Prosodic Fillers and Discourse Markers–Discourse Prosody and Text Prediction

Chiu-yu Tseng, Zhao-yu Su, Chun-Hsiang Chang and Chia-hung Tai

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

Abstract

Mandarin Chinese fluent speech prosody is characterized by a hierarchical multiple-phrase structure that specifies how speech paragraphs are constituted via Prosodic Phrase Grouping. Hence we view spoken discourse prosody as yet another higher node treats PGs (Prosodic Phrase Groups) as sister constituents. The goals of present study are two fold: one is to study how speech paragraphs are connected in speech flow; another is to derive prosody prediction from text analysis. Investigating cross-phrase F0 range narrowing and F0 reset with boundary information, we further conducted corresponding text analysis for prosody prediction. Results revealed two types of PG connectors, one is redundant Prosodic Fillers (PF) that are mostly duration triggered and manifested through narrowed F0 ranges; another is obligatory Discourse Markers (DM) that are lexically and/or syntactically triggered and manifested through widened F0 ranges and resets. Both could be predicted from text analysis. We believe this is a significant step forward towards understanding the organization of discourse prosody. It could also be applied to speech synthesis and/or unlimited TTS for prosody enhancement.

1. Introduction

In the hierarchical fluent speech prosody framework we proposed [1, 2], we focused on establishing cross-phrase prosodic relationships of speech paragraphs in narratives and/or spoken discourse, and treated Prosodic Phrase Group (PG) as a speech unit where phrases under grouping became prosodic sister constituents thus requiring individual phrase intonations to systematically adjust and modify. The hierarchy also specifies layered contributions that cumulatively make up output prosody. In short, the PG framework is based on perceived units located inside different levels of boundary breaks across speech flow, and specifies boundary breaks systematically as well. It also specifies how respective levels of prosodic units and boundary breaks cumulatively contribute to final prosody output. By the same logic, we view narratives and discourse prosody as yet another higher node/layer that associates PGs into sister constituents. The question then is what some of the prosodic features of narratives and spoken discourse are and how we could systematically account for the derivation of discourse output prosody from multiple-speech paragraphs.

Based on the PG framework described above, we have further investigated the acoustic domains of prosodic characteristics within and across PGs in narratives. In an earlier study [3] of F0 range variation and F0 reset locations, we found that PGs may sometimes begin with a short period of F0 narrowing while the PG-initial phrase typified by a F0 reset was shifted forward to the next following phrase. We called these transitional prosodic words (PW) or prosodic phrases (PPh) Discourse Markers (DM), and claimed that they were prosodic fillers that can be deleted from speech flow. In the following presentation, we will show that further analyses of transitional PW or PPh also revealed widened F0 range with noticeable F0 reset, thus further thus revised our definition by distinguishing Prosodic Fillers (PF) from DM. In addition, we added corresponding text analyses of PF and DM and found interestingly, both could be predicted from text. Together with prediction of boundaries and boundary breaks, prosody prediction from text analysis now accommodated more and more prosodic features.

Figure 1 is a schematic representation of how the afore-mentioned multi-phrase prosodic framework PG could be elaborated to further accommodate building-up of narratives or spoken discourse, and how the framework accommodates PF and DM. Units postulated were perceived prosodic entities rather than phonetic and/or phonological components. 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 due to space limit.). B1 denotes syllable boundary at the SYL layer where usually no perceived pauses actually 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 corresponding modular acoustic model had been constructed [2]. 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 one layer, namely, the PG layer. This 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.

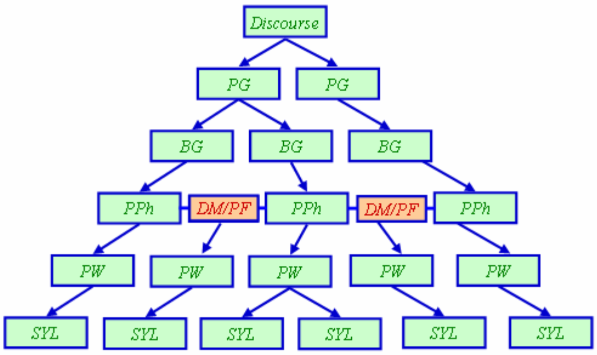


Figure 1: A schematic representation of how PGs form spoken discourse and where DM (Discourse Marker) and PF (Prosodic Filler) are located.

