

Fluent Speech Prosody and Discourse Organization: Evidence of Top-down Governing and Implications to Speech Prosody

Chiu-yu Tseng
Institute of Linguistics
Academia Sinica, Taipei, Taiwan

cytling@sinica.edu.tw

http://www.ling.sinica.edu.tw/member/fulltime_08.html

Understanding and Simulating Discourse Prosody

- Discourse position, chunking/segmentation, boundary breaks and tones, F0 reset...etc.
- Lehiste, 1975 –English
- Fujisaki, 1980's –Japanese
- Tseng since 1999 --Mandarin Chinese
 - Language specific and need to be tested on other languages.
- Oliveira, 2006 --Suya

Acoustic Correlates Examined

- F0
- Duration
- Intensity of voice source
- Pauses and boundary breaks
- F0 range variation
- F0 reset

Mandarin Chinese Discourse Prosody

- 1. Higher Level Organization and Discourse Segments
 - Fluent continuous speech vs. discrete/isolated single phrases
 - Domains, units and boundaries--multiple-phrase fluent speech paragraph
 - Macro/Top-down vs. micro/bottom-up (why **spoken discourse**)
 - Association vs. isolation
- 2. Hierarchical framework of fluent speech prosody and implications.—Speech planning
 - e.g. production undershoot, perceptual overshoot

- Phonetic Investigation via corpus approach
- Sinica COSPRO and Toolkit
 - <http://www.myet.com/COSPRO>
 - Collection of speech data
 - Development of analysis platform
 - Toolkit development and prosody modeling

Sinica COSPRO (Mandarin Chinese Continuous *S*peech *P*rosody Corpus) (2/1)

The corpus-

11.99GB recorded speech

mostly read narratives (organization of spoken discourse)

80MB spontaneous speech

7.7GB annotated

- 1. Phonetically Balanced Speech Database (COSPRO 01, 2047.8MB, 18:38, 6 speakers),
- 2. Multiple Speaker Speech Corpus (COSPRO 02, 2141MB, 19:29; 100 speakers)
- 3. Intonation Balanced Speech Corpus (COSPRO 03, 3441MB, 31:10; 5 speakers),
- 4. Stress-pattern Balanced Speech Corpus (COSPRO 04, 244MB, 48m; 2 speakers),

Sinica COSPRO (2/2)

- **5. Lexically-balanced Speech Corpus (COSPRO 05, 568.3MB, 35:50; 2 speakers),**
- **6. Focus-balanced Prosody Group Speech Corpus (COSPRO 06, 2060MB, 7:30; 2 speakers),**
- **7. Multiple Text-type/Speaking-style Speech Corpus (COSPRO 07, 626.7MB, 1:32; 2 speakers), including continuous word salads**
- **8. Prosody-unit Balanced Speech Corpus (COSPRO 08, 1500MB, 2 speakers), and**
- **9. Comparable Spontaneous/Read Speech Corpus (COSPRO 09, 80MB, 42m; 2 speakers).**

Features of COSRPO

1. Speech data collection

- a. Fluent continuous speech vs. canonical (discrete) phrase intonation
- b. Reading of text to reflect speech planning
- c. Design considerations in text composition

2. Annotation design—perceptually based (gap between speech and transcription of speech)

How do we listen?

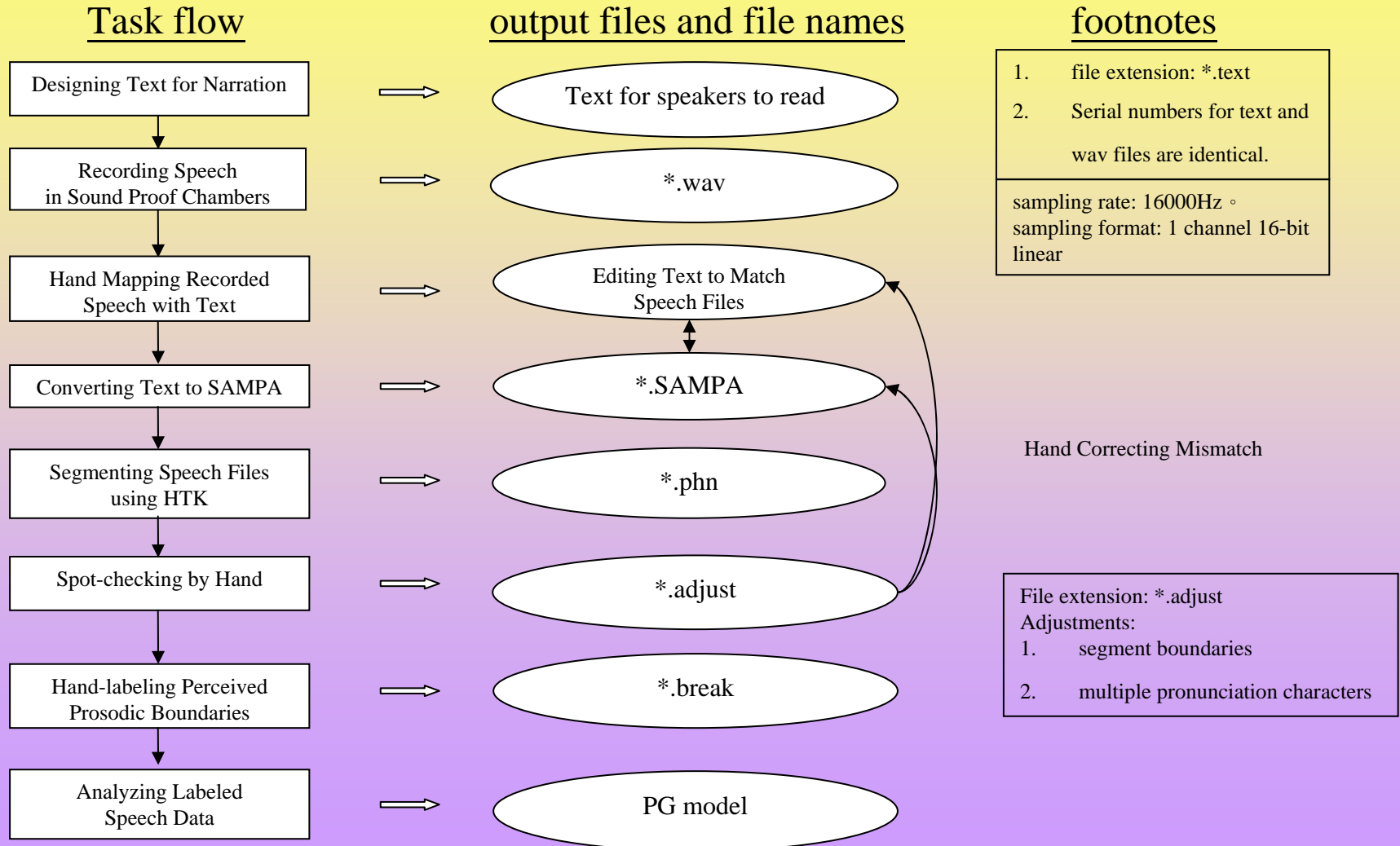
What do we listen to?

What do we hear?

Goals:

1. To construct the organization of fluent speech prosody from corpus analyses.
2. To account for the prosody of coherent multiple-phrase speech paragraphs in fluent speech as a discourse unit/segment.
3. To specify cross-phrase prosodic relationships and patterns systematically.
4. To derive cross-phrase prosody templates in relation to prosody organization.

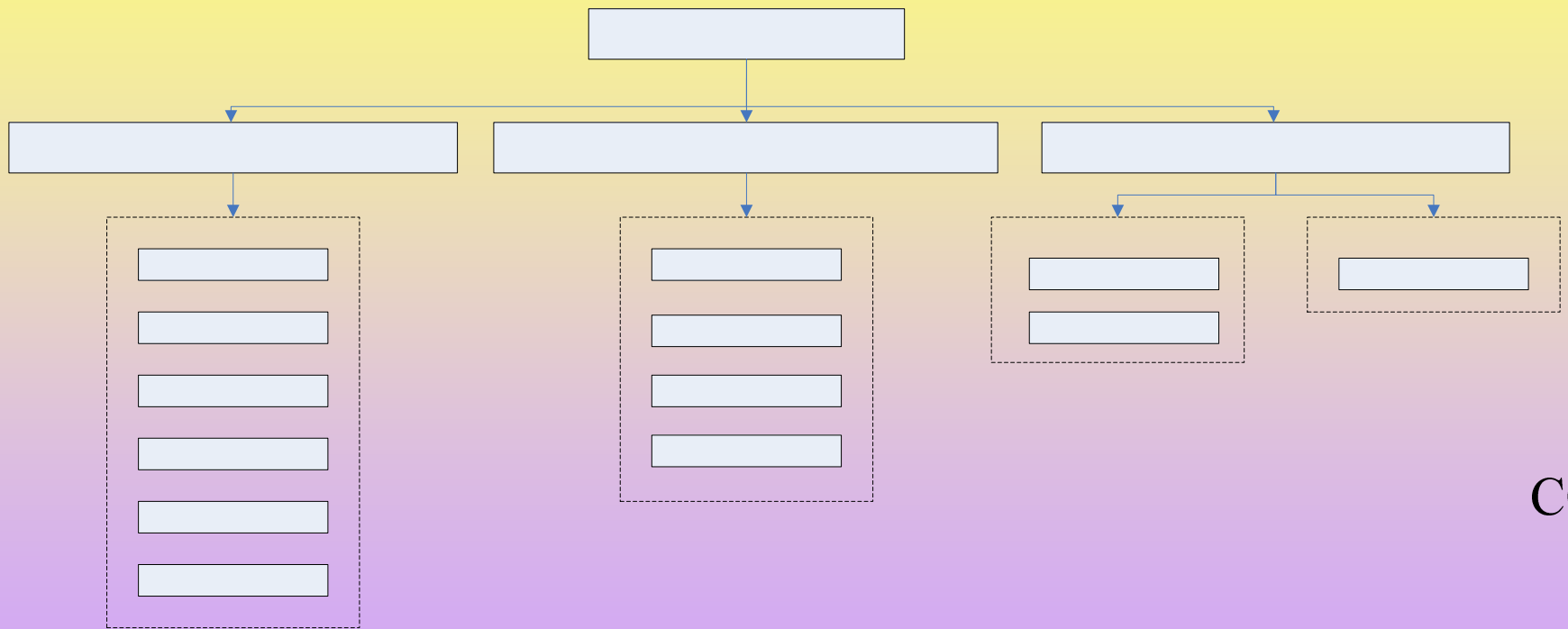
Flow chart of speech data processing and annotation - Read speech



Hand Labeling Perceived Boundary (Tseng et al, 1999)- Break Index (Transcription Consistency Maintained)

	Definition	Characteristics
B0	reduced syllabic boundary	Syllable truncation often occurred in fast or informal fluent speech.
B1	normal syllabic boundary/SYL	Usually with no identifiable pauses, but more of a psycholinguistic unit for native speakers.
B2	prosodic word boundary/PW	Perceived as a boundary where a slight tone of voice change usually follows.
B3	prosodic phrase boundary/PPh	A clearly perceived pause.
B4	breath group boundary/BG	Perceived end of exhale cycle followed by inhaling to begin another breathing cycle. It could be where a speech paragraph ends where trailing occurs with final lengthening coupled with weakening of speech sounds. But the speaker may still go on by breathing but not ending the speech paragraph.
B5	prosodic group boundary/PG	A complete speech paragraph ends by final lengthening coupled with weakening of speech sounds. The speaker makes a complete stop, take a new breath, and begin a new speech paragraph.

Sinica COSPRO Toolkit



Performing Acoustic Analysis Function

Labeling Continuous

COSPRO

Fluent Speech Prosody through Corpus Phonetic Investigations

1. Phrase or IU as paragraph unit.

Why is IU NOT enough to generate speech prosody?
How do effects from top-down governing affect IU?

2. Discrepancy between prosody and grammar

How PGs may or may not always correspond to boundaries of :

- Punctuations in text – language specific to Chinese
- syntactic boundaries – not language specific

3. Multiple-phrase phrase speech paragraphs as a discourse prosody unit.

Prosodic Phrase Grouping (PG) --The global melody and rhythm that constitute fluent speech prosody.

Example of Fluent Continuous Mandarin Chinese Speech



雖然，機械式思考為我們解開了不少零件與功能的疑結，但整體運作與成長演進等大問題卻益顯支離破碎。

而機械觀伴同來的理性自大，更為這個地球帶來許多慢性病症與不治之癌。

After annotation



雖然，〈B3/114ms〉機械式思考為我們解開了不少零件與功能的疑結，〈B4/151ms〉

但整體運作與成長演進等大問題〈B3/219ms〉卻〈B3/35ms〉益顯支離破碎。〈B5/232ms〉

而〈B3/25ms〉機械觀伴同來的理性自大，〈B3/291ms〉更為這個地球〈B3/53ms〉帶來許多慢性病症〈B3/34ms〉與不治之癌。〈B5/299ms〉

Note:

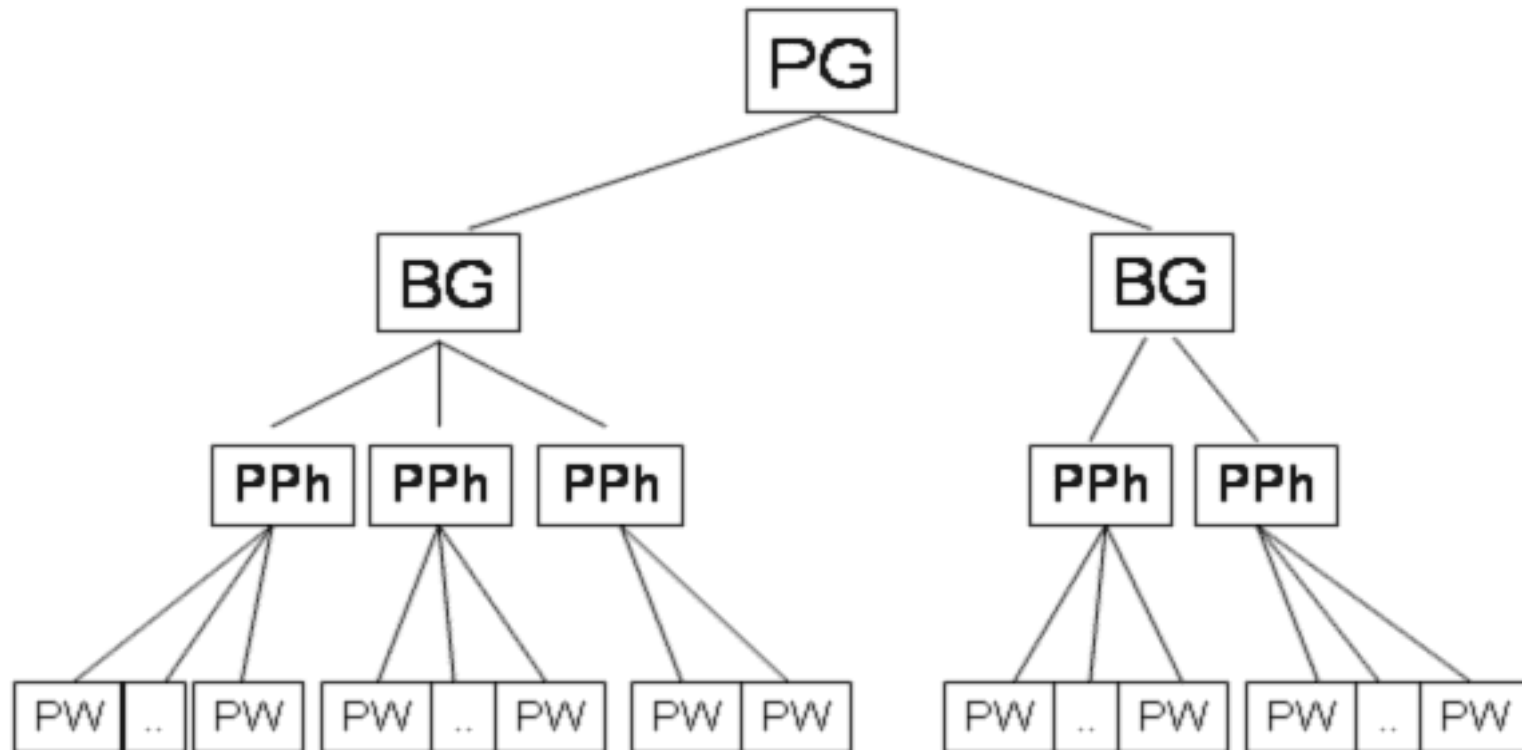
1. Where the boundaries do not correspond to any punctuation marks (B3).
2. How the boundary pauses differ in duration.

