From Ripples to Waves, Tides and Beyond

Chiu-yu Tseng and Chao-yu Su

LANGUAGE AND LINGUISTICS MONOGRAPH SERIES 54

Peaches and Plums

Edited by C.-T. James Huang and Feng-hsi Liu

Institute of Linguistics, Academia Sinica, Taipei, Taiwan

2014

Peaches and Plums, 257-278
2014-1-054-011-000277-1

From Ripples to Waves, Tides and Beyond

Chiu-yu Tseng[1] and Chao-yu Su[2]
Institute of Linguistics, *Academia Sinica*[1]
Taiwan International Graduate Program, *Academia Sinica*[2]

Mandarin tones are known to vary a great deal in an utterance from when they are uttered in isolation. This paper addresses why the 'speech paragraph' is a realistic speaking unit and why discourse prosody is an intrinsic part of naturally occurring continuous speech. We argue that while Chao's (1968) well known ripple-wave analogy describes how sentence intonation is layered over its lower units and triggers tones to modify, the same layering over of global discourse prosody from higher level discourse information also occurs in sentence intonation and triggers more modifications. To empirically prove the existence of prosodic modulations at the discourse level, we adopted a perception based multi-phrase discourse prosody hierarchy that specifies the multi-phrase associative relationship and tested the hypothesis with corpus analysis and computational modeling of data of continuous speech. Results of acoustic analyses show that output discourse prosody can be derived through multiple layers of higher level modulations thereby confirming that tone and intonation alternations are the result of interactions with multiple layers of higher level information. The study also shows how abundant traces of global prosody can be recovered from the speech signal. Therefore, the seemingly random variations in output speech are in fact systematic and predictable, and explain why phonological processing can be achieved.

Key words: HPG, tones, intonation, higher-level contributions, prosody context, F0 contour, cross-over, adjacency

1. Introduction

Tones and intonation have been considered the two most significant prosodic features of Mandarin speech. The most well-known analogy describing their interacting relationship is how tone (the smaller unit) is viewed as riding on intonation (the larger unit):

> "…The question has often been raised as to how Chinese can have sentence intonation if words have definite tones. The best answer is to compare syllabic tone and sentence intonation with small ripples riding on larger waves (though occasionally the ripples may be "larger" than the waves.)…"

This perception-based analogy also specifies the interaction between tone and intonation as additive:

> "The actual result is an algebraic sum of the two kinds of waves. Where two pluses concur, the result will be more plus; when a plus meets a minus, the algebraic addition will be an arithmetical subtraction." (Chao 1968:39)

Inspired by the above analogy, but adopting a top-down perspective, the above interaction can be interpreted as layering over from larger units onto smaller unit. In the following study presented below, we will show that in continuous speech, the layering is more complex than simply tone-(to) intonation, and the largest unit certainly does not stop at sentence intonation.

The issue under discussion is the role of higher level discourse information in fluent continuous speech, how it takes both tones and intonation as sub units and triggers more prosodic modifications, and why phonological processing is possible when tones and intonation appear so varied in output speech. We argue that speech communication rarely takes place in isolated simple sentences, but rather in multi-phrase paragraphs known as fluent continuous speech. As a result, higher level discourse information needs to be included and speech units larger than syntactically defined sentences need to be examined. This allows a top-down perspective that brings the global prosodic phenomenon, beyond lexically specified word tones and syntactically specified intonation, into the prosodic picture, and at the same time requires a systematic account to show how phonological processing is possible.

In the following presentation, we will show that contributions to output prosody come from manifold sources. In addition to contributions from segmental, lexical, phonological and syntactic constraints, global paragraph/discourse prosody is also an intrinsic part of naturally occurring speech. Essentially, global prosody reflects higher level discourse association through patterns of linear chunking and cross-over phrasing, a relationship which cannot be specified by syntactic structures of individual phrases/sentences within a speech paragraph, or pinned down from examining output phrase intonations individually. In other words, global discourse prosody provides a clearly audible prosodic context to the native ear that cannot be entirely represented by corresponding text transcription and punctuation marks. In comparison to well-known literature stating that sentence prosody reflects syntactic structure through overall declination, mid-sentence continuation rise and terminal fall (Halliday 1967; Crystal 1969; Ladefoged 2006), relatively much less phonetic studies have been reported on how discourse prosody reflects phrase and sentence association through global patterns of topical resets, continuation flattening and terminating echo. A discourse perspective makes it clear when speaking in paragraphs, not only are word tones constrained by phrase intonation, phrase intonations are also constrained by discourse association. Therefore, both tone and intonation are required to adjust, and tone adjustments are multifold. As a result, tone, being the smallest prosodic unit, is derived more than once before appearing in output speech; its deviations from the canonical form are but to be expected. Similar deviations also occur to the intonation

of individual phrases. However, the listener is totally unaware of such deviations, and there must be reasons for it. We will show that all such deviations, if they could be justifiably termed so, are collective reflections of discourse structure and speech planning above the sentence level.