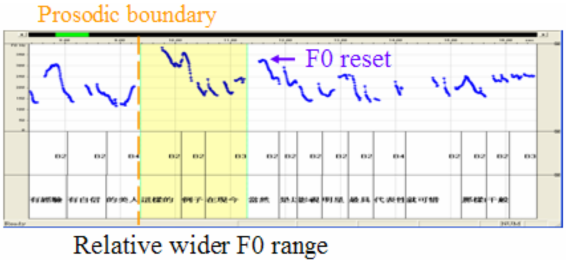
2. Materials

Speech data and corresponding text are analyzed and compared. Mandarin Chinese speech data from one male (M051P) and one female (F051P) speakers were used (Sinica COSPRO Database [4]) was analyzed. Both speakers are radio announcer by profession and aged under 35 at the time of recording. Each speaker read text of 26 discourses (11602 syllables in total at 85 to 981 characters per paragraph) in sound proof chambers at normal speaking rate of 200 ms/syllable. Pre-analysis annotation included automatically labeled segmental identities by the HTK toolkit in SAMPA-T notation, followed by subsequent manual tagged boundary breaks and manual spot-checked segmental alignments by trained transcribers using the Sinica COSPRO Toolkit [3]. The extracted features including pause, duration, F0, and intensity of the speech data had been normalized. Text of the 26 discourses was also analyzed for boundary breaks and compared with speech analysis results.

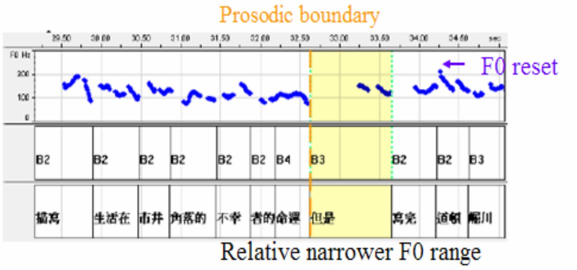
3. Method and Analysis

3.1. Overall F0 Range Narrowing vs. Widening

Our focus was on how successions of PGs or BGs across spoken discourse, in particular, whether any transitional units exist in between. Figures 2(a) and 2(b) are examples of speech data where portions of the final end of a preceding PG and the initial portion of the following PG are shown. The yellowed block indicates a transitional phrase between the PGs. Figure 2 (a) illustrates a transitional phrase with widened F0 range whereas Figure 3 illustrates a transitional phrase with narrowed F0 range.



(a) A connective phrase with relative wider F0 range



(b) A connective phrase with relative narrower F0 range

Figure 2: The transitional phrase between PGs as shown in yellow block shows it could be produced with either wider F0 as in (a) or narrower F0 range as in (b).

3.2. Acoustic Analysis of Speech Data

In an earlier study [3] we observed that F0 range narrowing sometimes occurred after a PG boundary. If we treat the narrowed section as PG-initial phrase, then contrary to BG- or PG-initial prosodic specifications [1, 2], no apparent F0 reset occurred. Figure 3 shows two F0 contours: 1. Solid line represents normal or regular PG strings, namely, no F0 range narrowing between two BG's/PG's. Or, no F0 range narrowing occurred after PG-final or BG-final. 2. Dotted line represents PG strings where narrower F0 range and smaller F0 reset after PG-final or BG-final boundaries. We observed one parameter that defined these narrowed F0 ranges in the F0 contour in Figure 3. $\triangle F0_{Head}$ is the difference between the maximum values of the second F0 contour and the third F0 contour in the speech flow. When $\triangle F0_{Head}$ is smaller than -1.5, as illustrated by the dotted red line in Figure 3, we define it as a PF. We reported that the semantic load of PF is insignificant and could be removed from speech without causing confusion [3].

Now we further define that when △F0head is bigger than -0.3, as illustrated in blue, we define the phrase with widened range as a DM.

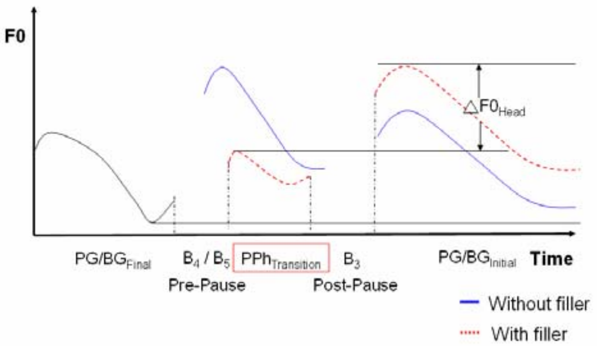


Figure 3: A schematic representation for consecutive PGs with and without a transitional phrase.

