flow worthy of more detailed investigation.

2. Speech Material:

Microphone speech recorded in sound proof chambers were used for the present study. One male (M051) and one female (F051) radio announcer, both under 35 years of age, read text at normal speaking rate of 200 ms/syllable [6]. The speech data consisted of readings of 26 paragraphs (11592 syllables in total) of text ranging from 85 to 981 characters per paragraph. Recorded speech data were first segmented using the HTK toolkit and human spot checked, then hand tagged by trained transcribers for perceived boundary breaks using the Sinica COSPRO Toolkit [4].

3. Method

We analyzed changes between F0 contours before and after the three bigger boundary breaks (B3, B4 and B5) [5]. Figure 3 shows a schematic diagram of an observed F0 contour. We observed two parameters in the F0 contour in Figure 2: (1.) the difference between the F0 contours before and after bigger boundary breaks, denoted as △F0 and (2.) the difference between the maximum values of the first F0 contour in the speech flow series before and after bigger boundary breaks in different positions, denoted as $\triangle F0_{Head}$.

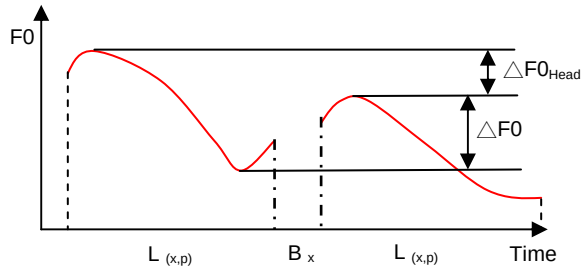


Figure 3*: A schematic representation of F0 contour. x is an integer ranging from 3 to 5.*

Only boundary breaks B3, B4 and B5 were further analyzed with duration patterns of prosodic units at both pre- and post-reset positions. That is, prosodic phrase (PPh), breath group (BG) and prosodic group (PG), respectively. At the BG layer, if there are only two PPhs, we then assumed the position relation between the two PPh lengths to be BG-Initial and BG-Final. To eliminate the variation between the male and female speakers, each set of data was normalized with the mean and standard deviation of the entire class. [3]

4. Analysis and Results

We analyzed F0 reset in multiple-phrase prosodic phrases by breaks and prosodic units, and further examined modifications of F0 range in relation to reset.

4.1. Statistics in △F0 and Pause by Break:

Break counts listed in Table 1 were analyzed. Figure 4 shows the prosodic features of M051 and F051 in △F0 and pause (break). Similar results are found of M051 and F051. The results indicate that both B5 and B4 have a higher △F0 than B3, but B5 has a longer pause than B4 or B3. Figure 4 also shows very little difference in the prosodic characteristics of B5 and B4 in △F0.