There has been reported acoustic evidence on French and English with respect to the correlation between duration and the planning scale, cognitive and psycholinguistic functions of phrase groups such as pauses and timing structure (Keller & Zellner 1996; Zellner Keller & Keller 2001), but not on the melodic properties of F0 (fundamental frequency or pitch in perceptual term) patterns which Mandarin tone and intonation are best known for, except Tseng (Tseng & Su 2008 in Chinese). In the following sections, we will use quantitative analyses to present some of the major F0 features of tone, phrase intonation and global discourse prosody to illustrate why specifying modulations by individual or combined tones and individual intonation outputs are insufficient to portray the prosody of fluent continuous speech, as well as why modifications of both tones and phrase intonations are necessary to point out that some of them are often no longer distinct in output speech. Most importantly, we will argue for the hierarchical nature and structure of discourse prosody above the sentence level. Consequently, prosodic modulations are seen as multi-layered from higher level units to lower ones, and the prosodic relationship is seen as both same-level linear smoothing and higher-level layering at the same time. Finally, we show that the greatly varied output prosody exhibited in continuous speech is not at all random and that is why phonological processing can be achieved.

2. Framework of paragraph and discourse organization

A hierarchical discourse prosody framework of perceived chunking and phrasing units will be used to analyze the speech data. The framework is termed HPG (Hierarchy of Prosodic Phrase Group) (Tseng et al. 2004, 2005a; Tseng 2006) and specifies paragraph and discourse associations beyond phrases and sentences. The HPG specifications state not only adjacent (between-units) sister relationships but also accommodate cross-over (among-units) relationships from larger-scale units and higher-level constraints. Accordingly, the framework and units make identifications of layer-dependent prosodic contributions possible, and at the same time also account for more contributing sources of overall prosodic information from different sized discourse units. Figure 1 is a schematic representation of the HPG.

The HPG hierarchy consists of 5 levels of perceived boundary breaks, B1 through B5, using ToBI notations. Prosodic units are defined by corresponding chunks located inside each level of boundary breaks across the flow of fluent speech. The layered HPG prosodic units and corresponding boundary breaks from the lowest level are the syllable (SYL)/B1, the prosodic word (PW)/B2, the prosodic phrase (PPh)/B3, the breath group (BG)/B4 and the multiple phrase group (PG)/B5 which corresponds to a speech paragraph. A physio-linguistic unit BG correlating with an audible and complete change of breath is included (Lieberman 1967;

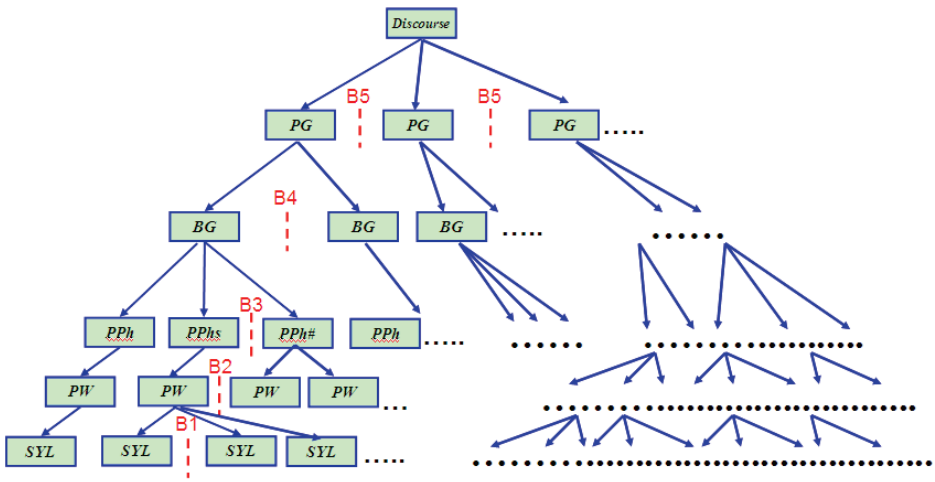

Figure 1: A schematic representation of HPG (Hierarchy of Prosodic Phrase Group). The prosodic units from the lowest level are the syllable (SYL), the prosodic word (PW), the prosodic phrase (PPh), the breath group (BG) and the multiple phrase group (PG) or paragraph. DM (Discourse Marker) and PF (Prosodic Filler) are located within and across PG. Not shown are the boundary breaks at the SYL level (B1), PW level (B2), PPh level (B3), BG level (B4) and PG level (B5).

Tseng 2002) to accommodate global planning due to speaker's regulating of breathing during continuous speech production. The top-down perspective also suggests that discourse prosody context is more than single-unit neighborhood concatenation. However, depending on speaking rate and the size of the paragraph, one change of breath is sufficient when the paragraph is short while several changes of breath are required when the paragraph is long. And this is indeed the case with the speech data we collected over time. Since the HPG framework is a hierarchical one, by default it allows prosodic layers to collapse whenever necessary. Therefore, it is not unusual that BG and PG are sometimes collapsed into one layer, represented as BG.

In the next sections, corpus analysis and modeling of the F0 contours by the HPG framework and various speech genres are conducted by the prosodic units Syllable (SYL), PW, PPh and PG to illustrate (1) that F0 patterns could be derived at each and every unit, (2) why no single unit accounts for output prosody, and (3) how quantitatively these units contribute to output F0.