We note here that PF and DM assume different roles in discourse prosody. PF appears to be more surface prosodic variation, its function more like filled pauses. DM appears to be attention-calling devices used by speakers to draw attention; its function thus both prosodic and semantic. Both PF and DM are transitions that connect PGs and are therefore major boundary features of discourse prosody.

3.3. Boundary Effect on F0 Reset

In this section, we compare preceding PG boundary effect on following F0 reset with and without transitions. We defined the value of F0 reset as the difference between the maximum F0 values of a PG-initial or BG-initial and the minimum F0 values of the preceding PG-final or BG-final. When a transition PPh appears after a BG- or PG-final, maximum F0 value often moved to next PPh, namely, the second F0 contour of dotted line in Figure 4.

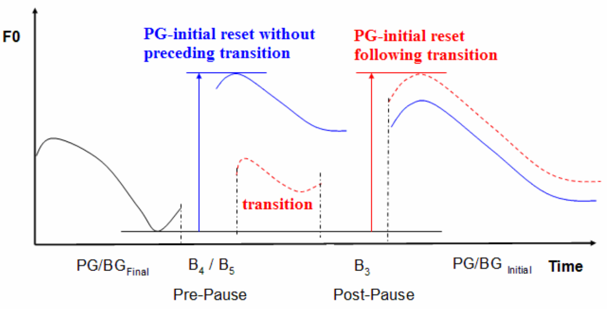


Figure 4: Schematic representation of F0 reset in normal and transitional phrases

Figure 4 also illustrates how we compared F0 reset of the boundary information between two PGs with and without transitional phrases. In other words, when two PGs occurred in succession without filler as transition, the F0 reset after a PG boundary is the reset of another PG-initial. When a filler transition occurred between PGs, we defined F0 reset at the phrase after the filler which was in fact the initial phrase of the following PG.

Figure 5 shows the comparison of F0 reset with and without transitional phrases for both speakers.

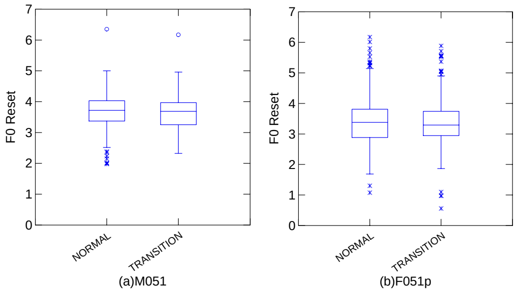


Figure 5: Comparison of F0 reset between normal and transitional situations from two speakers (a) M051 and (b) F051

Both the male and female speakers exhibited similar reset patterns in conditions with and without transitions shown in Figure 5. Filler transitions occurred in speech data produced by both speakers. Therefore it would be interesting to further study the semantic and grammatical properties of these transitions through corresponding text analysis.

3.4. Text Analysis to Predict PFs and DMs

Text analysis of the 26 discourses was performed, focusing on fillers and markers in relation to lexical, syntactic and discourse structure.

3.4.1. PF

Lexical and syntactic analyses of corresponding text showed that PF demonstrated the following features:

(1.) Semantically triggered fillers. When a post-B4 or B5 PPh contains lexical items meaning "說 say", "指出 point out", or "表示 indicate", concrete content usually follows, and in speech it is likely to produce these lexical items with narrowed F0 range. In speech, this implies the upcoming of concrete contents. Figure 6 is an example.

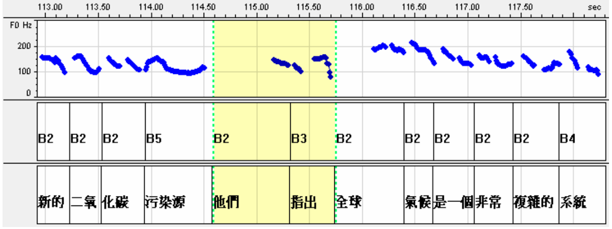


Figure 6: An example from speaker M051P where yellowed background indicates filler "they point out". Note how F0 reset occurred during filler.

(2.) Syntactically triggered fillers. When a prepositional phrase occurred at post-B4 or B5 PW or PPh is, it is usually followed by the main phrase of a sentence. The prepositional phrase tended to be produced with narrowed F0 ranges in speech and hence a PF while the following main phrase would usually receive more stress and begin with a F0 reset. Figure 7 is an example where the PF is an 8-character phrase.