Table1*: Break counts in data-M051, F051.*

M051	B3	B4	B5
Count	798	174	61
F051	B3	B4	B5
Count	724	168	68

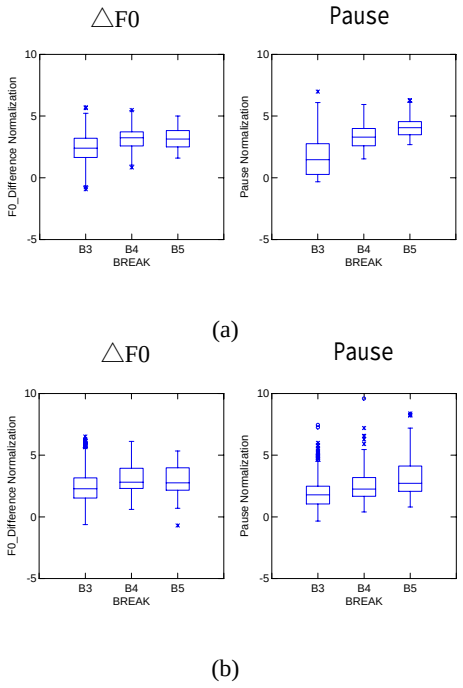


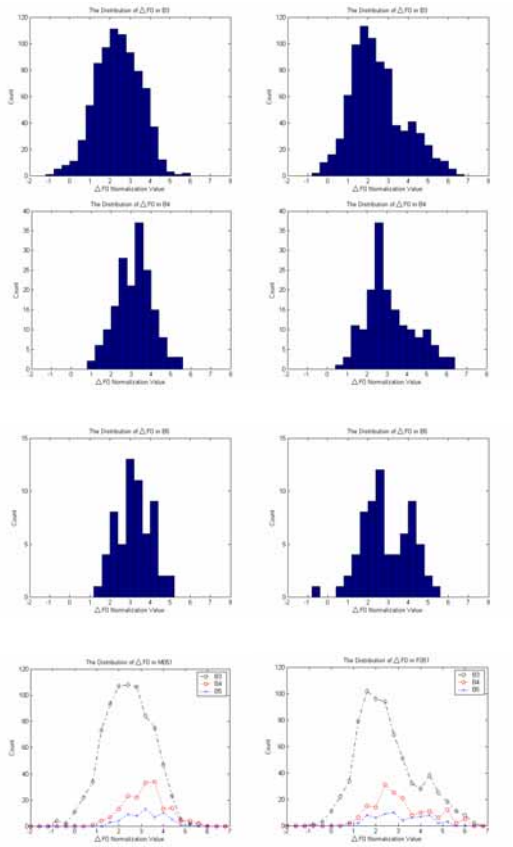
Figure 4*: (a) Prosodic characteristics of M051. (b) Prosodic characteristics of F051.*

Although there is no significant difference with respect to △F0 between B5 and B4 (p-value>0.01), significant difference was found with respect to break ($z = 6.58$, $p<0.01$ in M051 / $z = 2.958$, $p<0.01$ in F051). However, we noted that the distinction between B4 and B5 is not clear in △F0. Therefore, it is reasonable to assume that that the pauses themselves are is used as the main cue of distinction irrespective of F0 reset between B5 and B4.

Figure 5 shows the distribution of △F0 by break in M051 and F051. Here we can see that the distributions of △F0 in M051 are different from those of F051. The distributions of △F0 in M051 tend to symmetrical, whereas the distributions of △F0 in F051 tend to be positively skewed. Figure 5(b) shows that F051 has a negative △F0 value in B5. In addition, the range of the distributions of △F0 in F051 is wider than that in M051. We found that △F0 of F051 in B5 is a negative value compared to M051 in the same sentence.

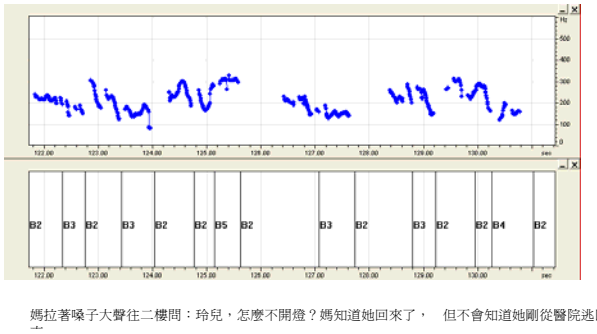
We note with interest here that male and female speakers use different styles and patterns in reading. Further more, the

speakers also parse and interpret the same text differently, which was further manifested through different choice of boundary breaks and perhaps voice quality. To illustrate the point, Figure 6 shows that F051 and M051 used different boundary breaks in the same sentence read. Figure 6(a) shows that F051 used a voice quality technique which led our transcribers to determine the boundary as B5 in the sentence. However, Figure 6(b) shows that M051 used a modal method to read the sentence; so our transcribers did not perceive the boundary break as B5 at the same position.

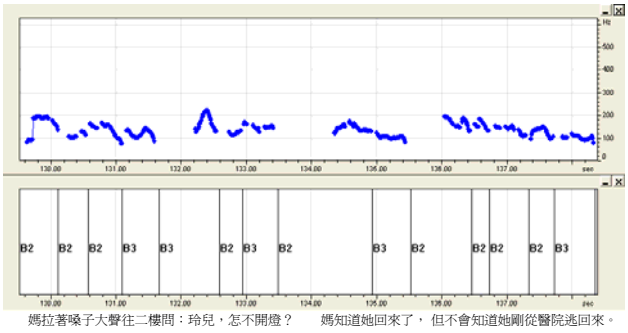


(a) (b)

Figure 5*: (a) The distribution of △F0 by Break in M051. (b) The distribution of △F0 by Break in F051.*



(a)



(b)

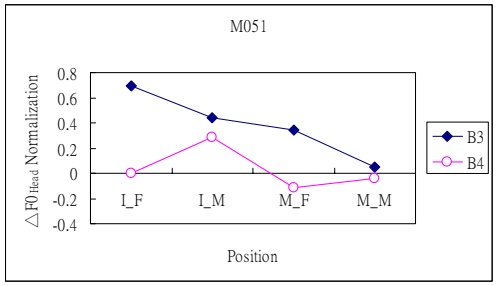
Figure 6: (a) An example of F051. (b) An example of M051.