Hierarchy of Fluent Speech Prosody

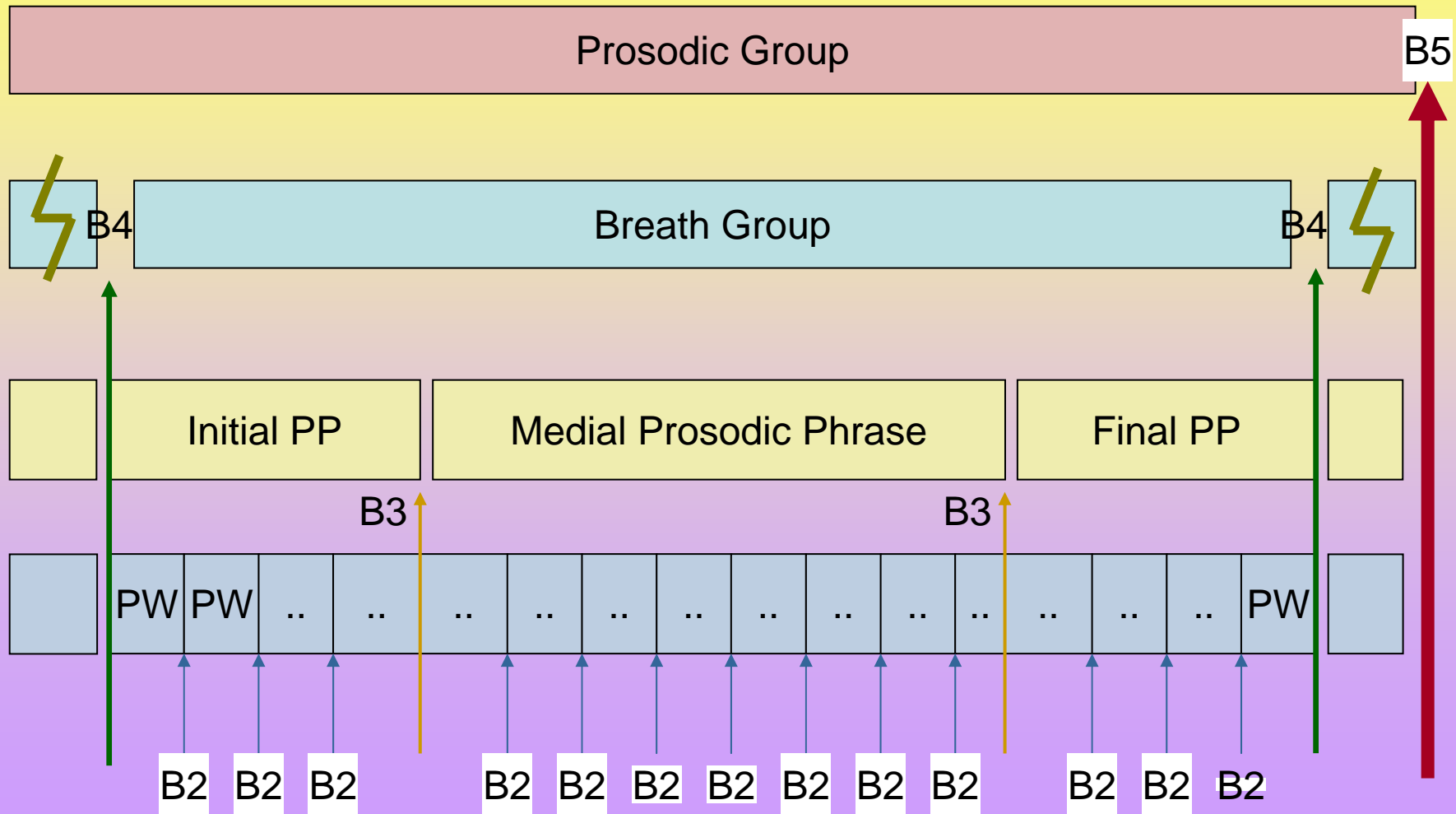
- Characteristics
 - Top-down government/constraints involved in continuous speech
 - Perceptual
 - Physiological (breath group Lieberman 1976)
 - Cognitive (Units of on-line planning and processing) and unintentional

Fluent Speech Prosody (Chao, 1968; ripples and tides; Fujisaki, 1980's)

Tseng's framework (2004, 2005)



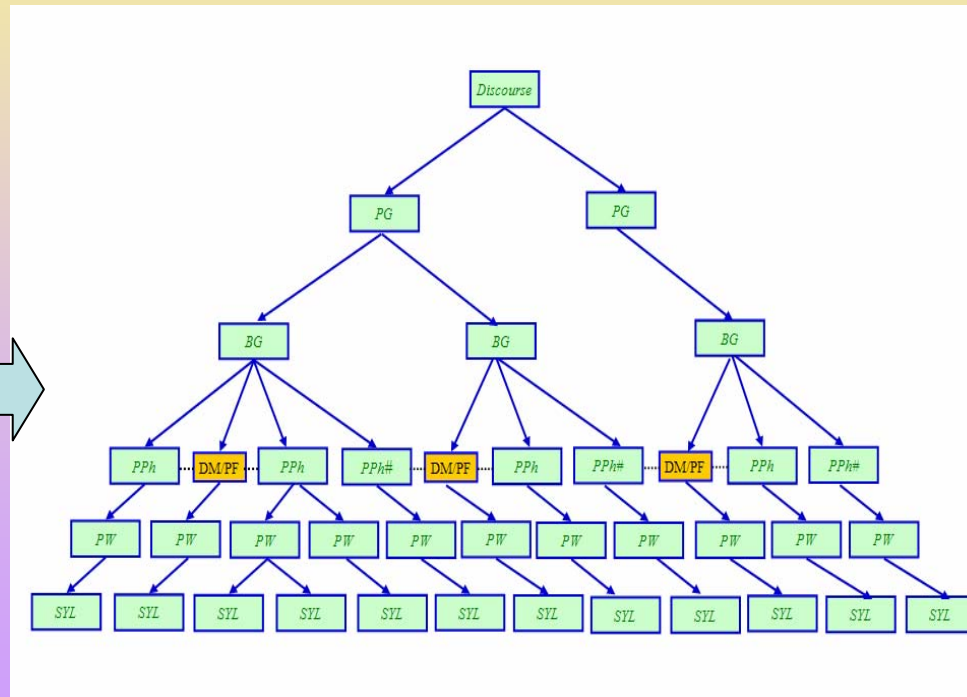
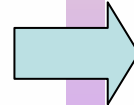
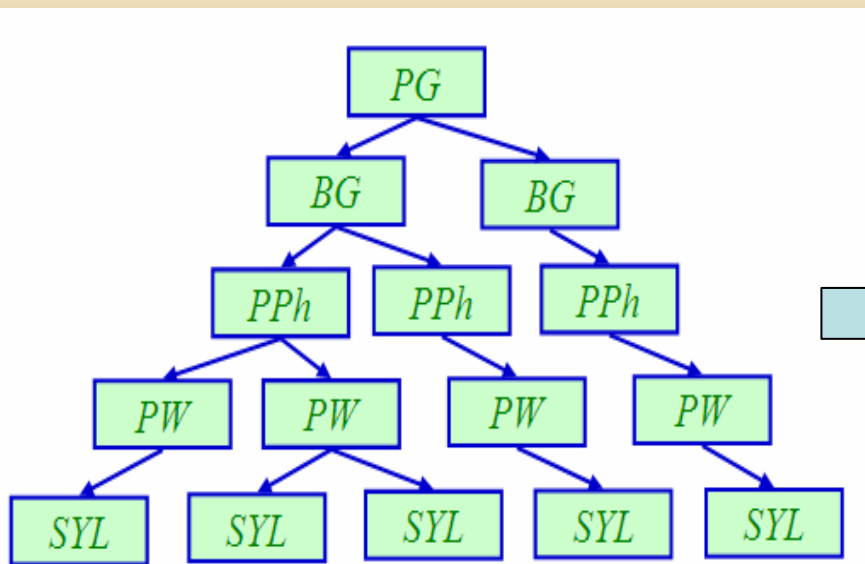
Prosodic Units and Boundaries in the Framework



From PG to Spoken Discourse

Tseng et al(2004a, b; 2005a)

How PPhs are IC of PG and PGs are IC of discourse. Both PPh and PG are therefore discourse units



Organization and framework (Tseng et al 2004a, 2005a)

1. Goes beyond IU (phrases and/or sentences)
2. Treats phrases not as unrelated discrete prosody units, but as sister constituents under PG.
3. Specifies how phrases modify respective intonations in PG.
4. Forms a hierarchy that governs phrases under with prosody units and their corresponding boundaries.
5. Accounts the overall output of multiple-phrase prosody that corresponds to speech paragraphs in spoken discourse.

What happens to phrase/sentence intonations or IU subject to higher level information?

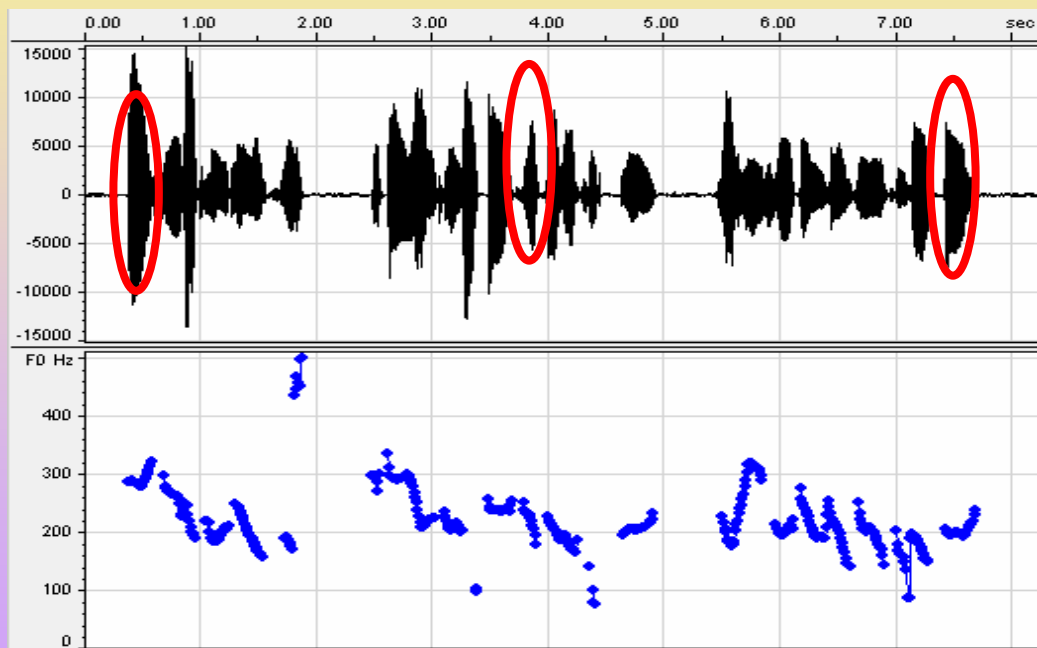
1. Assume respective positions in PG.
2. Make PG-specified modifications.

Basic Features of Fluent Speech Prosody

- How a speech paragraph is perceptually defined by position related features across phrases, i.e., how it begins, continues and ends.
 - PG related positions:
 - PG-Initial (beginning)
 - PG-Medial (continuation)
 - PG-final (termination)
- Unit:
PPh
- What corresponding patterns exist in relation to prosody hierarchy.
- How the prosody hierarchy contributes to prosody output.

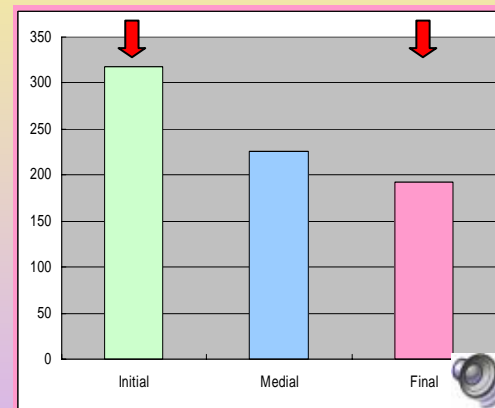
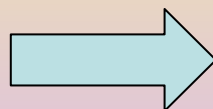
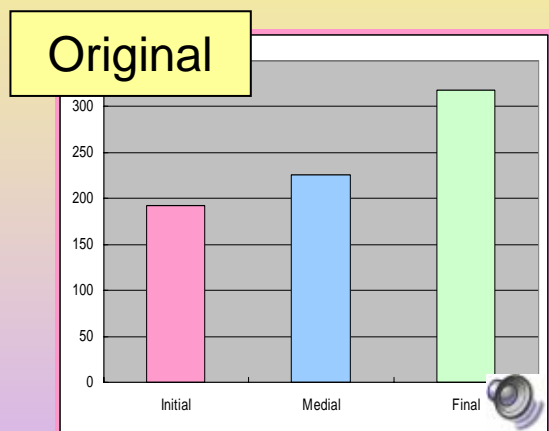
PG Positions in Relation to Prosody Organization—Higher Level Information