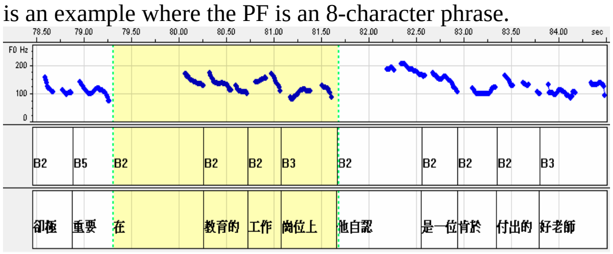


Figure 7: An example from speaker M051P where yellowed background indicates prosodic filler "on the ground as an educator". Note how F0 reset occurs after the filler.

(3.) Duration triggered fillers. When a post-B4 or B5 PW or PPh is a short word, usually 2 to 3 syllables at most, it is usually produced with narrowed F0 range in speech. The reason may perhaps be the fact that when short PW that ix not the focal point of expression it could become fillers in speech. Figure 8 is an example.

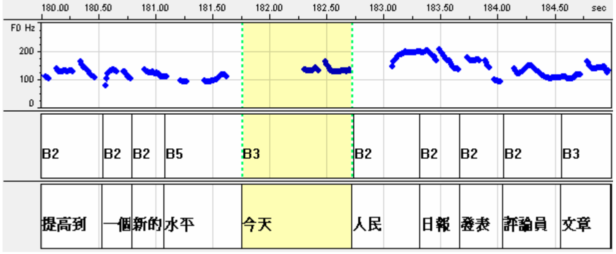


Figure 8: An example from speaker M051P where yellowed background indicates prosodic filler "today". Note how F0 reset occurs after the filler.

3.4.2. DM

Transitional words and phrases such as “but, however, at the same time, by the same logic…” that connects two parallel structures are usually produced as DM. These features could easily be derived from lexical analysis. When transitional words occurred after a paragraph boundary usually signaled by punctuation mark period in text, and after BG or PG boundary B4 or B5 in speech, they could be cue phrases to signal contrast and call attention, and therefore are produced with widened F0 range and even reset. Figure 9 is an example.

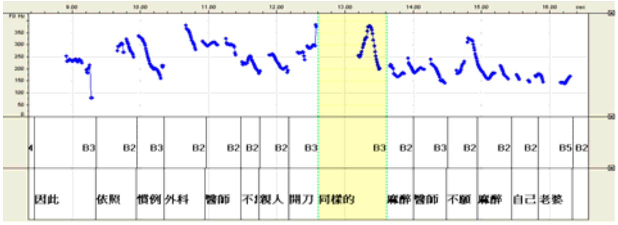


Figure 9: An example from speaker M051P where yellowed background indicates discourse mark “in the same way”. Note the main phrase occurs after the mark.

3.5. F0 Range in Speech vs. Punctuation Marks in Text

Corresponding text analysis vs. speech data also included punctuation marks. Our hypothesis is that when reading text, punctuation serves as indicators of speech planning, in particular boundary and boundary breaks, and therefore may affect F0 range variation in speech output. We found that the punctuation mark colon in text tended to trigger the preceding phrase to become a PF in speech as Figure 10 illustrates.

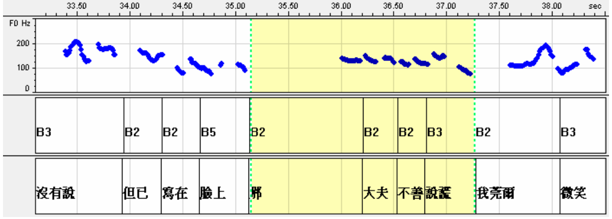
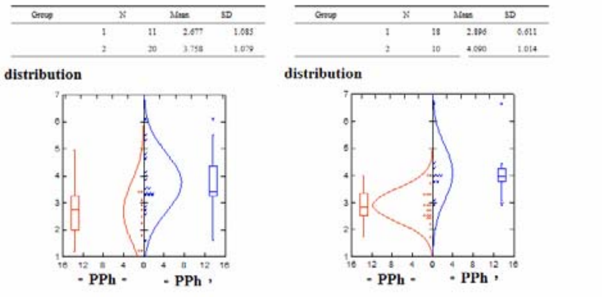


Figure 10: An example from speaker M051P where yellowed background indicates prosodic filler “then”.

Comparisons are made between punctuation analysis and speech data, as shown in Figures 11 and 12.