3. Speech materials and preprocessing

Mandarin speech data were selected for HPG layer-dependent analysis and genre-dependent analysis. Speech data elicited for HPG layer-dependent analysis were: (1) reading of

text of 26 random discourse pieces from Sinica COSPRO (http://www.aclclp.org.tw/corp_c.php) (COSPRO 05 or CNA in the present paper, approximately 6700 syllables, produced by 1 male and 1 female radio announcers M051 and F051) and (2) reading of Chinese Classics (CL, approximately 3,500 syllables, produced by 1 male and 1 female untrained native speakers M056 and F054). Speech data elicited for genre-dependent analysis were: (3) simulated reading of weather broadcast (WB, approximately 7,000 syllables, produced by1 male and 1 female untrained speakers M054 and F054) and (4) the same CL data classified by degrees of rhyming regularity (a) **R**egular (Han and Tang poetry), (b) **S**emi-**R**egular (Tang Ballads from Music Bureau and Song lyrics) and (c) **IR**regular (Qin, Tang and Song classic prose).

Pre-analysis included automatic annotation and manual tagging. At the segmental level, the HTK toolkit was used to automatically obtain consonant and vowel identities by the SAMPA-T notations. Trained transcribers then spot-checked segmental alignments and made manual corrections. At the prosody level, perceived prosodic units and boundary breaks from SYL to PG were manually labeled using the Sinica COSPRO Toolkit (Tseng et al. 2005b).

4. Methodology for F0 analysis

4.1 The command-response model

Because our hypothesis states that there are layers of linear and global contributions from concurrent smoothing and immediate upper level constraints in the composition of the F0 contours, we must use a computational model that could separate multiple layers of waves from the ripples in the F0 and to mathematically account for the interactions. The physiology based command-response model, commonly known as the Fujisaki model (Fujisaki & Hirose 1984) works exactly by the ripple-wave rationale. The model is composed of three components; the major parameters are base frequency, a Phrase Command Ap representing the magnitude of global contour of a phrase and Accent Command Aa representing local humps of accentuation by smaller domain. Mathematically, Aa is superimposed onto Ap to derive the ultimate output F0 contour. Thus the model inherently assumes that F0 is generated from more than one component differing in size and scale, as defined 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

Initially the model was applied to Japanese and English to separate contributions of phrase intonation from accentuated words. When it was used to analyze Chinese, a syllable-timed tone language, the Ap command was applied to phrase units to represent intonation contour pattern; and the Aa command was modified to apply to each syllable to represent the contour pattern of each syllabic tone (Wang et al. 1999). The application was manual and the speech samples were examined phrase by phrase; but the results were positive: overall contour patterns and individual syllable contours could be derived to illustrate how tones and intonation collectively make up the output contour of a short phrase. Computer software was then developed to make automatic extraction of these parameters possible, while tone and intonation patterns were obtained from Chinese, Thai and Vietnamese (Mixdorff 2000; Mixdorff et al. 2003). By this time the model was established as a standard tool to mathematically account for contributions from tones and intonation. In the present study, since we had to extract Ap and Aa values from large amounts of speech data, and make the extractions by HPG specified units and layers, we further modified Mixdorff's method and developed tailored software to accommodate our needs (Tseng & Pin 2004; Tseng & Su 2007). Our software specifies fitting one Aa to the syllable while the scale selected for Ap is one PPh each time. Measurements of these two parameters were obtained for subsequent analysis.

4.2 Modeling layered contributions by the HPG framework

A multi-layer linear regression technique (Keller & Zellner 1996; Zellner Keller & Keller 2001) was adopted for the HPG framework where layer dependent contributions and cumulative results could be obtained. A linear model was developed to predict F0 contour with the Fujisaki parameters for each layer as Figure 2 shows. The regression procedure was executed in two separate parts for each layer, namely, a higher part Ap for predictions of the higher level global values and a lower part Aa for the lower level local values.

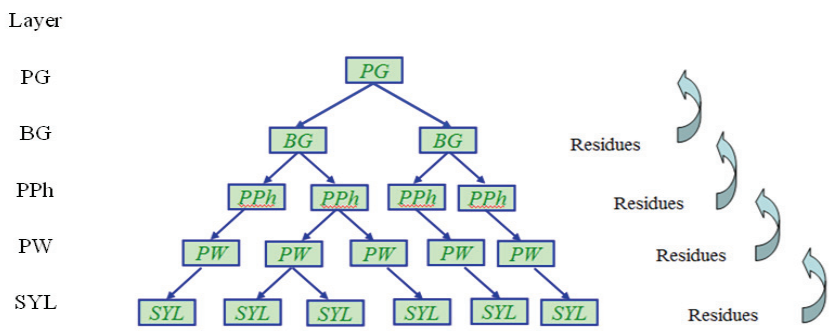

Figure 2: A schematic representation of linear regression predictions from the Syl level upward; whereby each level contributes to final output independently and cumulatively.

For lower part Aa, the multi-layer linear regression starts from the SYL layer. The regression models at SYL, PW and above layers are listed below. ***f*** means the linear regression by function.

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

Since Aa values were extracted from each syllable, predictions were made from the SYL layer; the conditions specified the current, preceding and following tones to account for sisterhood effects. The accuracy at SYL layer is regarded as the contribution from the syllable layer, and the errors (derived residues) are regarded as the predicted contributions from higher layers instead of variation. The errors were further included in the next round of prediction at the immediately higher PW layer with linear regression to account for the contribution at the PW layer. Fine tunings including boundary context and boundary properties of the PW, PPh and PG are calculated to derive higher-level contributions as well. The accuracy sum of SYL, PW and boundary properties is regarded as cumulative contribution for final Aa output.