PG-I vs PG-M vs PG-F

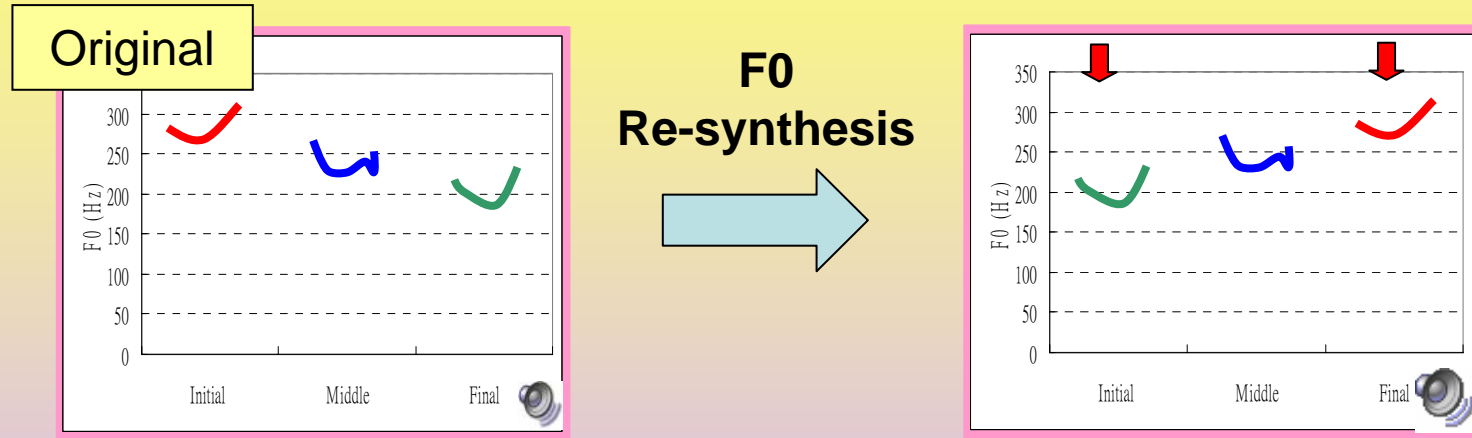


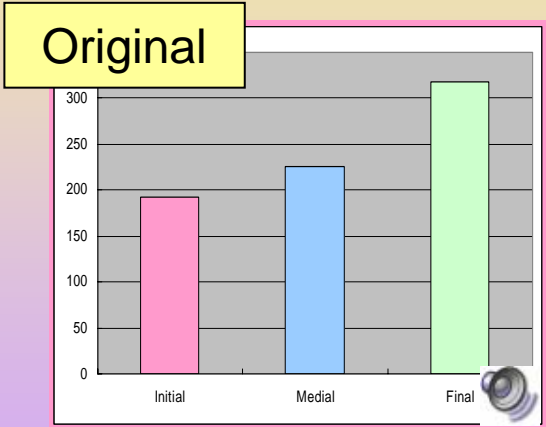
Carrier PG with embedded -ba1 : 巴是一個常見的字，一般人常把巴字掛在嘴邊，講話時動不動就會提到巴。

Switching embedded target syllable in PG-initial and PG-final phrases

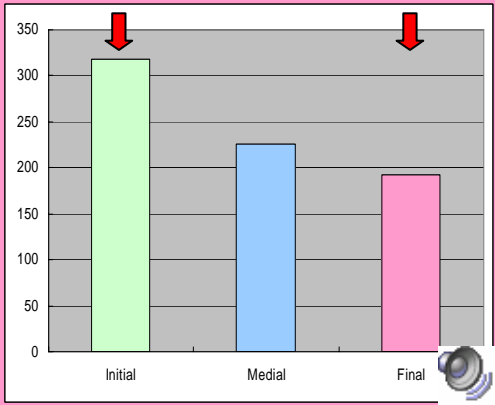


Switching F0 contours of PG-initial and PG-final target syllables





**Duration
Re-synthesis**



Higher Level Cadence and Templates

- Corpus investigations and quantitative analyses enabled us to derive cross-phrase hierarchical templates corresponding to every prosodic layer in the following 4 acoustic correlates.
 - 1. F0 **contour** templates
 - 2. Duration **cadence** templates
 - 3. Intensity **distribution** patterns
 - 4. Boundary **breaks** and **lengthening** templates
- Evidence of cumulative contributions (Tseng et al, 2004, 2005)
- Further studies on F0 range variation and reset.

Template 1. Syllable Duration Cadence of Fluent Speech — Cross-phrase Tempo and Rhythm (2/1)

1. Analyses of speech corpora reveal layered duration adjustment and overall temporal allocation patterns that accounts for higher level rhythm in fluent speech.
2. Interactions of syllable durations prosody units are found in overall temporal allocation patterns.
3. Duration adjustments of syllables (lengthening and shortening across time domain) are patterned in relation to prosody organization.

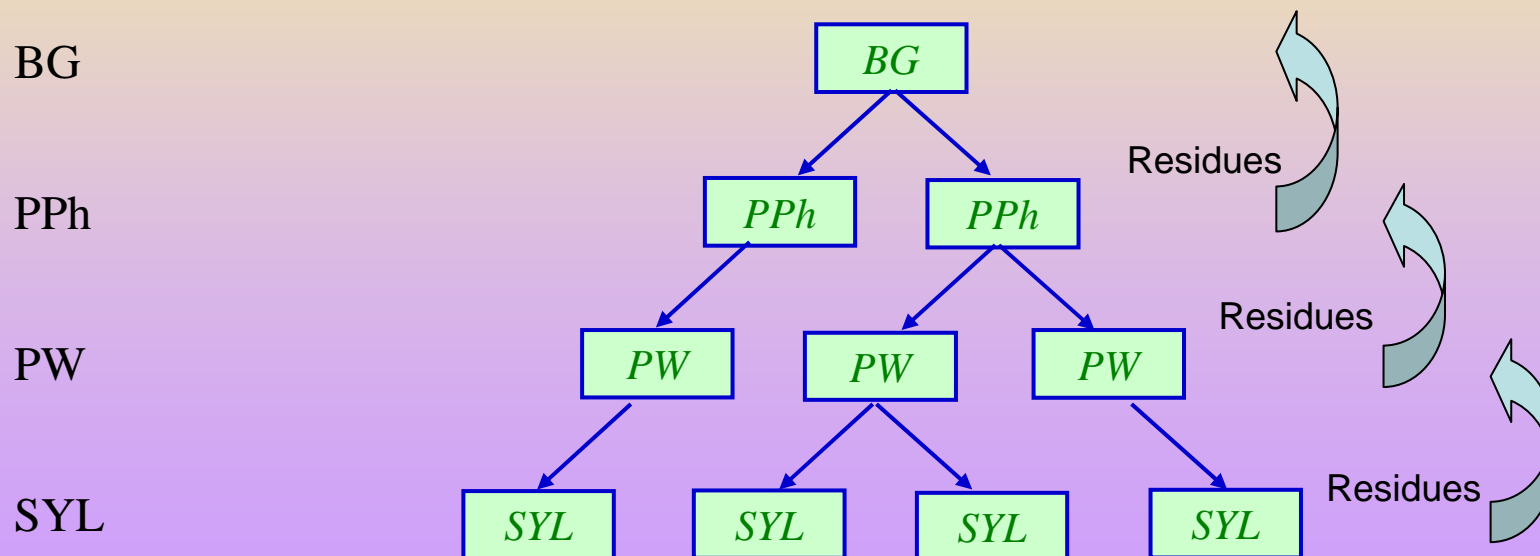
Template 2. Syllable Duration Cadence of Fluent Speech — Cross-phrase Tempo and Rhythm (2/2)

4. Evidence of corresponding governing relationship between PG and its subordinate layers were found.
5. Prediction can be made by number of syllables and status within a PG
6. Ultimate cross-phrase output **speech rhythm** of speech paragraph is the **cumulative results of layered contributions** from the prosody hierarchy.

Quantitative Analyses

Linear regression

- Using a step-wise regression technique **DataDesk™** from **Data Description, INC**, a linear model with four layers (Zellner, 1994) was modified and developed to predict speakers' timing behavior with respect to different speech rate.



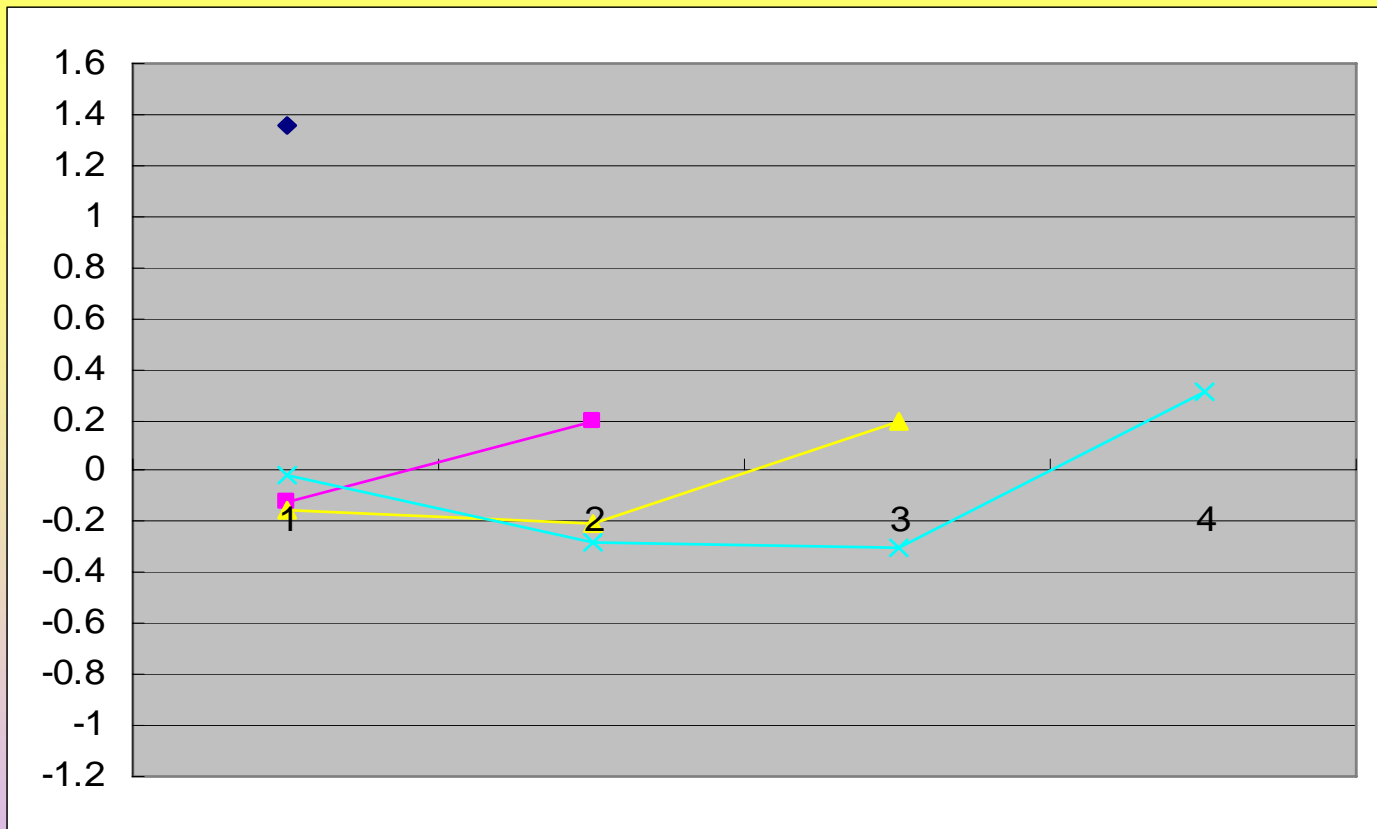


Figure1. Coefficients of M051 from the PW Model. The horizontal axis represents the position of each syllable within a PW; the vertical axis the coefficient values.

positive coefficients: lengthened syllable durations at the PW layer;
 negative coefficients: shortened syllable durations at the PW layer;
 Coefficients of p-value smaller than 0.1 were marked with the 'X' label

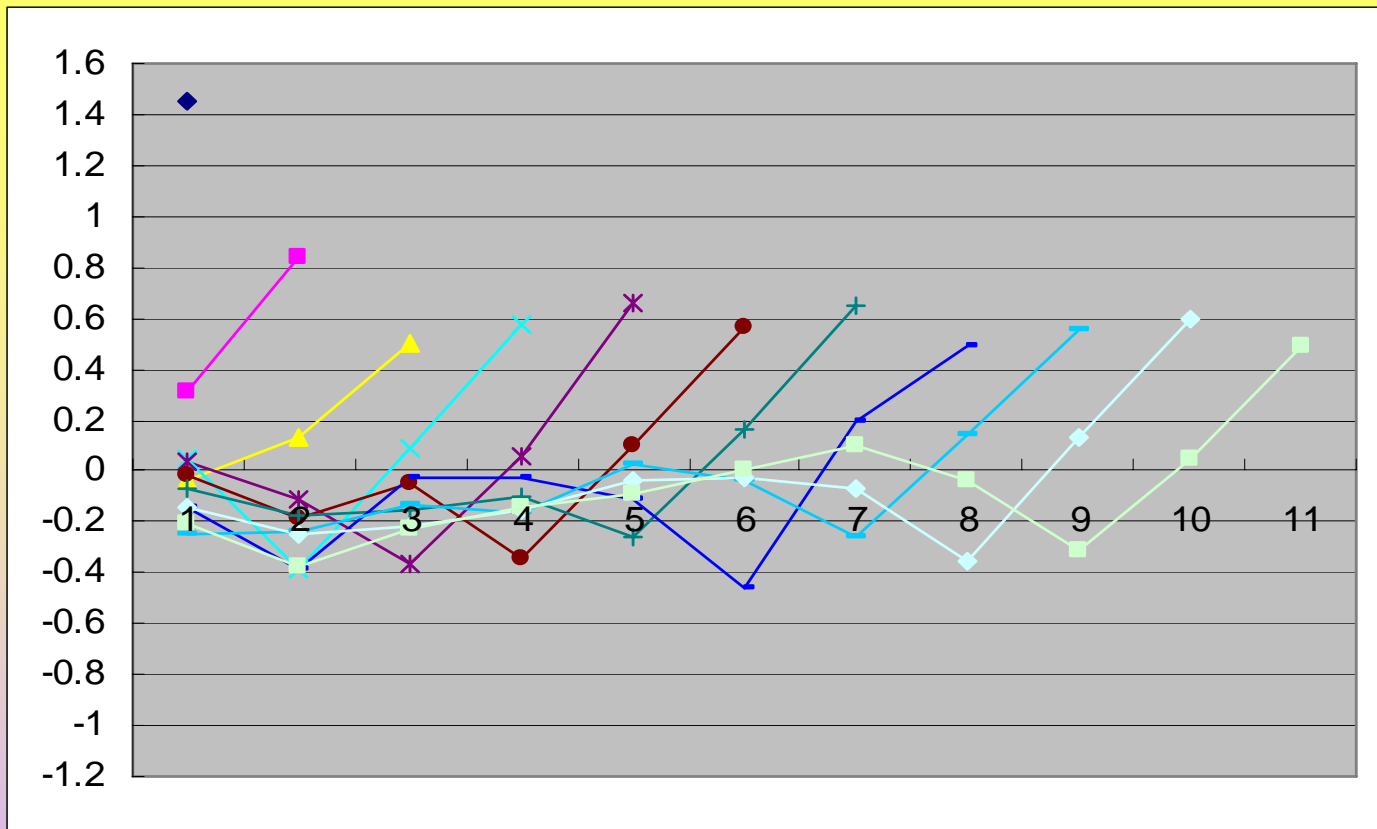


Figure3. Coefficients of M051 from the PPh Model

Findings at PW and PPh Layers

- 1. Boundary effects and final lengthening.
- 2. A clear cadence of speech **rhythm**. Final lengthening and backward shortening make up most of the **rhythmic patterns** of fluent speech.
- 3. Cadence was found across speaker, speaking rates and dialects.
- 4. Final syllable lengthening at the PPh layer was found to be twice as long for faster speech (FFS)
- 5. A **complementary** effect of final syllable lengthening was found between the PW Layer and the current PPh Layer. **Whenever the final syllable of a PW is lengthened, the same degree of final syllable lengthening could NOT be found at PPh level.**