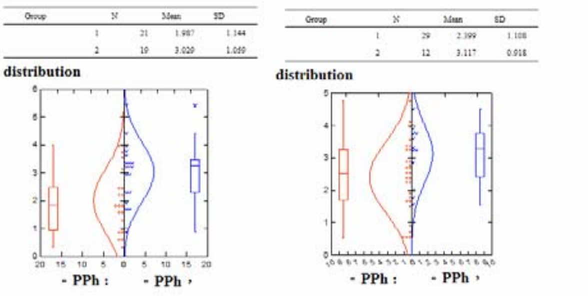


Group	N	Mean	SD
1	11	2.677	1.085
2	20	3.758	1.079

Group	N	Mean	SD
1	18	2.896	0.611
2	10	4.090	1.014

(a) (b)

Figure 11: Comparison of punctuations and speech data for two speakers. Blue lines indicate distribution of relative short portion of speech data with wider F0 range where in text it is an actual initial of a new paragraph which is preceded by a period and followed by a comma. Red lines indicate the distribution of relative short portion of speech data with narrower F0 range where transitional word or phrase between a preceding period and a following colon.



Group	N	Mean	SD
1	21	1.987	1.144
2	19	3.029	1.059

Group	N	Mean	SD
1	29	2.399	1.108
2	12	3.117	0.918

Figure 12: Comparison of punctuations and speech data for two speakers. Blue lines indicate distribution of relative short portion of speech data with wider F0 range where in text it is an actual initial of a new paragraph which is preceded by a period and followed by a comma. Red lines indicate the distribution of relative short portion of speech data with narrower F0 range where transitional word or phrase between a preceding and a following period. In other words, a transitional sentence between two long paragraphs in text.

These results were then used to build a model to predict PFs and DMs from text for speech synthesis application.

3.6. Constructing a Model to Predict PFs and DMs from Text

We reported earlier [3] that discourses markers carried little linguistic/semantic weight and are optional in speech. However, analyses presented above demonstrated that only duration-triggered fillers would not cause the sentence to become ill-formed whereas semantic and syntactic fillers would. Therefore, duration PF [5] is optional prosodic enhancers but other PFs and DMs carry different degree of semantic and/or syntactic loads as well. Hence we further experimented whether it was possible to predict DM and PF from text as part of prosody prediction. Figure 13 illustrates how we integrated the analyses performed to build the prediction model.

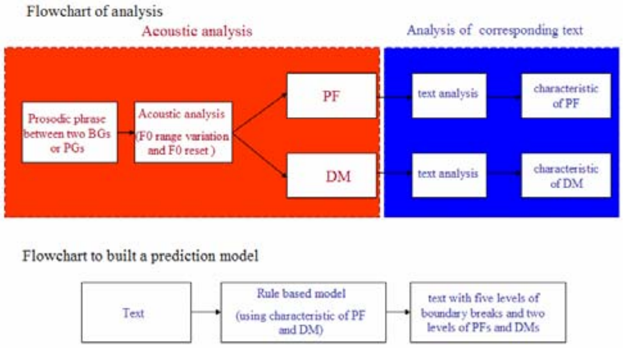

Flowchart of analysis
Acoustic analysis
Analysis of corresponding text
Prosodic phrase between two BGs or PGs
Acoustic analysis (F0 range variation and F0 reset)
PF
DM
text analysis
characteristic of PF
text analysis
characteristic of DM
Flowchart to built a prediction model
Text
Rule based model (using characteristic of PF and DM)
text with five levels of boundary breaks and two levels of PFs and DMs

Figure 13. The upper column summarizes acoustic analyses (shown in read block) and text analyses (shown in blue block) performed; the lower column summarizes how a prediction

model utilizes both results to further constructing a prediction model for PF and DM.

A rule-based model was constructed to predict two levels of fillers from text. Figure 14 summarizes the algorithm:

```
ProduceDM/PF
(1)     Input Data= text piece with boundary breaks
        for each PPh in Input Data
(2)         if (CheckKeyWord(PPh) is true)
            Add to Candidate List
        for each PPh in Candidate List
(3)         DetermineLegality(PPh)
        Output Data: text with PF, DM and boundary breaks
```

Figure 14: Procedures to predict two levels of transitions PF and DM.