The same rationale was applied to Ap predictions where the regression procedure begins from the PPh layer. The models at each layer are listed below.

PPh
Ap=*f(FollowingPPh_Length, PrecedingPPh_Length,*
CurrentPPh_Length)+Delta4
BG
Delta1=f(BGSequence)+Delta5
PG
Delta2=f(PGSequence)+Delta6

The procedure starts from PPh and is repeated until reaching the top of the hierarchy, PG, to derive the cumulative contribution from PPh to PG.

5. Tone, intonation and global modulations

5.1 Tone modeling

Enabled by the Fujisaki model (Fujisaki & Hirose 1984), our first immediate goal is to model and predict Mandarin tones in continuous speech through quantitative corpus analysis at the SYL layer.

Speech data

Read Mandarin speech from 4 speakers (2 males and 2 females) of two speech genres, prose (CNA) and varied rhymes (CL), were analyzed. Table 1 summarizes the results of accuracy of tone prediction by the predictions of the Aa component from SYL, PW and boundary effects above PPh.

Table 1: Cumulative accuracy of Aa predictions from SYL, PW and boundary effect above PPh

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%
CNA	F051	38.40%	45.00%	48.43%	51.27%
	M051	41.61%	47.96%	51.33%	54.53%

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%

Procedure and Results

To start with, a tone model was constructed to predict individual tone identities from the speech data. The result shows that the accuracy of cross-speaker prediction ranged from 38% to 46% only. In other words, if tones in output speech are modeled by their F0 patterns syllable by syllable at face value, less than half of the tones could be correctly identified. Next we added the linear context of each tone under consideration, namely, information from a current tone's immediate neighborhood, defined as the context of forward and backward tone smoothing. The prediction of accuracy was increased to a range of 45% to 55% across the 4 speakers. Subsequently, the layering of PW information is added by two factors: (1) PW boundary information that separates pre-boundary tokens from the others. The prediction accuracy was increased to a range of 48% to 61%. (2) PW position sequence that specifies the exact location of the current syllable inside a PW. The prediction accuracy was increased to a range of 51% to 67%. The same rationale was applied to additional layers of higher level boundary information. After considering the factor of PPh boundary information, the prediction accuracy was increased to a range of 54% to 73%. After considering the factor of PG boundary information, the prediction accuracy was increased to a range of 56% to 74%. To demonstrate how boundaries have effects on the preceding prosodic unit, their contribution to tone modeling

was also tallied and a range from 5% to 7% was derived. Finally, cumulative accuracy of Aa prediction was derived at a range from 56.25% to 73.80%.

Discussion

In short, prediction accuracy from single tone to 5 layers of prosodic information was consistent across the two speech genres and 4 speakers. For female speaker F054 producing rhymed classics speech (CL), accuracy of tone prediction improved from 46.21% (by single tone) to 73.80% (after 5 layers of modulations); for male speaker M056 prediction accuracy improved from 39.12% to 66.89%; for female speaker F051 reading prose (CNA), the improvement was from 38.40% to 56.25%; and for male speaker M051 from 41.61% to 59.32%. Cumulative effects from discourse boundaries by speaker ranged from 7.19%, 5.43%, 4.98% and 4.79%, respectively (Tseng & Su 2008 in Chinese). The cumulative accuracy of Aa prediction ranges from 56.25% to 73.80%. The above results demonstrate that tones are in fact hardly identifiable in continuous speech; their immediate linear neighborhood helps little. However, by including multiple layering from various sizes of higher level information, the cumulative predictions of tones improved layer by layer. The same cumulative accuracy also shows how Mandarin output speech is not simply tone strings.

5.2 PPh intonation & global intonation modeling

To confirm that intonation contours by phrase are affected by higher level information, the same quantitative approach is adopted and applied to the intonation and discourse levels for accuracy and contribution analysis.

Speech data

The same elicited Mandarin speech from 4 speakers of two speech genres, prose (CNA) and varied rhymes (CL) used in §3, was analyzed.

Procedure and Results

The same prediction procedure was repeated. A phrase model was constructed to predict the magnitude of the individual phrase from the speech data. Table 2 summarizes the accuracy of intonation prediction (Ap prediction) from the phrase unit PPh, the multi-phrase sub-paragraph at the change of breath BG, and the highest paragraph unit PG.

Table 2: 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%

The results show that accuracy of prediction was also improved across speech genre, speaker and gender when additional contributions from higher levels were added. For female speaker F054 reading rhymed classics speech (CL), accuracy of intonation prediction improved from 58.79% (by single phrase) to 76.66% (after 5 layers of modulations); for male speaker M056 prediction accuracy improved from 37.87% to 73.66; for female speaker F051 producing a prose reading (CNA), improvement was from 80.17% to 87.71%; and for male speaker M051 from 81.53% to 88.20%. The results confirmed that the overall magnitude of individual phrases, namely, intonation contour patterns, is affected by levels of information above the phrase level.

Table 3 shows cumulative tone prediction Aa (73.66% to 88.20%), cumulative intonation prediction Ap (56.25% to 73.80%) and the average of combined predictions of cumulative Aa and Ap predictions (70.28% to 75.23%).