4.2. Statistics in $\triangle F0_{Head}$ by Position:

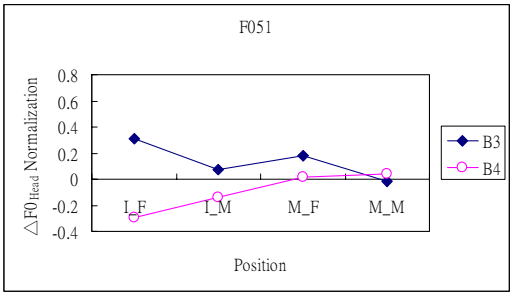
Boundary breaks were grouped into the classes by position and listed in Table 2. PG related position tags PT–Initial, -Medial and –Final were grouped into four classes by the prosodic hierarchy: Initial-Final (I_F), Initial-Medial (I_M), Medial-Final (M_F) and Medial-Medial (M_M). B5 did not appear in the data as often as B3 and B4, resulting scarce of the number of B5 in position "I_F". Hence, the statistics of B5 should be ignored in following analysis.

Table 2: *Break counts by position in data-M051, F051.*

M051				
Position \ Break	B3	B4	B5	Sum
I_F	83	34	1	126
I_M	164	42	12	210
M_F	193	45	14	336
M_M	358	53	34	445
Sum	798	174	61	1033
F051				
Position \ Break	B3	B4	B5	Sum
I_F	78	27	1	125
I_M	156	46	13	196
M_F	207	47	14	321
M_M	283	48	40	371
Sum	724	168	68	960



(a)



(b)

Figure 7: *(a) $\triangle F0_{Head}$ of the different M051 position. (b) $\triangle F0_{Head}$ of the different F051 position.k*

Figure 7 shows the mean of $\triangle F0_{Head}$ by position in speakers M051 and F051. We can see that the $\triangle F0_{Head}$ of B3 in position "I_F", "I_M" and "M_F" is bigger than B4. $\triangle F0_{Head}$ of "I_F" in B3 is the biggest among the four positions. In addition, $\triangle F0_{Head}$ of position "M_M" of B3 and B4 is similar for both M051 or F051.

4.3. Change of F0 Range:

The base form of our framework [1, 2] specifies that the most notable F0 reset occurs in the BG initial position [7, 8], i.e., the first PPh into a PG. However, more detailed analysis revealed from the speech data that quite a few initial phrases were produced with a narrower F0 range where the actual F0 reset is moved forward to the next phrase. Figure 8 shows an example of a narrower initial F0 range followed by a forward shift of initial F0 reset with wider F0 range. We noted that these initial phrases that did not contain a notable F0 reset are usually produced with a relatively much narrower F0 range, and are short in duration. Further observation showed that these short units were usually transitional or carry-over phrases in the nature of "on the other hand", "in other words" instead of phrases of concrete content. Future and further syntactic and semantic analysis should yield comparable account and/or rules to explain the phenomenon. For the time being, we performed statistical analysis on the probability of the occurrence of these 'transition phrases' on the basis of observed acoustic characteristics only. When distinct F0 resets (medial phrase F0 Head >1.5*initial phrase F0 Head) took place in the medial position of BG, as illustrated in Figure 9, we defined the situation as a 'transition induced shift'. Results of statistical analysis are shown below. We found that the transition occurs in short phrases (usually one PW) frequently and the probability of such 'transition' appearing in long phrases (above two PW) diminishes significantly.