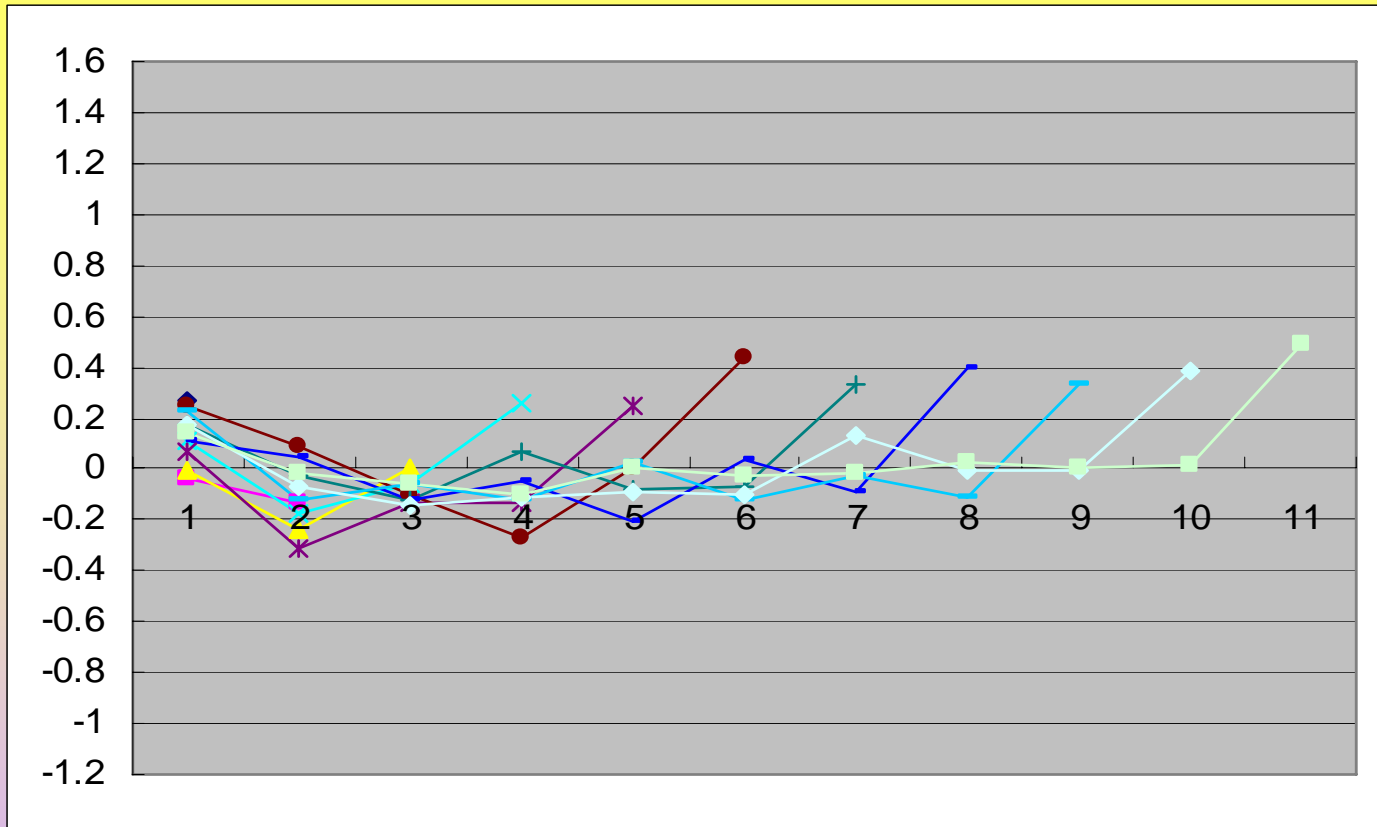


Figure 5. Coefficients of M051 from Initial PPh of BG layer Model
 i.e., effects of the BG layer on BG-initial PPhs.

Note:
 lengthening on the first and last syllable

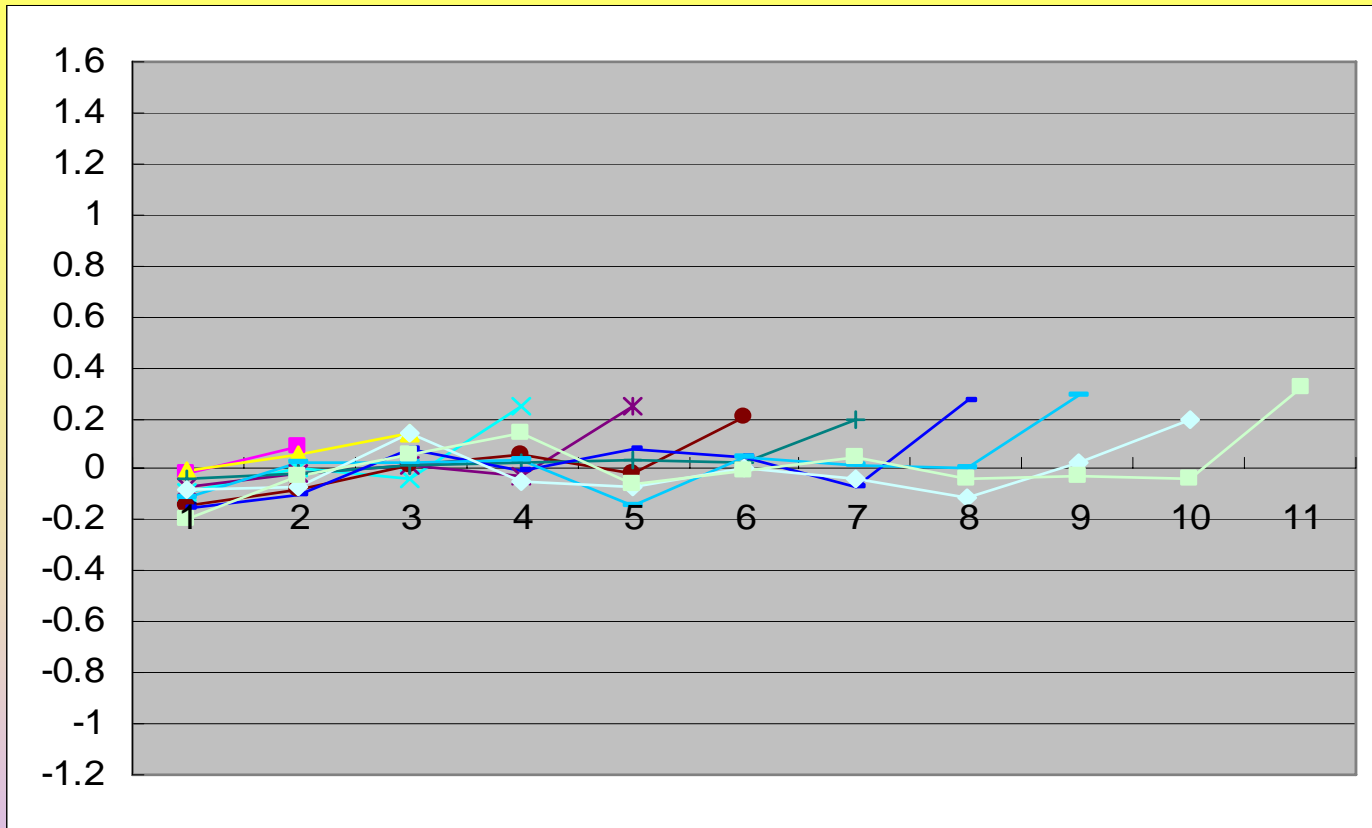


Figure7. Coefficients of M051 from Medial PPh of BG layer Model
i.e., effects of the BG layer on BG-medial PPhs

Note:

First syllable shortened; final syllable lengthened.

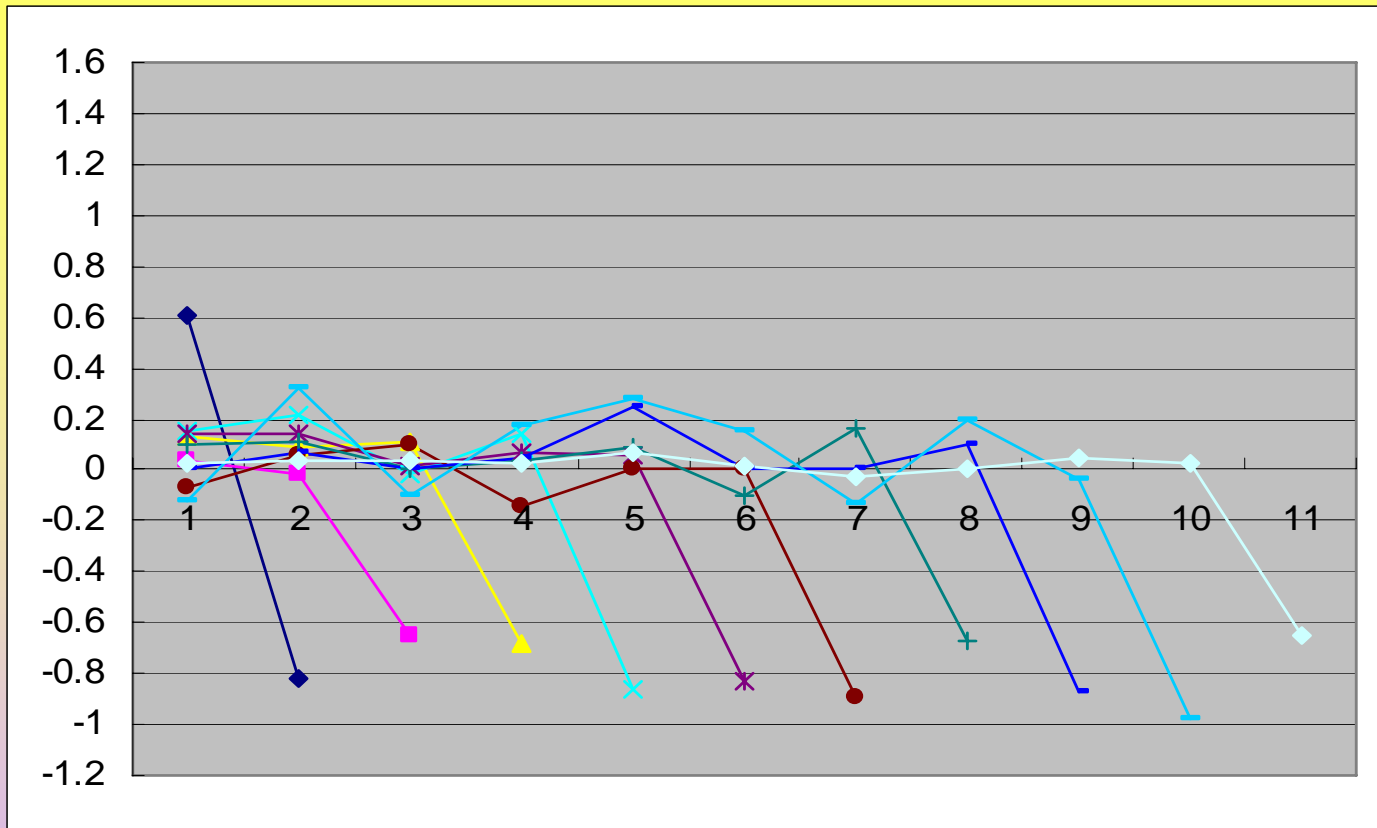


Figure9. Coefficients of M051 from Final PPh of BG layer Model
i.e., effects of the BG layer on BG-final PPhs

Note:

1. Final syllable is shortened.
2. Negative coefficients reflect a clear contrast between BG-initial and BG-final PPh.
3. Cumulative overall final-syllable lengthening still exists after trade-off .

Cumulative prediction of vs. original speech output

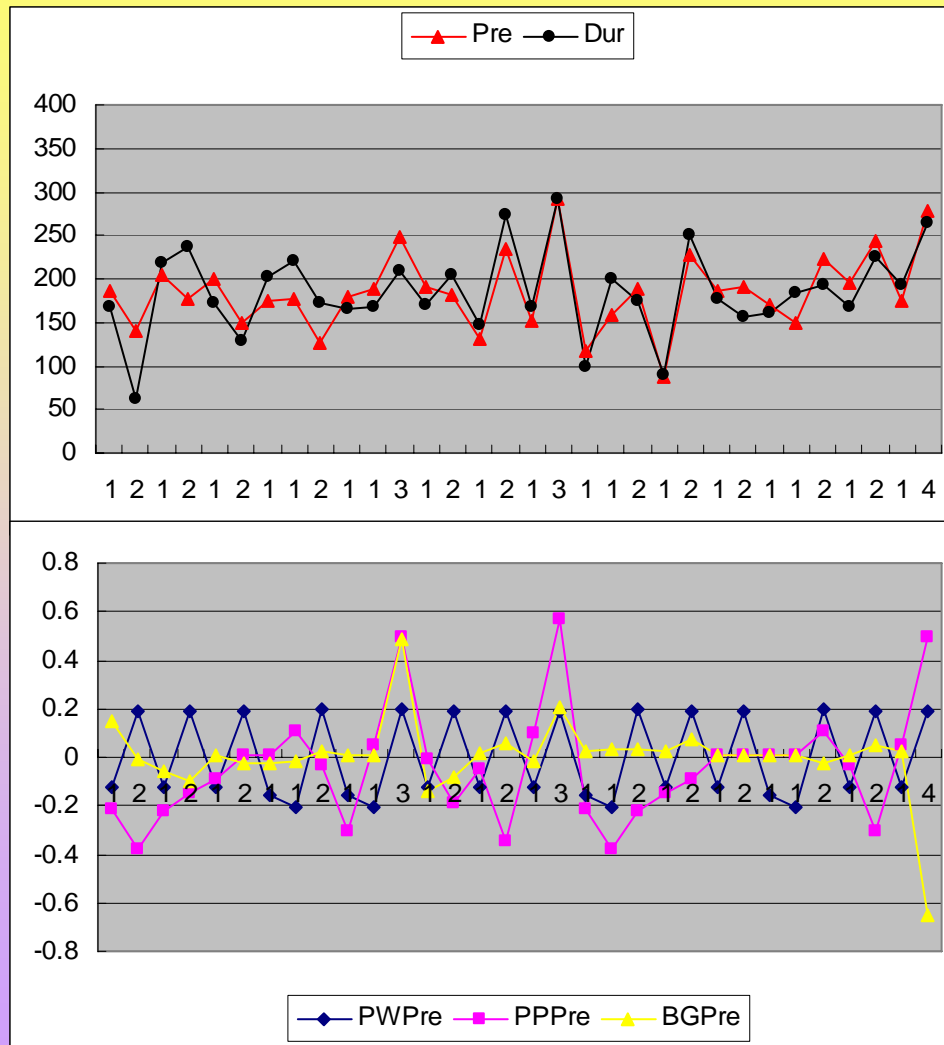
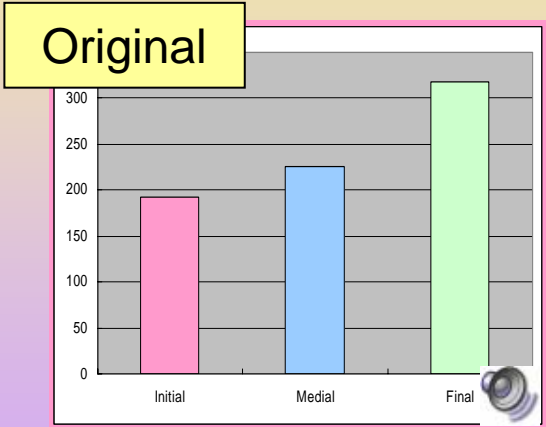
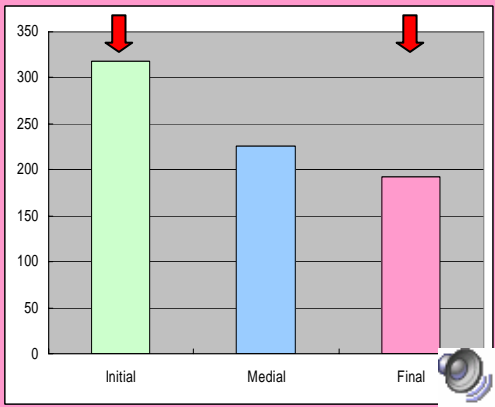


Figure 11. Comparison between speech data and predictions for M051



**Duration
Re-synthesis**



Significance: Higher Level Speech Rhythm

1. Higher level PG effects are found across syllables and phrases.
2. Every prosodic layer contributes to the final cross-phrase output. **Trade-off effects were found between prosody levels.**
3. PG positions are crucial prosody information.
4. Intrinsic segment/syllable durations can NOT account for speech rhythm in fluent speech.