Predictions are two-fold. The model first specifies prediction of boundaries and breaks, then further predicts where DM and PF may occur. In Step (1) we read in text of a multiple-phrase paragraph with punctuations to predict boundary breaks B1 to B5. Elaborating an earlier statistical model [6] that predicted one level of boundary break (B2), we further predicted all 5 boundary breaks, thus converting the text with punctuations only into text with predicted boundary breaks. In Sept (2) the algorithm checks every PPh to see whether it is a cue phrase such as “不過 however”, “但是 but”, “在 in….” “而 save for....”, “說 say”, “表示 state, indicate”…etc. (See 3.1(1.)) add to the candidate lists. In Step (3) we determine whether a candidate is PF or DM. To avoid unrelated uses of cue phrase, we also use parts of speech as constraints. The end result after applying these three steps are now text with five levels of boundary breaks and two levels of discourse transitions PF and DM.

4. Results and Discussion

4.1. Results of Predicting Prosody from Text

We tested the model on text and compared the results with speech data to see how successful the prosody prediction fit the speech data. Performance evaluation is based on precision, recall and F-score.

$$\text{Precision} = \frac{\text{numbers of correctly predicted PW boundary}}{\text{numbers of predicted PW boundary}} \quad (1)$$

$$\text{Recall} = \frac{\text{numbers of correctly predicted PW boundary}}{\text{numbers of real PW boundary}} \quad (2)$$

$$F-score = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

Table 2 shows the results from the baseline whereby boundary breaks B4 and B5 are used to predict F0 Reset position without considering fillers and markers.

Table 2. Baseline without fillers

	Recall	Precision	F-Score
F051	0.602	0.616	0.609
M051	0.571	0.541	0.556

Table 3 shows predictions from the rule based filler predictions. The predictions are using predicted discourse marker, B4 and B5 to predict F0 Reset position.

Table 3. Baselines with fillers

	Recall	Precision	F-Score
F051	0.667	0.551	0.603
M051	0.635	0.486	0.550

Comparing Tables 2 and 3, we report that the F-Score value of text with fillers was comparable to that of only using B4 and B5. Although the F-Score value of text with only using B4 and B5 was slight greater than that of text with fillers, by considering fillers we were in fact taking facts from speech data into account. Note also that the recall of text with fillers was greater than the without case. As for the precision rate, the reason could be due to speaker variation or intension variation and merits further study in the future.

In Table 4 we present comparison of cross-speaker consistency using speaker M051 as correct answer. Results showed that speakers may interpret of the same text somewhat differently while our prediction results were in fact better than speaker consistency. The reason may be that we have grasped the core content of text under analysis at the current stage, but did not include alternative interpretation to accommodate speaker intention.

Table 4. Consistency of two speakers

Recall	Precision	F-Score
0.577	0.535	0.555

5. Conclusions

Our earlier analyses of F0 range variation as well as F0 reset of narratives showed F0 narrowing across PGs. We studied these narrowed ranges and argued that there were Prosodic Fillers (PF) comparable to fillers and/or filled pause in spontaneous speech. Though these PFs connected speech phrases into speech paragraphs and paragraphs into spoken discourse, they could be treated as redundant fillers that provided more output prosodic variation [3]. Further analyses presented above of transitions between PGs showed that distinctions should be made between PF from Discourse Markers (DM) for DMs are obligatory semantic- or syntactic-components between PGs and therefore are not fillers in speech output. We have further showed how both PF and DM could be predicted from text analyses, thus making prosody prediction closer to natural speech output. We believe that any framework of discourse prosody should include fillers and markers, account for their respective prosodic functions, and provide output specifications as well. Our prediction also stated two levels of filler prediction that allowed fillers and makers to exist within a PG. Systematic account and prediction of PF and DM in is significant step forward towards understanding the organization of discourse prosody. These findings could also be applied to speech synthesis and/or unlimited TTS for prosody enhancement.

6. References

[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, Chun-Hsiang Chang and Zhao-yu Su. "Investing F0 reset and range in relation to fluent speech prosody hierarchy". Technical Acoustics 2005, Vol.24, 279-285.*

[4] *http://www.MyET.com/COSPRO*

[5] *Daniel Marcu and Abdessamad Echihabi. 2002. An Unsupervised Approach to Recognizing Discourse Relations, Proceeding of the 40th Annul Meeting of the Association for Computational Linguistics (ACL), Philadelphia, 368-375.*

[6] *HuaJui Peng, Chiching Chen, Chiuyu Tseng, Kehjiann Chen ,"PREDICTING PROSODIC WORDS FROM LEXICAL WORDS--A FIRST STEP TOWARDS PREDICTING PROSODY FROM TEXT", International Symposium on Chinese Spoken Language Processing, ISCSLP, 2004.*