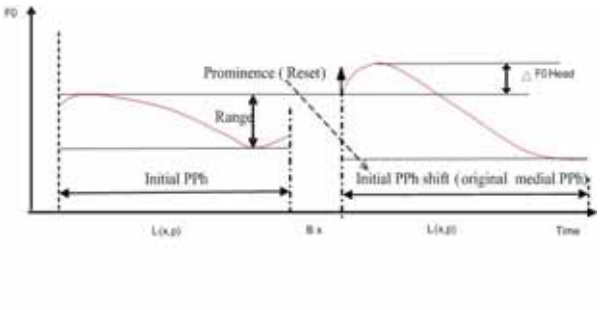
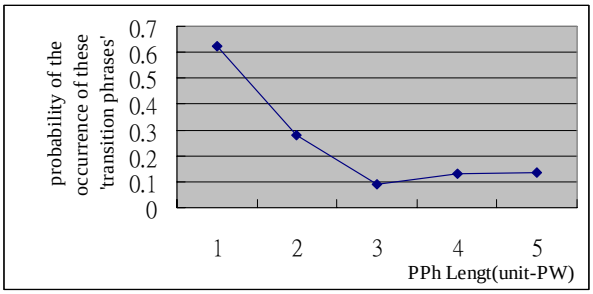
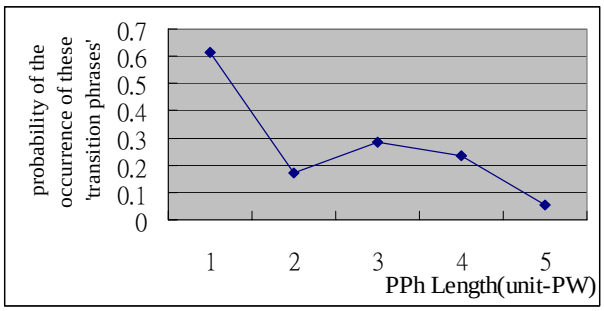


Figure 8: *A schematic representation of narrower initial F0 range followed by a forward shift of initial F0 reset.*

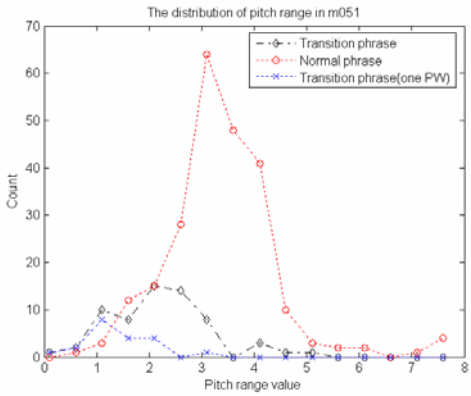


(a)

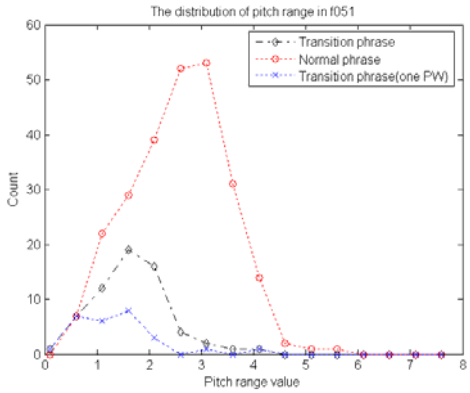


(b)

Figure 9: Probability of 'transition phrases' among different PPh Length in speaker m051 (a) and speaker f051 (b)



(a)



(b)

Figure 9: Histograms of distribution of F0 range where Figure 9 (a) and (b) show 'Transition Phrase' and 'Normal Phrase' produced by speaker m051 and f051, respectively. Note that the distribution of the 'transition PPh range' tends to be smaller than normal situations.

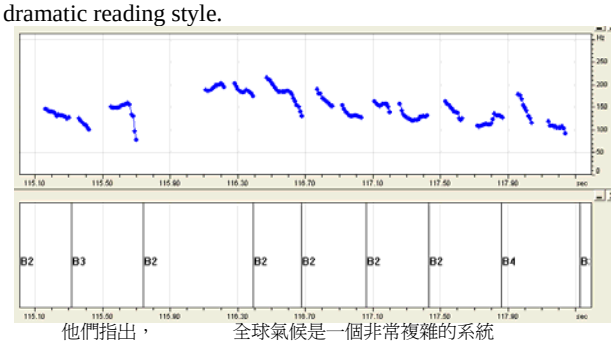
We noted that the probability of 'transition phrase' occurring in short BG-initial PPh is large, especially when the initial PPh into a BG or PG marked by boundary break B3 was in fact only one PW. We believe this feature could be incorporated into our prosody framework to accommodate text analysis for prosody prediction on the one hand, and further enhance prosody model on the other hand. It can also be applied to improve output naturalness in unlimited TTS.