Speech Rhythm: **Units** and Significance

- 1. A hierarchical organization does function during speech production; templates of **chunking and specifying shortening and/or lengthening (LS(S)(S)L)** during speech production.
- 2. Higher level cadence templates must also be used in on-line speech processing (perceptual overshoot, look-ahead and forecast).
- 3. *Lower level lexical tone information aside, temporal allocation patterns and timing structures are crucial to fluent speech. Syllable duration cadence patterns are just as significant as F0 contour patterns to prosody.*

Template 2. F0 cadence of Multiple-phrase PG— Perceived Fluent Speech Melody and Cumulative contributions to output F0 (Tseng et al, 2004)



Most significant implications: Higher Level Association

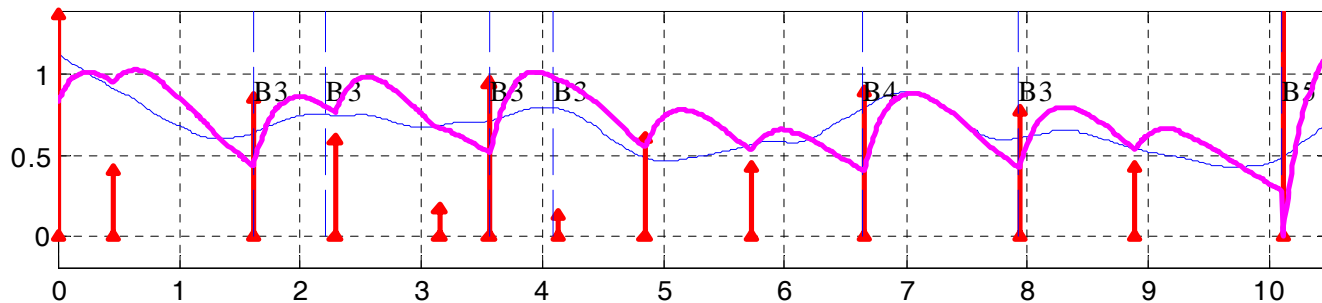
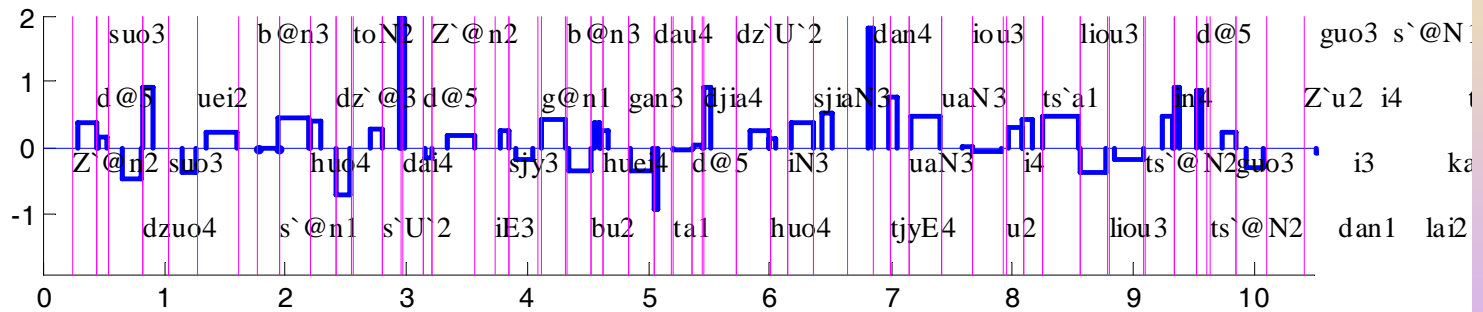
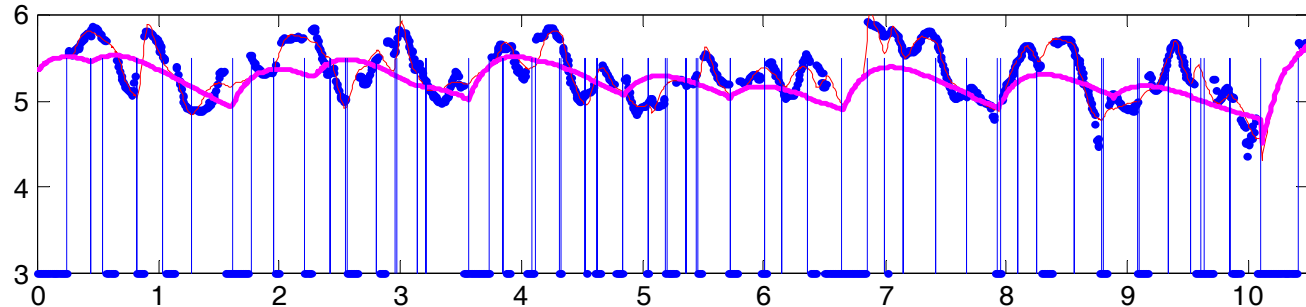
1. Phrase intonations are **NOT unrelated** prosody units, but rather, immediate constituents of paragraph and sister constituents that form speech paragraphs PG (Prosodic Group).
2. Phrase/sentences intonation and boundary breaks may not always correspond to syntactic boundaries in speech.
3. Prosody framework of fluent speech involves establishing and specifying **higher-level organization and cross-phrase** prosodic relationship, as well as reflecting planning units of narratives/spoken discourses.

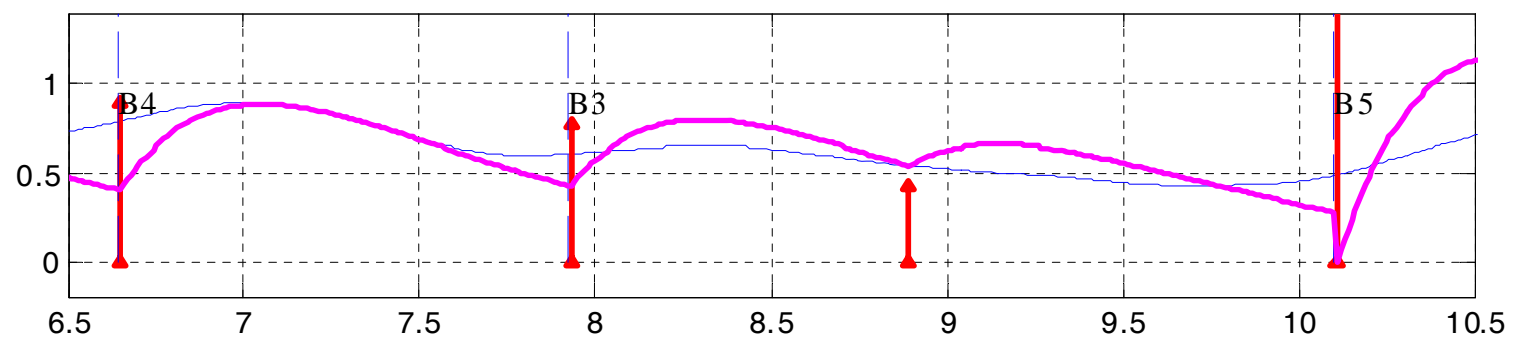
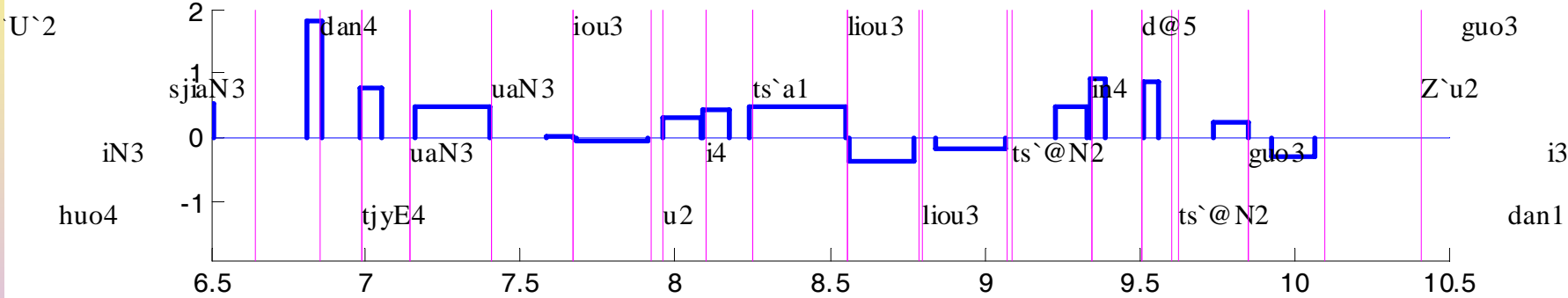
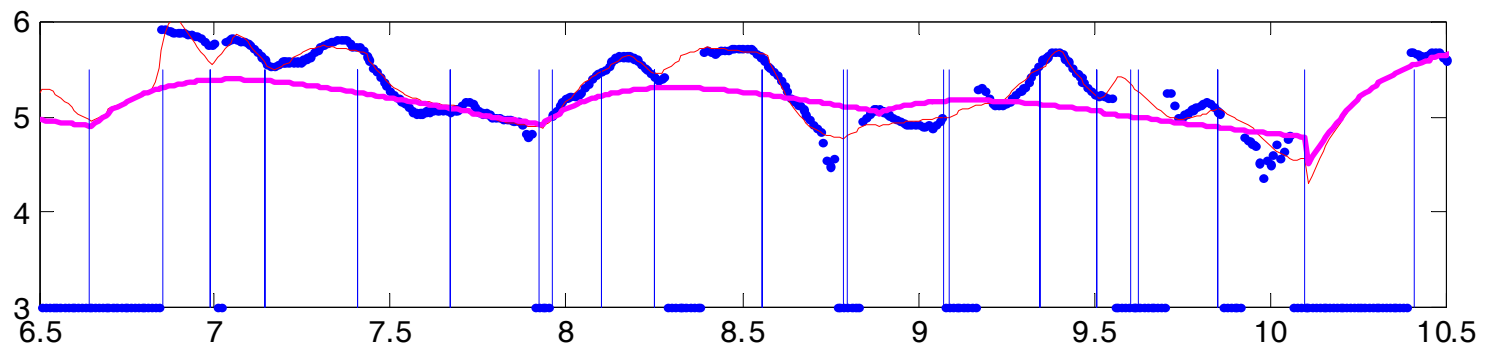
Phrasal Intonation Statistics

- Each phrase command is in effect by a linear combination model of several local factors:
 - pause
 - previous A_p effect
 - position in PPh (boundary depth)
 - level of the Fujisaki model parameter F_0 base

$$\text{Phrase command } A_p = \text{constant} + \text{coeff1} \times \text{pause} + \text{coeff2} \times \text{pre_phr} + \text{coeff3} \times f0_{\text{min}} + i_{\text{Syl_PPh}}$$

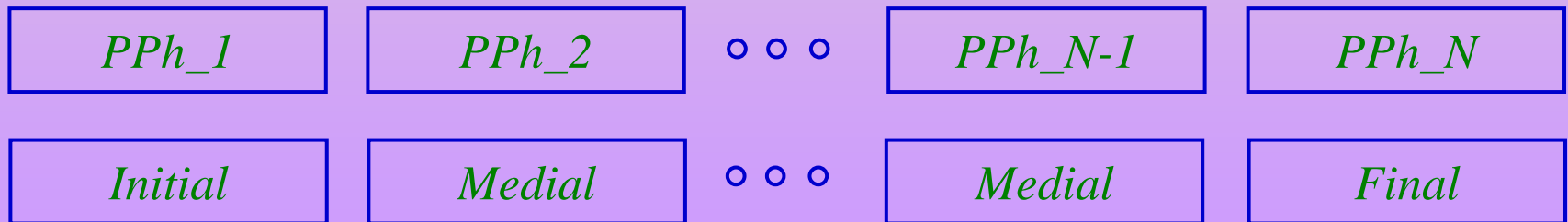
Automatic Extraction





P.G. Intonation Model

- Each phrase command is also in effect by a global dominator P.G.. in either of the following factors:
 - Index of prosodic phrase in P.G.
 - Initial, medial and final positions in P.G.



P.G. Intonation Model

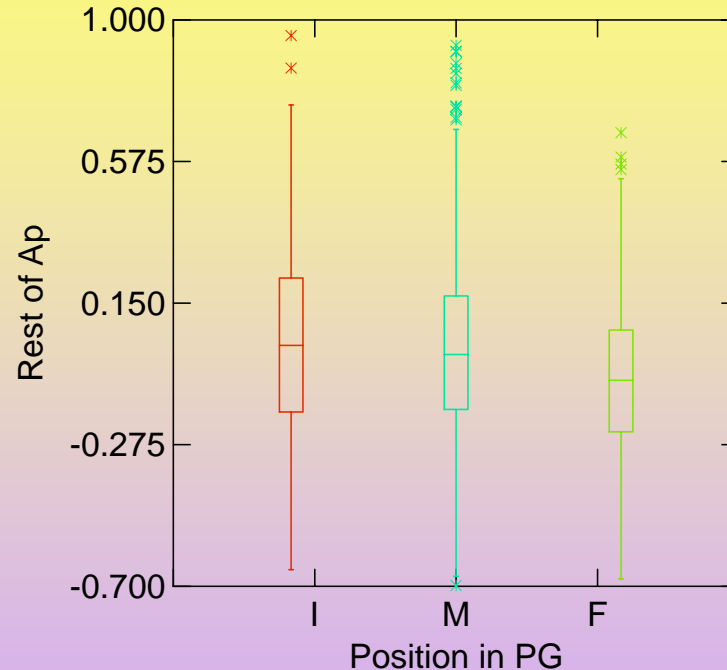
	Mean-Square	F-ratio	P
Model	0.035	1.068	0.159(Failed)
Error	0.033	-	-

ANOVA table for the test of significance in index of PPh

	Mean-Square	F-ratio	P
Model	0.572	12.127	0.001(Significanc
Error	0.047	-	e)

ANOVA table for the test of significance in initial, medial and final PPhs

Phrase Command Ap in Effect of P.G. Initial, Medial and Final



The rest of Ap is accounted for residuals predicted by local phrasal intonation model in previous section, and now it has significance in positions governed by P.G. initial, medial and final.