Table 3: The ultimate accuracy of prediction by Aa, Ap and average of Aa and Ap

Corpus	Speaker	Aa	Ap	m/Aa and Ap
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%

Discussion

The results from overall cumulative tone prediction Aa suggest that the tones produced by untrained native speakers (F054, M056 for CL) are more varied (76.66% to 73.66%) than those produced by trained radio announcers (F051, M051 for CNA, 87.71% to 88.20%). One possible reason is speaker dependent, namely, untrained native speakers produced syllabic tones with more variation than trained radio announcers. However, overall cumulative intonation prediction Ap suggests that the prediction of Chinese Classics CL (73.80% to 66.89%) is considerably better than the prediction of prose reading CNA (56.25% to 59.32%). One possible reason is genre dependent, namely, Classics CL are read with more uniformed intonation patterns than discourse pieces. The following analyses were designed to address these possibilities.

5.3 Mandarin tones by Aa patterns

The above results demonstrate how greatly varied the F0 contours of tones in the output of continuous speech are from their canonical forms, and how much their identities depend not on the patterns restricted to the syllable unit alone, but on the interactions from both linear smoothing of the same level sister units and layering of higher level units. The next question is: given the fact that tones do not maintain their canonical contours in output speech, how could phonological processing by the syllable be achieved? We therefore further compare the Aa patterns of Mandarin output tones to see if they remain distinct from each other in continuous speech.

Results

Figure 3 shows the plotting of the modeling of Aa by speaker and genre at the SYL level where neutral tones were treated separately as a fifth tone. The results show that when all levels of higher level information are not considered, though correct prediction of Aa by tone identities is only about 40~45%, the five tones are distinct from each other across speakers and genres. The Aa patterns of each tone are similar across the 4 speakers, as shown in previous studies using the Fujisaki model (Fujisaki & Hirose 1984).

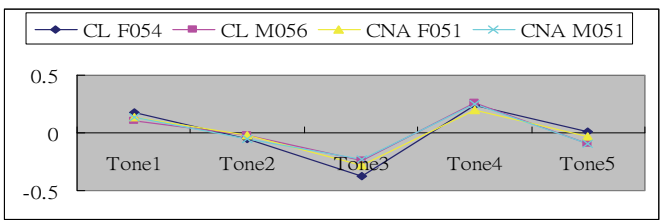

Figure 3: Tone model of Aa. The horizontal and vertical-axis indicate the tone type and average Aa value, respectively.

Discussion

The results confirm that in linear terms, in spite of the fact that over half of the tones produced deviated considerably from their respective canonical forms, the so-called deviations are similar across speakers and genre for each tone and distinct contrasts among the tones remain. Moreover, tones 1, 4 and the other tones could be distinguished by tone height, thus maintaining a HL contrast as well. We believe that these relative F0 contrasts in rapidly changing continuous speech serve the purpose of differentiating tones from each other and are the major reason of why phonological processing can be achieved.

5.4 Mandarin tones by Aa patterns and PW positions

Because the HPG discourse prosody framework (§2) specifies more prosodic units than the syllable for tone and phrase for intonation, a more detailed analysis of the HPG units and levels is necessary to yield a more comprehensive picture of higher level contributions to lower level units. Therefore, in addition to tone relations in linear terms presented in bottom layer, Figure 1, we further examine how tones are modified by the immediately higher level multi-syllable unit, the PW, by paragraph information.

Results

Figure 4 shows plotting of the PW models of Aa by speaker and three PG positions, PG-Initial, PG-Medial and PG-Final. The results show that most of PW-final syllables exhibited pre-boundary F0 lowering except for cases of 4-syllable PW at the PG-Final position by speaker F051. We note here that by definition, the minimum size of a speech paragraph PG