4.4. F0 Contour Observation in relation to F0-reset forward shift:

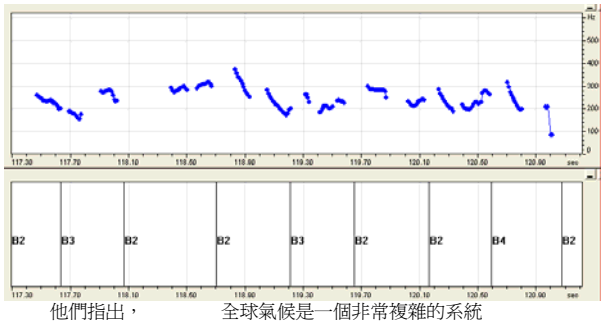
We have shown in our analysis [see Figure 6] that male and female speakers use different F0 reset patterns and ranges, but behaved similarly regarding PG-initial transition phases. Further observations of overall F0 contour patterns also revealed less speaker variation, as illustrated in Figure 10. Figure 10 show similar overall F0 contour patterns produced by M051 and F051. Both speakers used narrower F0 range at the initial phrase, and moved the actual F0 reset forward to the next phrase. This indicates that constraints from syntax and semantics override speaker and/or speaking-style differences.

Figure 11 show that M051 and F051 chose different focus [9] and boundary break B3. The results are different global output patterns. The male speaker m051 (shown in Figure 11 (a)) read the sentence with apparent overall declination. However, the female speaker made us bigger F0 reset in B3 to express her emphasis in phrase.

Figure 12 shows that M051 and F051 produced opposite F0 contour patterns for the same speech paragraph. We observed that the male speaker used a modal pattern of reading. Contrary to the male speaker, the female speaker decreased the initial of F0 while shifting the highest F0 reset much more forward. The different choice of focal points thus rendered different output effect. The female speaker's choice has thus created a more dramatic reading style.

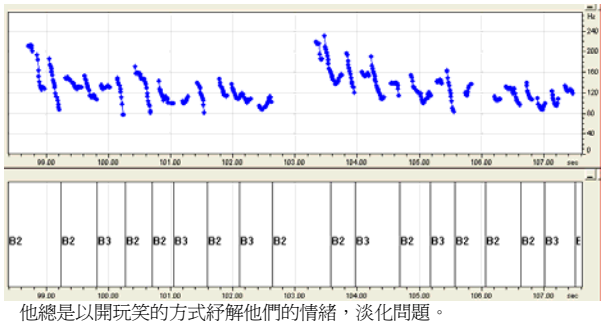


(a)

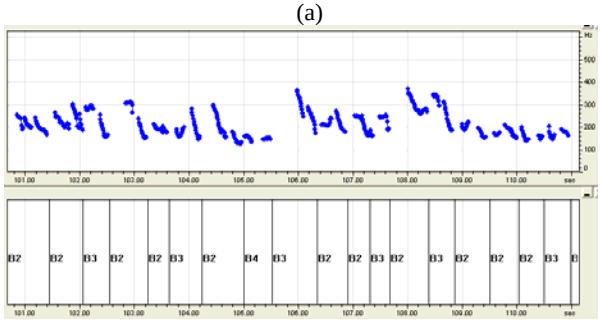


(b)

Figure 10*: (a) An example of M051.* *(b) An example of F051.*

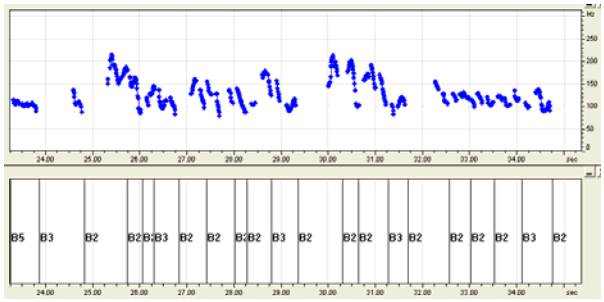


(a)