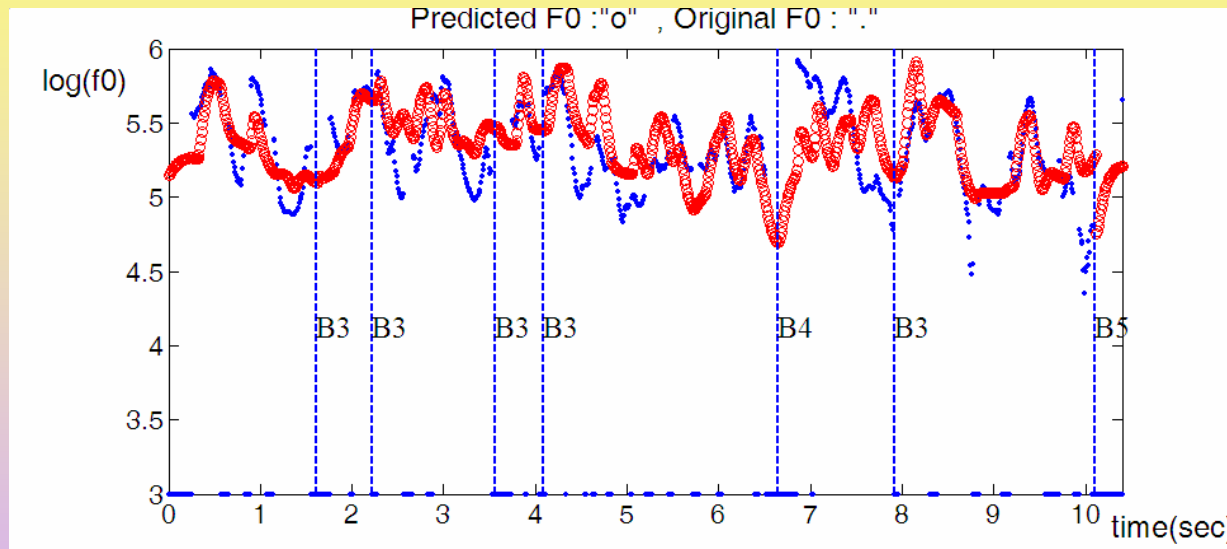
Overall Phrase Command Model

Modeling of overall phrase command:

Phrase command $A_p =$

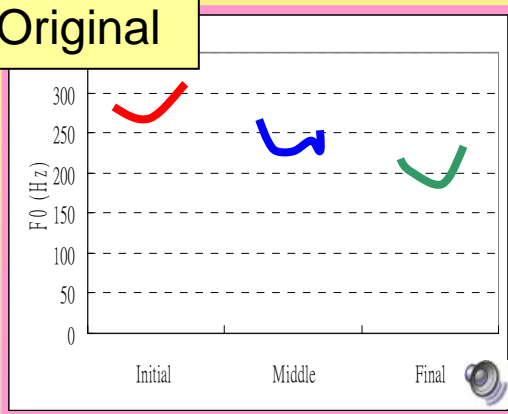
$$\begin{aligned} & \text{constant} + \text{coeff1} \times \text{pause} + \text{coeff2} \times \text{pre_phr} + \text{coeff3} \times \\ & f0_{\min} + iSyl_PPh \text{ (syl position in PPh)} \\ & + \text{PG effect coefficients} \end{aligned}$$

F_0 Prediction in P.G.

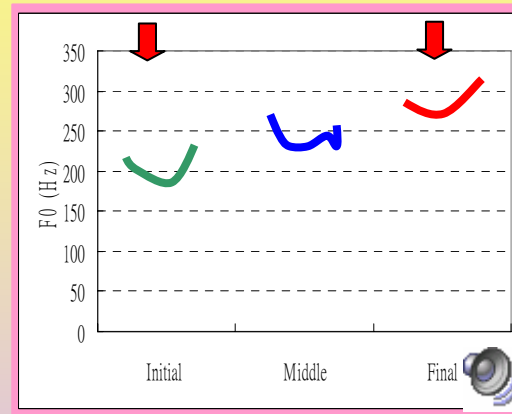


Simulation result of global intonation modeling of a PG. The red line represents simulated global contours; the blue represents contours of the original speech data.

Original



F0
Re-synthesis



Template 3. Cross-phrase Intensity Distribution Cadence Patterns

- 1. Units: PPh
 - Patterns are found only from the PPh layer up.
- 2. The longer the PPh is, the more energy required at the beginning.
- 3. Also relative to prosody organization.

Template 4. Cross-phrase Boundary Break Patterns

- Hierarchical
- Predictable
- At least 3 levels
- Relationships with speaking rate:
 - Faster speech:
 - more major breaks (B3, B4, B5) less minor breaks, longer duration and bigger differences among them.
 - Slower speech:
 - Less major breaks but more minor breaks, less duration difference (irregular B3s)

Fluent Continuous Speech One More Time



雖然，〈B3/114ms〉機械式思考為我們解開了不少零件與功能的疑結，〈B4/151ms〉

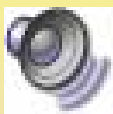
但整體運作與成長演進等大問題〈B3/219ms〉卻〈B3/35ms〉益顯支離破碎。〈B5/232ms〉

而〈B3/25ms〉機械觀伴同來的理性自大，〈B3/291ms〉更為這個地球〈B3/53ms〉帶來許多慢性病症〈B3/34ms〉與不治之癌。〈B5/299ms〉

Note:

1. Where the boundaries do not correspond to any punctuation marks (B3).
2. How the boundary pauses differ in duration (from 25 to 299 ms).

How Important Are Boundary Breaks?



What if we remove all breaks ?

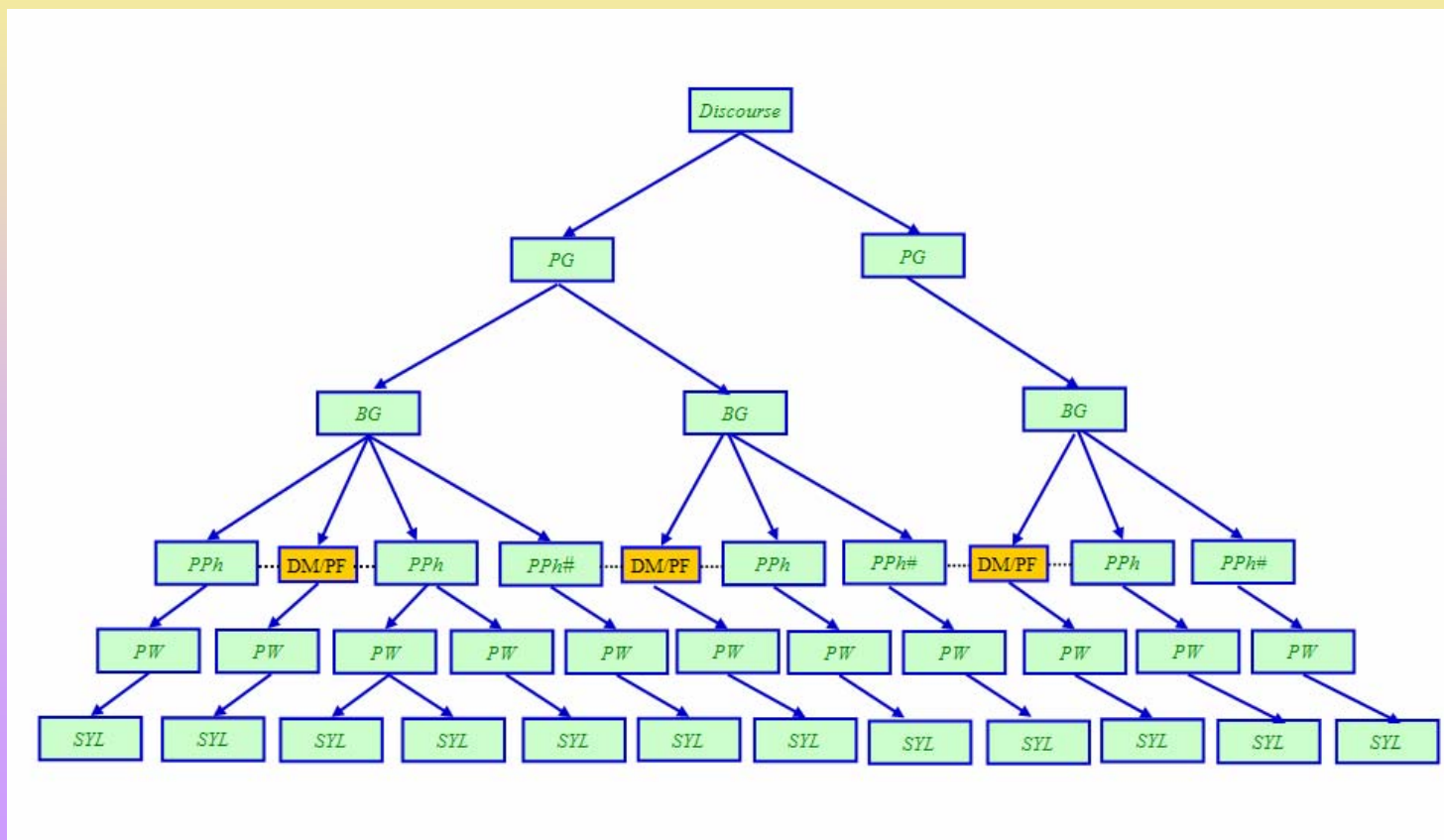


What if we swap the longest and shortest breaks ?

From PG to Spoken Discourse

Tseng et al(2004a, b; 2005a, 2006)

How PPhs are IM of PG and PGs are IM of discourse. Both PPh and PG are therefore discourse units



How Paragraphs Form Discourse? What Other Discourse Units Exist?

Materials (2/1)

- ✓ Sinica COSPRO(COSPRO_08)

- Speakers

One male (M051P) , One female (F051P)

- Text

26 discourse pieces (85- to 981-characters, 11592 characters)

- Speech data

Reading of text at normal speaking rate (200 ms/syllable)

(1.5-hr recorded speech, 170Mbytes after annotation)

➤ Annotations:

1. Segmental identities were first automatically labeled using the HTK toolkit and SAMPA-T notation (Tseng et al, 1999).
2. Perceived boundary breaks were hand tagged by trained transcribers using the Sinica COSPRO Toolkit ([url://www.myet.com/COSPRO](http://www.myet.com/COSPRO))
3. All labeling was also spot-checked by trained transcribers.

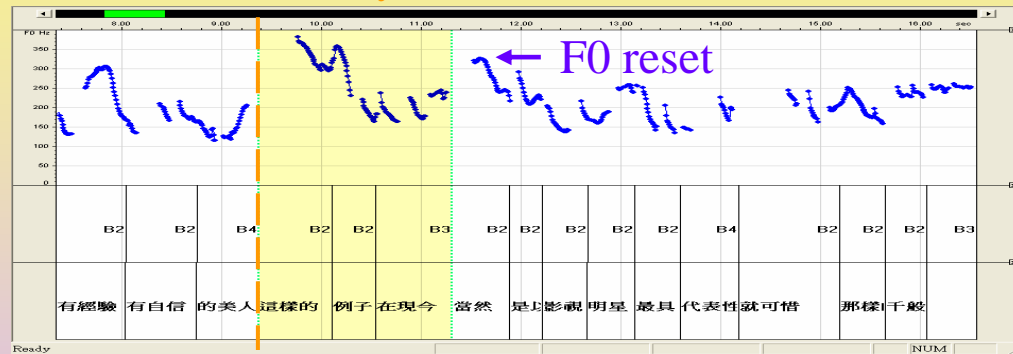
➤ Extracted features

1. maximum and minimum **value** of every syllable between two perceived boundaries.
2. F0 range within a prosody unit and F0 reset at a boundary are calculated.

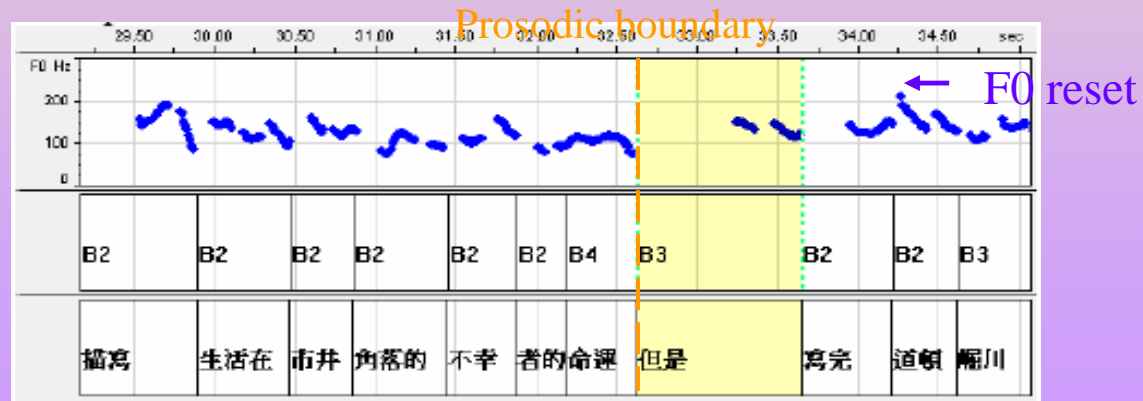
Between-PG or -BG F0 Range variation

Yellow block indicate F0 range variation of the connective phrase

Prosodic boundary



Relative wider F0 range



Relative narrower F0 range

Prosodic Fillers and Discourse Markers

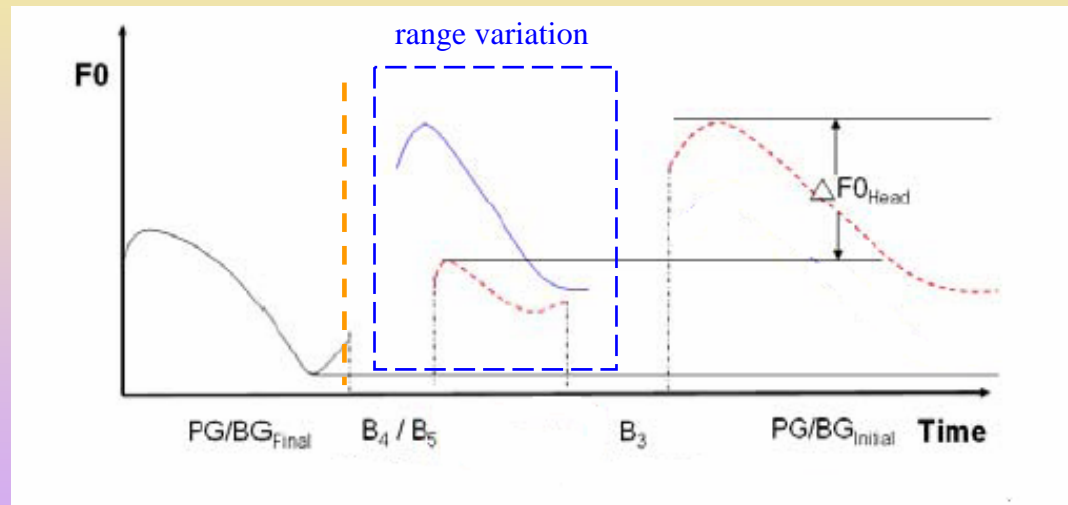
Prosodic filler (PF):