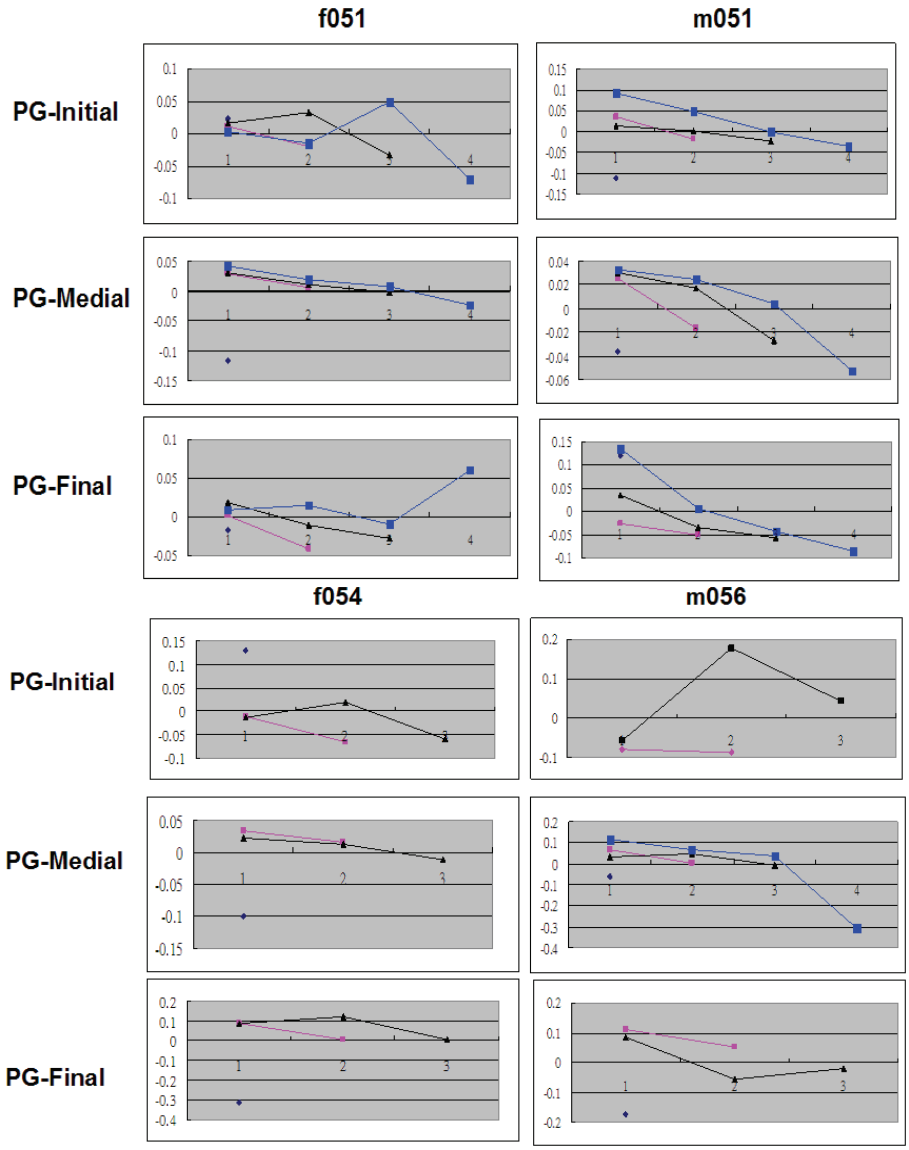

Figure 4: PW model of Aa by PG positions. Each trajectory denotes PW model for specific PW length. The horizontal and vertical-axis indicate the Syl sequence index in PW and average Aa value, respectively.

should consist of at least three PPhs, i.e., PG-initial, -medial and -final. At the same time, when the size of a PG increases, it is achieved by increasing the number PG-medial phrases. That is, the number of samples in the category of PG-medial PPh should always be greater than PG-initial and -final ones at one phrase each per paragraph. In the case of speaker F051, the distribution of 1-, 2-, 3- and 4-syllable PW is 1.16%, 56.89%, 37.40% and 4.56%, respectfully;

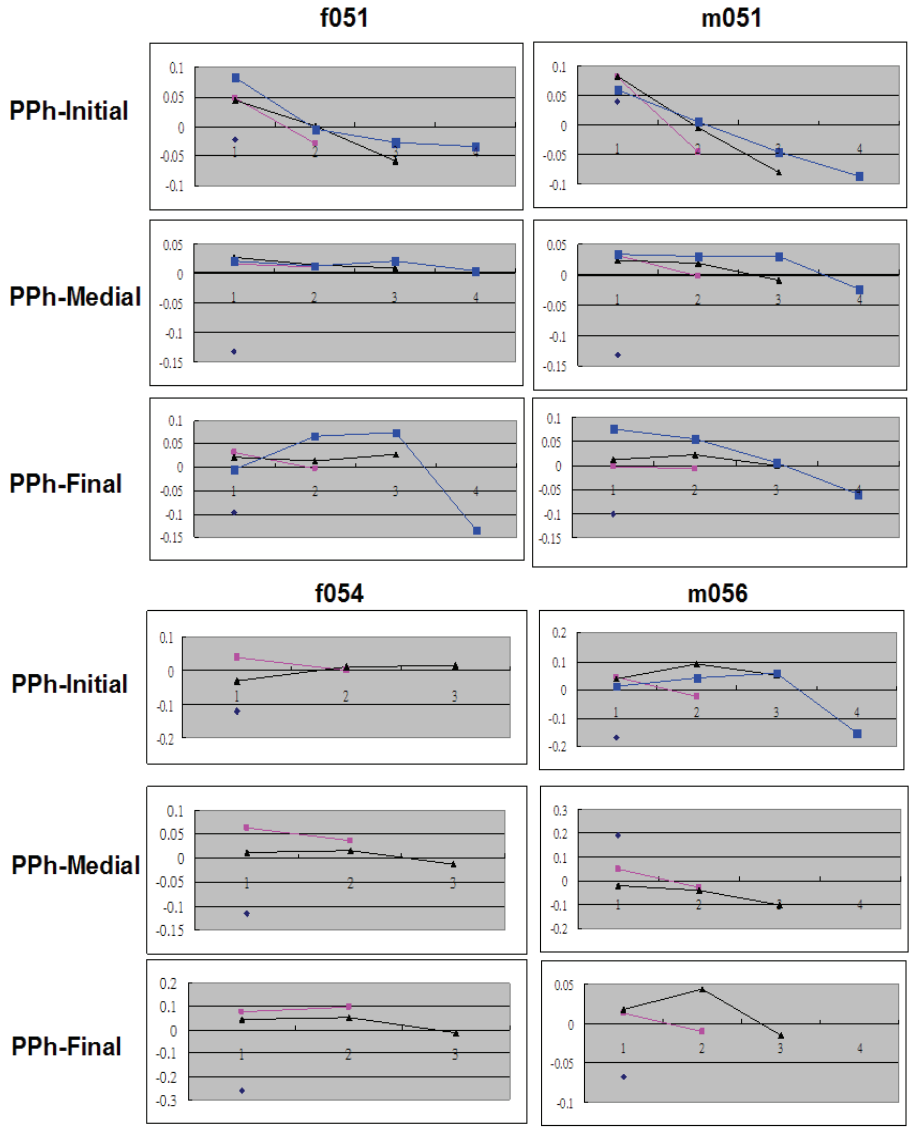

Figure 5: PW model of Aa by PPh position. Each trajectory denotes PW model for specific PW length. The horizontal and vertical-axis indicate the SYL sequence index in PW and average Aa value, respectively.