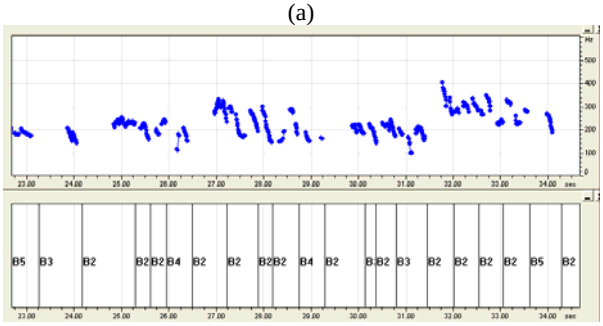


(b)

Figure 11*: Readings of identical text produced by male speaker M051 (a) and F051 (b). Note how each speaker begins at a different reset level, focuses at different points and how these factors affect the global contour patterns.*



(a)



是的，如果媽還有其他孩子，也許這小室會有活潑的氣息，
如果她有兄弟或姐妹，或許她今天不必從精神科逃出來。

(b)

Figure 12: Readings of identical text produced by male speaker M051 (a) and F051 (b). Note where different points of highlight of F0 reset occurred and how the global contour looked.

5. Discussion and Conclusion

We have demonstrated through analysis of F0 reset and F0 range that our model of fluent speech prosody could be further enhanced by incorporating an initial transitional phrase into the framework. These initial and relatively short (up to 5 syllables) transitions should be specified as narrower in F0 range and with no F0 reset. When initial F0-range narrowing occurs, the PG-initial F0 reset is moved forward to the next phrase. These narrowing and reset fronting allows the prosody framework to accommodate features necessary in narratives or spoken discourses, and renders more natural melodic output for running speech.

We have also found similarities and differences in F0 reset and F0 range for male and female speakers. For similarities, we observed expected structure-induced narrowing and reset occurring across speakers. For differences, we observed different locations and patterns of F0 resets. The male speaker read most of the provided text in a modal manner, with almost no focal points across the phrases, thus yielded a more monotonous reading. The female speaker, on the other hands, employs a wider F0 range and large magnitude of F0 reset, thus yielding a more theatrical reading of the same text. However, in our analysis we were still able to distinguish the transition PPh range from the normal PPh range in the male speech more easily than female speech by observing variability between the "transition PPh range" and "normal PPh range", as shown in Figure 9. In other words, normal vs. transition phrase (one PW) distinctions can be obtained from the distribution of transition for the male speaker, but not for the female speaker. We believe these initial investigations serve as refinement to our prosody framework; they can also be implemented speech synthesis for more natural and gender-oriented prosody output.

6. Reference

[1] Tseng, Chiu-yu, Pin, Shao-huang and Lee, Yeh-lin (2004). “Speech prosody: Issues, approaches and implications,” in *From Traditional Phonology to Modern Speech Processing* (語音學與言語處理前沿), Fant, G., Fujisaki, H., Cao, J. and Xu, Y., Eds Foreign Language Teaching and Research Press (外語教學與研究出版社),417-437, Beijing, China.

[2] Chiu-yu Tseng, ShaoHuang Pin and Yeh-lin Lee, Hsin-min Wang and Yong-cheng Chen, “Fluent Speech Prosody: Framework and Modeling” (in press and available on-line May 16, 2005) Speech Communication, Special Issue on Speech Prosody.

[3] Chiu-yu Tseng and Bau-Ling Fu(2005). "Duration, Intensity and Pause Predictions in Relation to Prosody Organization" In *interspeech 2005*.

[4] Sinica COSPRO and Toolkit http://reg.myet.com/registration/corpus/main.asp

[5] Tseng, Chiu-yu and Pin, Shao-huang (2004). "Modeling Prosody of Mandarin Chinese Fluent Speech via Phrase Grouping" .In *Proceedings of Oriental-COCOSDA2004.*

[6] Tseng, Chiu-yu, Cheng, Yun-ching, Lee, Wei-shan and Huang, Feng-lan (2003). "Collecting Mandarin Speech Databases for Prosody Investigations ".In *Proceedings of Oriental- COCOSDA 2003*

[7] G. MÖHLER, J. MAYER(2001), "A discourse model for pitch-range control". In *Proceedings of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis.*

[8] Möhler, Gregor, Mayer, Jörg (1999): "A method for the analysis of prosodic registers", In *EUROSPEECH'99*, 735-738.

[9] Li Aijun(2002): "Chinese Prosody and Prosodic Labeling of Spontaneous Speech" In *SP* 2002.