A phrase between two PGs or BGs with relative narrower F0 range

Discourse marker (DM):

A phrase between two PGs or BGs with relative wider F0 range

The connective phrase with F0

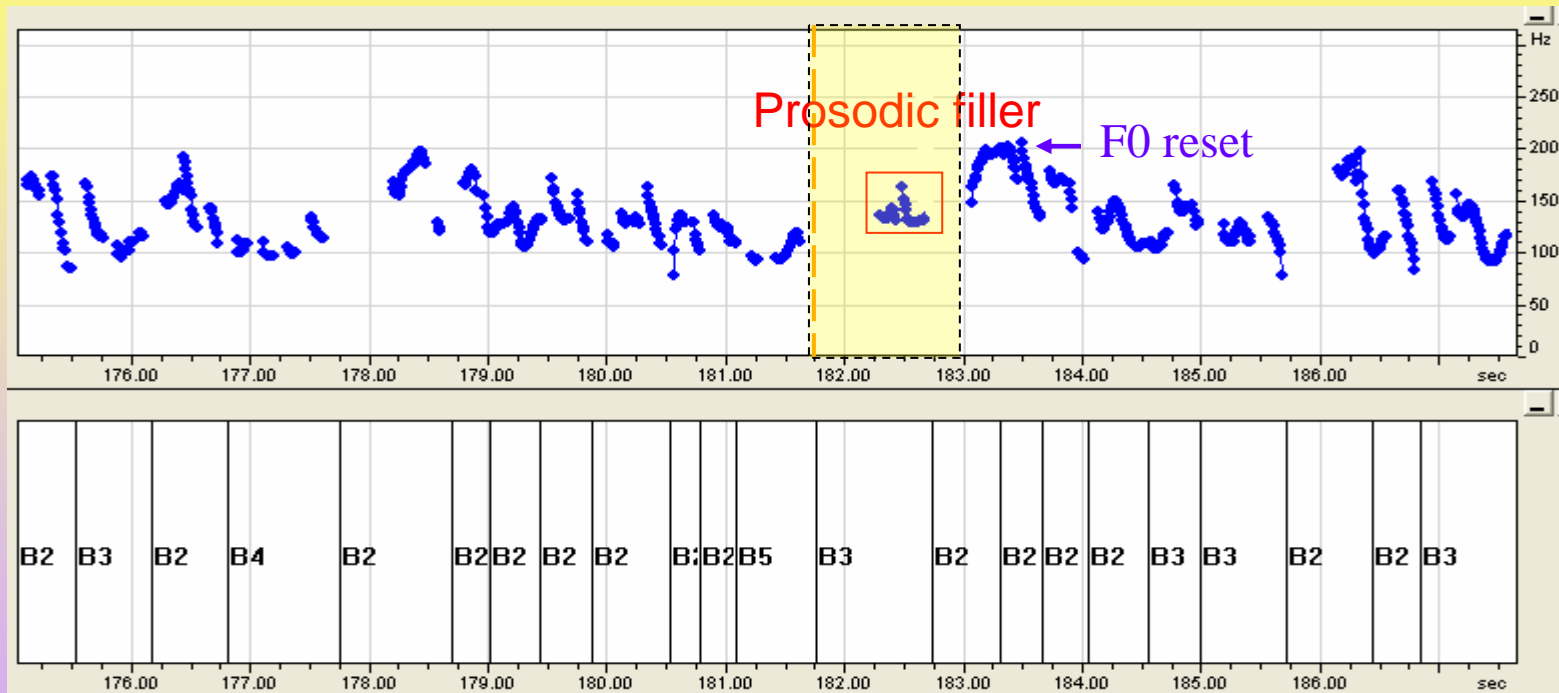


When $\Delta F0_{head}$ is smaller than -1.5, as illustrated in red, we define the phrase with narrowed range as a prosodic filler (PF).

When $\Delta F0_{head}$ is bigger than -0.3, as illustrated in blue, we define the phrase with widened range as a discourse marker (DM).

PF (Tseng et al, 2005)

Prosodic boundary

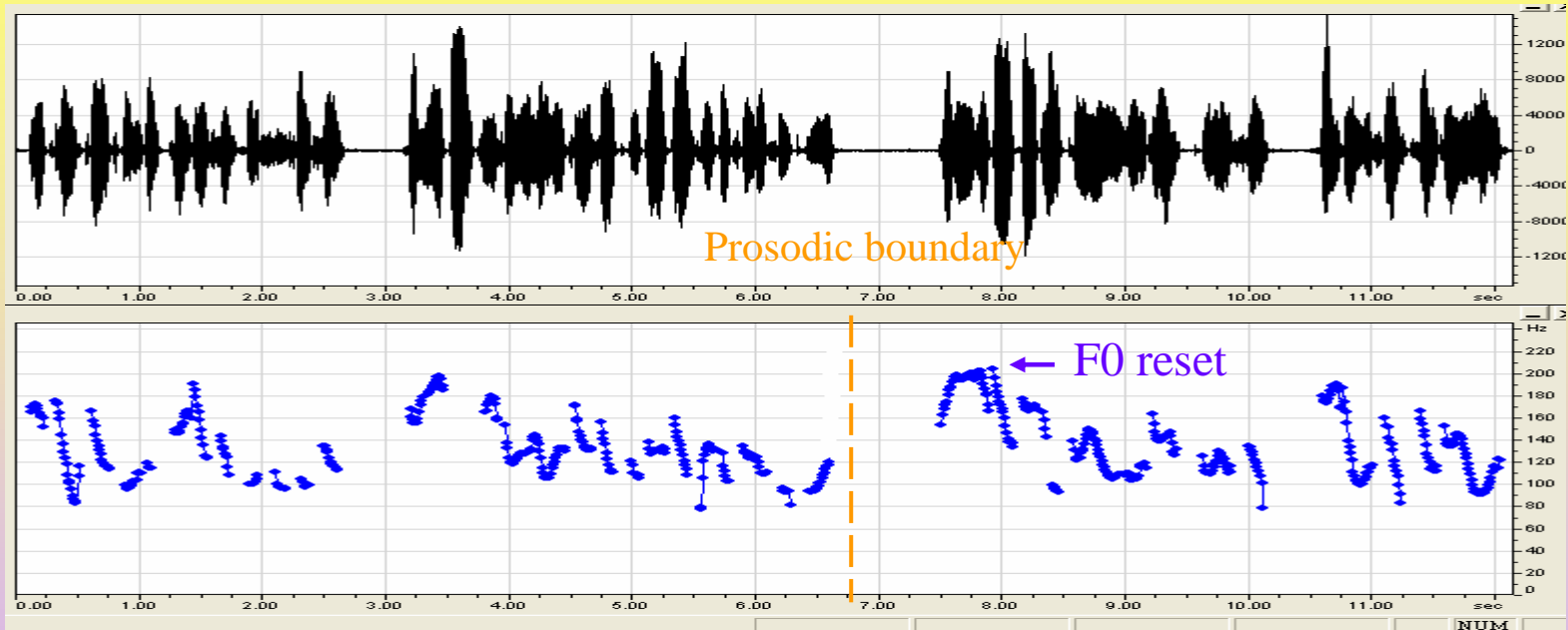


又具有<B2>較強的<B3>針對性<B2>和實效性，<B4>從而把<B2>精神
<B2>文明<B2>建設<B2>提高到<B2>一個<B2>新的<B2>水平。
<B5>今天today，<B3>人民<B2>日報<B2>發表<B2>評論員
<B2>文章，<B3>題目是：<B3>三點<B2>一線，<B2>看中原<B3>



Original


Preliminary experiment (2005)



又具有較強的針對性和實效性，從而把精神文明建設提高到一個新的水平。
人民日報發表評論員文章，題目是：三點一線，看中原，

 The waveform of the prosodic filler “今天today” was removed.

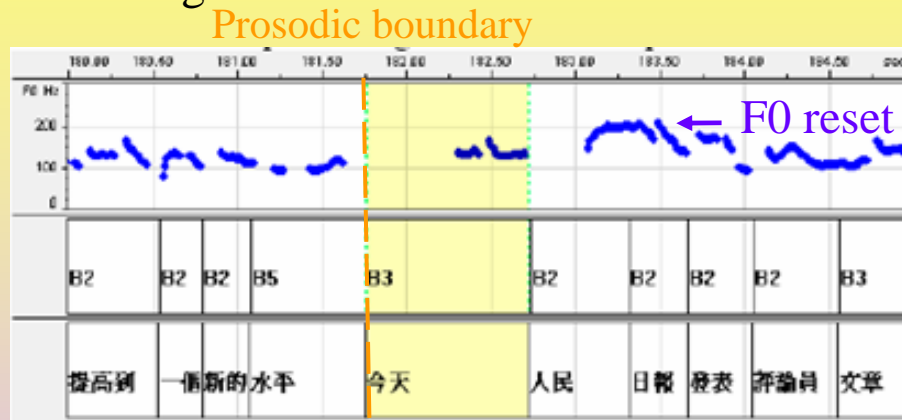
 The waveform of the prosodic filler “今天today” and following boundary break was removed.

 The waveform of prosodic filler “今天today”, preceding and following boundary breaks was removed.

PF and corresponding text analysis

PFs between PGs were demonstrated relative narrower F0 range or smaller F0 reset.

1. Relative narrower F0 range



Corresponding text analysis

When a post-B4 or B5 PW or PPh is 2 or 3 syllables such as “今天today”, it is usually produced as a filler.

2. Relative smaller F0 reset



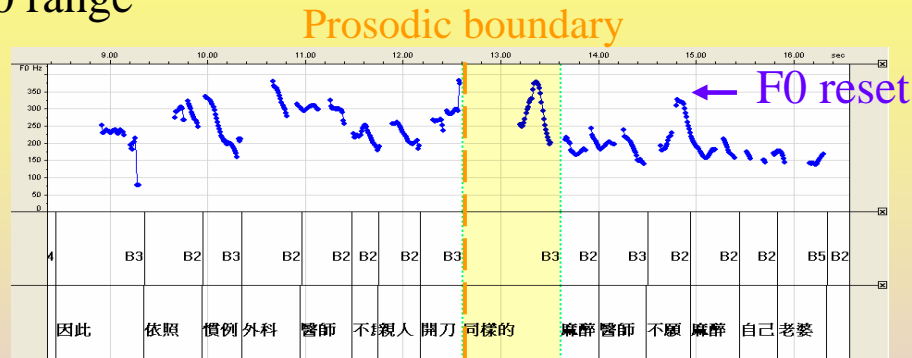
Corresponding text analysis:

Prepositional phrase

DM and corresponding text analysis

DMs between PGs were demonstrated relative wider F0 range or bigger F0 reset.

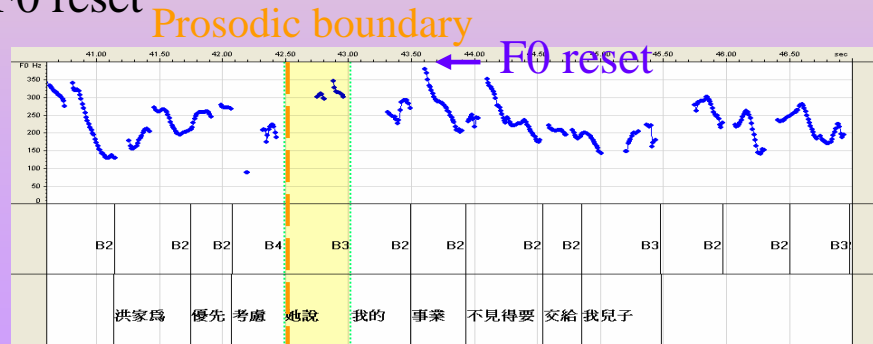
1. Relative wider F0 range



Corresponding text analysis:

A transitional word such as “同樣的in the same way”, “然而however”, “但是 but”.

2. Relative bigger F0 reset



Corresponding text analysis:

Lexical items meaning “她說she says”, “指出point out”, or “表示indicate”.

F0 Range of speech data in relation to punctuations in text

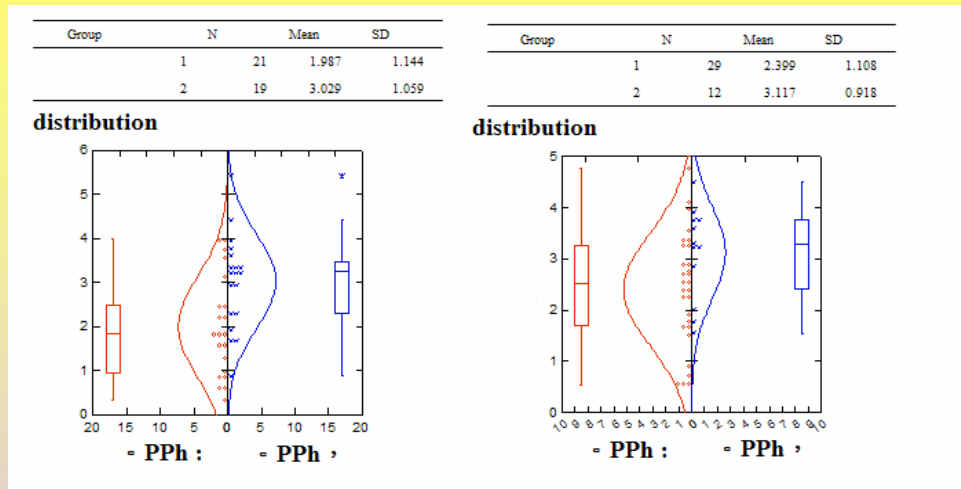


Figure 11: Comparison of punctuations in text for speakers (a)M051 and (b)F051

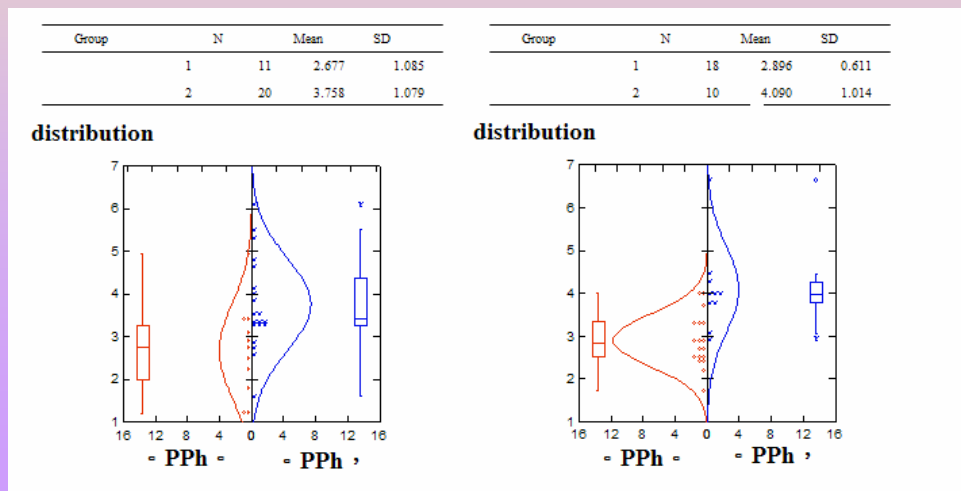


Figure 12: Comparison of punctuations in text for speakers (a)M051 and (b)F051.