while in the case of speaker m051, the distribution of PW by the same order is 1.17%, 66.22%, 29.16% and 3.44%, respectfully. In other words, the insufficient samples of 4-syllable PW at PG-final position may be the major reason that contributes to the irregular pattern shown in the case of speaker F051.

The same rationale was applied to the PPh information over tones and Figure 5 shows PW model of Aa by PPh-positions. Declination of Aa is found in PW-final of PPh-medial positions but by different degrees while PPh-medial positions exhibited a flatter pattern across speakers.

Discussion

The results from both the PG level and the PPh level show that higher level positions do have a lowering effect on the pre-boundary PW unit. At the PG level, the PW-final syllables exhibited pre-boundary F0 lowering in accordance with the PG positions except for the PG-Final position of one speaker (F051). At the PPh level, similar pre-boundary F0 lowering of the PPh positions is also found. In addition, the sharpest declination slope is found at the PPh-final position across the board. Results from both the PG and PPh levels indicate that pre-boundary lowering of syllable tone is systematic and at least twofold. Modeling and analysis of tone patterns in relation to different levels of higher level information is complete by now. In the next sections, we proceed to model global modulations of phrase intonation.

5.5 Mandarin phrase intonation by Ap patterns and BG & PG positions

In this section, we present modeling of global modulations of phrase intonation from higher levels of discourse information above the phrase and sentence. Ap modeling by BG and PG positions is carried out. The results of BG and PG modeling are plotted in Figures 6 and 7.

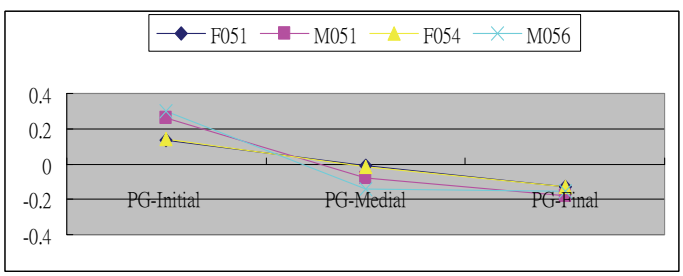

Figure 6: PG model of Ap by PG position. The horizontal and vertical-axis indicate the PG-position index and average Ap value, respectively.

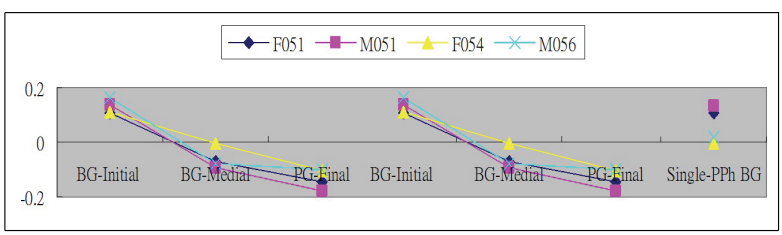

Figure 7: BG model of Ap by BG position. The horizontal and vertical-axis indicate the BG-position index and average Ap value, respectively.

Results

As shown in both Figures 6 and 7, similar patterns of global F0 left-to-right lowering of adjacent units at both the PG and the BG levels are found across the 4 speakers, namely, a within-unit down drift pattern from -Initial to -Medial to -Final, and a within-unit high-low contrast between -Initial and -Final. Figure 7 further shows that the between-paragraph cross-unit association is a low-high contrast between adjacent -Final and -Initial in Ap.

Statistical analyses were performed on all Ap values to see if higher level contributions are significant. Results are summarized in Table 4. All Ap values are classified by BG and PG positions. The number of category is 12 and the df is (1, 7). The results indicate that F0 pattern by higher-level/larger-unit BG- and PG- positions are significant.

Table 4: ANOVA for Ap in different PG & BG positions

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

Discussion

The above results show that in output speech there are contributions of F0 information other than syllabic tones and individual phrase intonation. These results also prove the existence of higher level discourse information in a BG or PG and their contributions to phrase intonation contours. Contributions from higher level information by large scale feature Ap are statistically significant. As shown in Figure 6, the same level within-unit associative pattern is both adjacent down drift and cross-unit high-low contrast, whereas as shown in Figure 7, the same level between-unit associative pattern is adjacent low-high contrast. The statistical significance of higher level contributions in the associative as well as cross-over patterns imply that such information is helpful for discriminating within-unit and cross-unit discourse boundaries. The results further demonstrate that in continuous speech there are more levels of layering than from intonation to tone. The output of both the tone and intonation units have to undergo multiple layers of superimposition and modulations, and are in fact the derived outcome of multiple interactions. Their deviations from the canonical counterparts are systematic and subject to discourse constraints in order for discourse information to be delivered through output prosody. According to the ripple-wave analogy (Chao 1968), there are different sizes of ripples, waves, tides and even bigger tides riding on each other; each sized unit makes contributions to the final output. Collectively, these results prove that modifications of overall intonation contour trajectories are constrained by higher level discourse information, and the modifications are systematic.