The blue line indicates the distribution of relative short portion of speech data with wider F0 range where in text it is an actual initial of a new paragraph right after a period and followed by the comma.

The red line indicates the distribution of relative short portion of speech data with narrower F0 range where in text it would be a transitional word or phrase right after a period and followed by the colon.

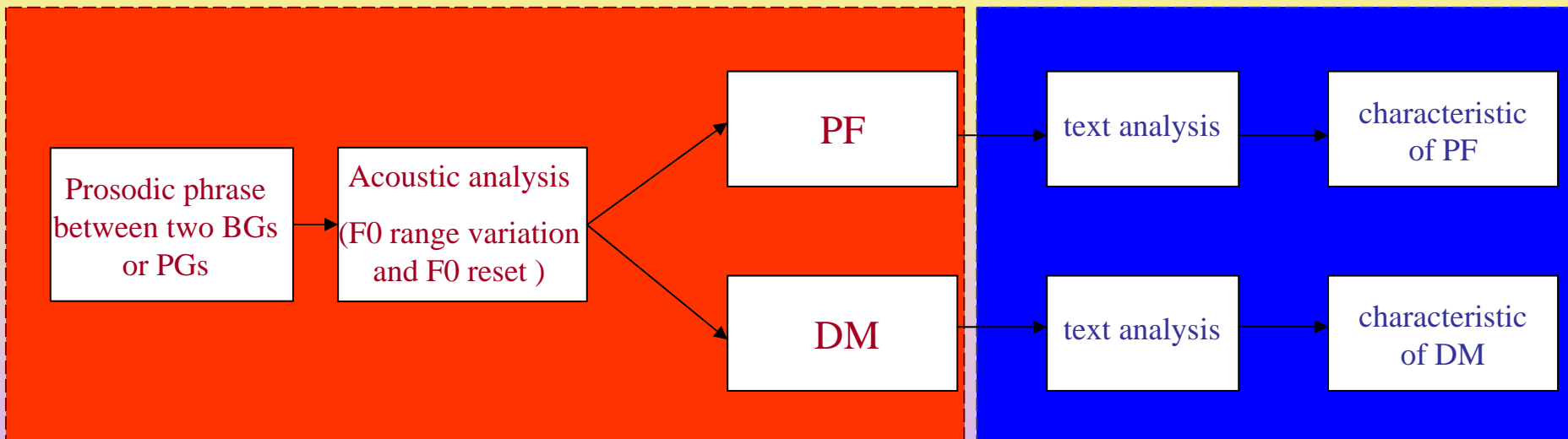
The blue line indicates the distribution of relative short portion of speech data with wider F0 range where in text it is an actual initial of a new paragraph right after a period and followed by the comma.

The red line indicates the distribution of relative short portion of speech data with narrower F0 range where in text it would be a short portion right after a period and followed by another period. (In other words, a transition between two long paragraphs in text).

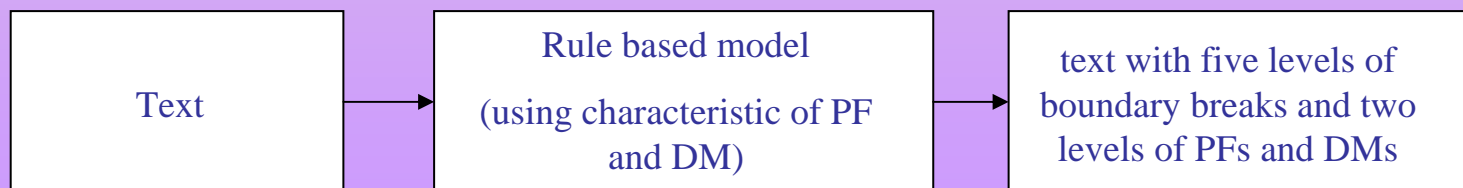
Building a model to predict PFs and DMs from text

Acoustic analysis

Analysis of corresponding text



Building a prediction model



Rule based model to predict PFs and DMs

We use phrase containing lexical construction and punctuations
to built this text model

ProduceDiscourseMarker/Fillers

- (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 discourse markers and
boundary breaks

Figure 7: Procedures to produce two levels of fillers

The end result after all steps are applied is text with five levels of boundary breaks and two levels of PFs and DMs(1.First level is between two PGs
2.Second level is between two PPhs)

Text prediction vs. Speech data

Comparing the result of text analysis and the result of speech data

Definition:

$$\text{Precision} = \frac{\text{numbers of correctly predicted PF and DM}}{\text{numbers of predicted PF and DM}}$$

$$\text{Recall} = \frac{\text{numbers of correctly predicted PF and DM}}{\text{numbers of actual PF and DM in the speech data}}$$

$$F - score = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Text prediction vs. Speech data

Table 2. Baseline without fillers

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

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

Table 3. Baselines with fillers

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

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

Table 4. Consistency of two speakers

Recall	Precision	F-Score
0.577	0.535	0.555

Table 4 shows comparison of cross-speaker consistency using speaker M051 as correct answer. Results showed that both speakers have their own interpretation of the same text.

Summary of Studying F0 Range Variation and Resets

- **1. Initial investigation (2005) of narrowed F0 range across speech flow showed that PFs occurred in fluent speech.**
- **2. Further analysis of between PG phrases showed that distinctions should be made between PFs that are semantically redundant and DMs that are attention callers.**
- **3. Both are discourse units that are prefixes of paragraphs that connects discourse segments PPh.**
- **4. Predicting PFs and DMs from text is possible.**
- **5. Any discourse prosody framework should include fillers and markers and their respective prosodic functions.**
- **6. These findings could be applied to speech synthesis and/or unlimited TTS to enhance prosody output.**

Conclusions

- Discourse Prosody and Top-down Information
 - 1. Higher organization must be accounted for.
 - 2. Hierarchical framework is necessary.
 - 3. Prosodic units must accommodate and correspond to higher nodes and higher information.
 - PPh Intonation (or IU) is subordinate and subjacent PG unit.
 - PG is subordinate and subjacent discourse unit.
 - 4. Cumulative contribution must be specified and explained.
 - 5. Generalization of discourse prosody is thus systematic.

Modeling Fluent Speech Prosody

- A modular acoustic model was constructed :
 - F0 contours
 - Duration patterns
 - Intensity distribution
 - Boundary lengthening and breaks

Modeling F0 contours on the Fujisaki model

➤ phrase command A_p

$A_p = \text{constant} + \text{coeff1} \times \text{Pause length before phrase command} + \text{coeff2} \times \text{Accumulated previous phrase command response} + \text{coeff3} \times \text{F0min in the Fujisaki model} + f(\text{Phrase command position in PPh})$

$$AccF_0 = \sum_{prev\ A_p} A_p \cdot \alpha^2 \cdot (t - T_{0i}) \cdot e^{(-\alpha(t - T_0))}$$

adjustment

F0 adjustment

Fujisaki model's phrase commands

$$G_p(t) = \begin{cases} = \alpha^2 t \cdot \exp(-\alpha t), & \text{for } t \geq 0 \\ = 0, & \text{for } t < 0 \end{cases}$$

$$\Delta A_p = \hat{A}_p - A_p = (P_c - P_p) \times \exp \times \alpha^{-1}$$

Modeling duration patterns

➤ 階層性韻律結構效應值 (Hierarchical prosodic effects)

DurS (ms) = Syllable intrinsic duration

+ *fPW(PW length, position in PW)*

+ *fPPh(PPh length, position in PPh)*

+ *fIFPPh(Initial/Final PPh length, position in PPh)*

➤ 音節固有時長 (Syllable intrinsic duration)

Syllable intrinsic duration = constant + CTy + VTy + Ton

+ *PCTy + PVTy + PTon + FCTy + FVTy + FTon + 2-way*

factors of the above factor + 3-way factors of the above factor

adjustment

- Duration adjustment

$$DurS_i^* = \begin{cases} OriDur(S_i) & , i = 1, m/2, m \\ OriDur(S_i) - DF_i & , 1 < i < m/2, m/2 < i < m, \end{cases}$$

DF_i

$$\begin{aligned} &= M_{TC} / M_{MC} \times [f_{PW}(PW \text{ length, position in PW}) - f_{PW}(2,1) \\ &+ f_{PPh}(PPh \text{ length, position in PPh}) - f_{PPh}(11,6) \\ &+ f_{IFPPh}(Initial / Final PPh \text{ length, position in PPh})], \end{aligned}$$

Modeling Intensity Distribution

➤ 階層性韻律結構效應值 (Hierarchical prosodic effects)

IntS (NRMS) = Syllable intrinsic in tensity

+ *f_{PW}(PW length, position in PW)*

+ *f_{PPh}(PPh length, position in PPh)*

+ *f_{I/PPh}(Initial/Final PPh length, position in PPh)*

adjustment

- Intensity adjustment

$$IntS_i^* = \begin{cases} OriInt(S_i) & , i = 1, m/2, m \\ OriInt(S_i) - DF_i & , 1 < i < m/2, m/2 < i < m, \end{cases}$$

DF_i

$$\begin{aligned} &= M_{TC} / M_{MC} \times [f_{PW}(PW \text{ length}, \text{ position in } PW) - f_{PW}(2,1) \\ &+ f_{PPh}(PPh \text{ length}, \text{ position in } PPh) - f_{PPh}(11,6) \\ &+ f_{IFPPh}(\text{Initial} / \text{Final } PPh \text{ length}, \text{ position in } PPh)], \end{aligned}$$

Modeling boundary breaks

- 階層性韻律結構效應值 (Hierarchical prosodic effects)

BreS (NRMS) = Syllable intrinsic break

+ *f_{PW}(PW length, position in PW)*

+ *f_{PPh}(PPh length, position in PPh)*

+ *f_{IFPPh}(Initial/Final PPh length, position in PPh)*

adjustment

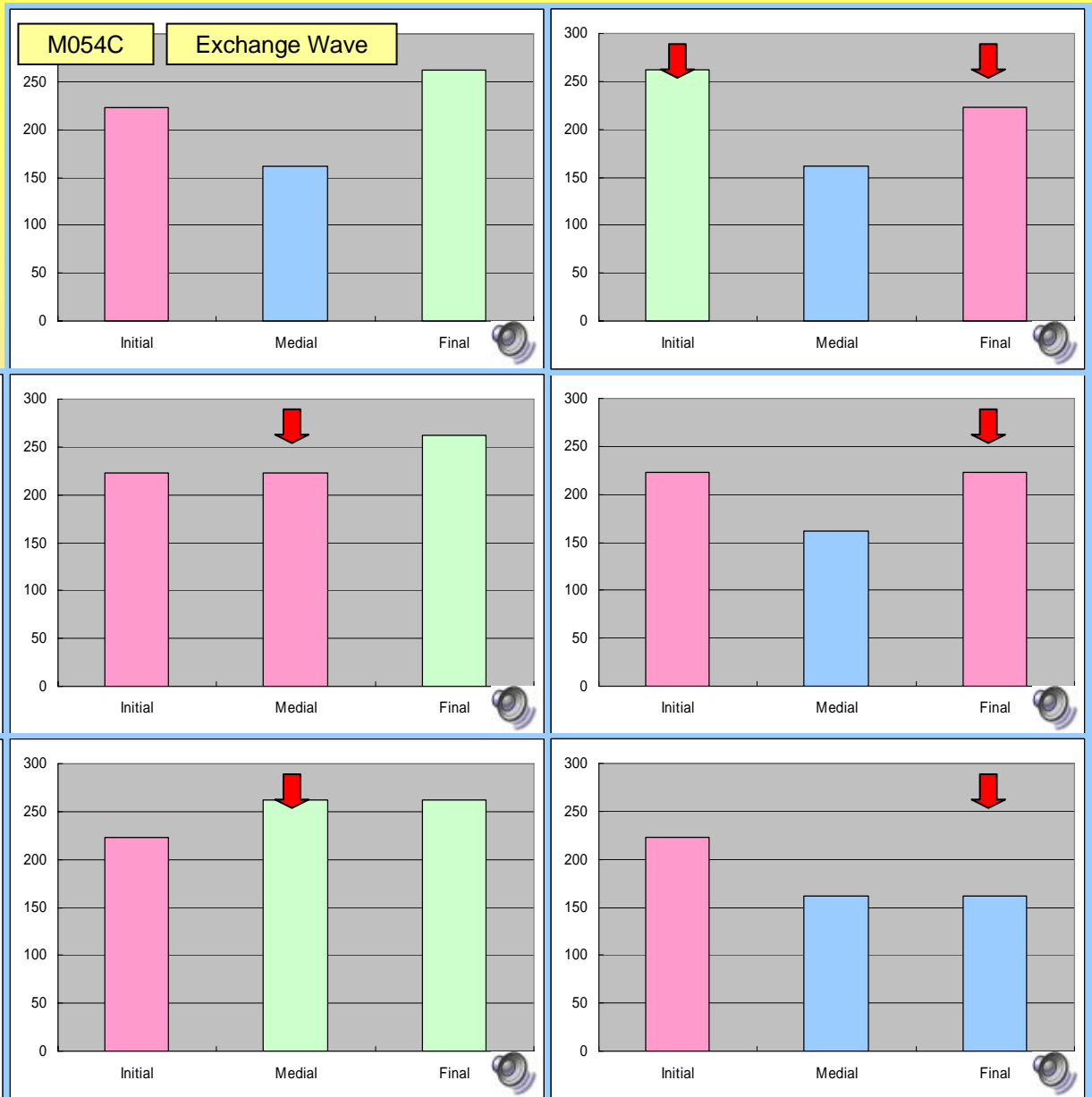
- Break adjustment

$$BreS_i^* = \begin{cases} OriBre(S_i) & , i = 1, m/2, m \\ OriBre(S_i) - DF_i & , 1 < i < m/2, m/2 < i < m, \end{cases}$$

DF_i

$$\begin{aligned} &= M_{TC} / M_{MC} \times [f_{PW}(PW \text{ length, position in } PW) - f_{PW}(2,1) \\ &+ f_{PPh}(PPh \text{ length, position in } PPh) - f_{PPh}(11,6) \\ &+ f_{IFPPh}(Initial / Final PPh \text{ length, position in } PPh)], \end{aligned}$$

Speech Synthesis



Speech Synthesis