5.6 Modulations of Mandarin phrase intonation by speech genre

Having proved the existence of higher level discourse information above the phrase and

sentence, our next questions were whether the distribution of higher level information is speech genre specific and whether stylistic variations could be accounted for by the same discourse framework. If the HPG framework is a productive prosody model, modulations of phrase intonation should be genre conditioned, and different patterns of contribution distribution from higher level discourse information could be found. To address the questions, more speech data were analyzed.

Speech Data

A total of 4 prosody genres were compared (§3), namely, CL classified by degree of rhyming regularity, including regular (**R**), semi-regular (**SMR**), irregular (**IR**), and a simulated weather broadcast (**WIR)** by 3 HPG higher prosodic layers, PPh, BG and PG. Speech data from 4 speakers, 2 males and 2 females, were used.

Procedure

The same analysis procedure from §4.1 was repeated; the feature analyzed was phrase intonation represented by the Ap component and by higher HPG nodes above the PPh level.

Results

Figure 8 shows respective contribution patterns and distributions within and across 4 prosody genres; Table 5 summarizes cross-speaker comparison of prediction percentage by the same parameters. Contributions from three prosody layers PPh, BG and PG are accounted for. Results confirm that contributions from higher level information to output prosody are genre-dependent across the speakers. Each prosody genre possesses distinct distribution patterns in the PPh and BG layers. Degree of regularity is higher-level dependent, as shown in how BG contribution decreases from **R, SMR, IR** to **WIR** (Figure 8). Note that for output prosody **R**, BG layer contributions by speaker account for 28.1% and 51.57%; for output **SMR** 20.39% and 22.67%; for output **IR** 8.7% and 16.52%; for output **WIR** 5.63% and 7.36%. Contributions from the PPh layer are in complementary distribution with the BG layer while contributions from the PG layer are insignificant. The gradation from **R** to **WIR** is also systematic in terms of prosody genre.

Discussion

The results demonstrate that distribution patterns are both genre-dependent and genre-specific. Contributions from both the PPh and BG layers are obligatory across prosody genres, and together they comprise output prosody. The results also confirmed that the HPG framework is a productive one, functioning as a default base form of paragraph and discourse prosody. Dynamic variations of prosody by genre can now be accounted for on an **R-to-IR** or even **-WIR** continuum by varying proportional contributions from the PPh to BG layers, rather than attributing each prosody genre independently.

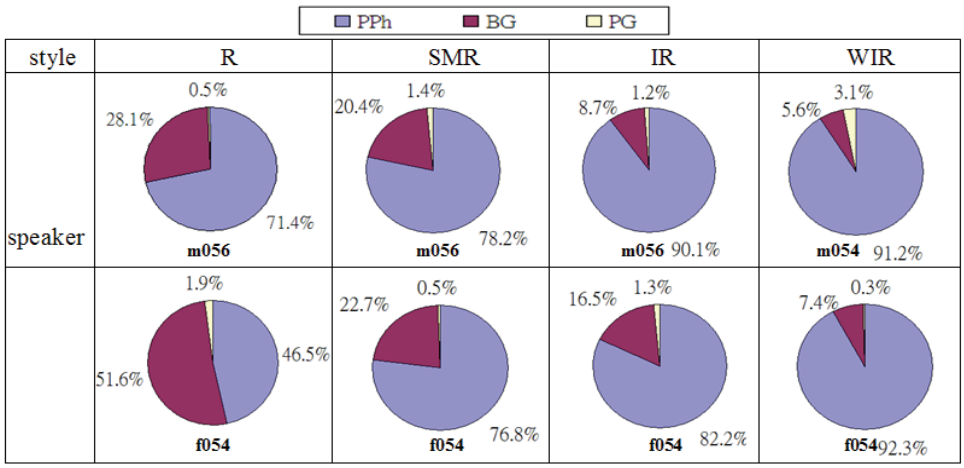

Figure 8: Cross-speaker (m056; m054 and f054) comparison of respective contribution distributions within and across 4 prosody genres regular (R), semi-regular (SMR), irregular (IR) and weather broadcast (WIR); and by 3 HPG prosodic layers, PPh, BG and PG.

Table 5: Cross-speaker (M056; M054 and F054) comparison of prediction percentage by 3 prosodic layers PPh, BG and PG; by 4 genres within and across 4 prosody genres regular (R), semi-regular (SMR), irregular (IR) and weather broadcast (WIR); and by 3 HPG prosodic layers, PPh, BG and PG.

speaker	genre	PPh	BG	PG
M056	R	71.44%	28.10%	0.46%
	SMR	78.22%	20.39%	1.39%
	IR	90.08%	8.70%	1.22%
M054	WIR	91.22%	5.63%	3.14%
F054	R	46.50%	51.57%	1.93%
	SMR	76.82%	22.67%	0.50%
	IR	82.16%	16.52%	1.32%
	WIR	92.30%	7.36%	0.34%

6. General discussion

The above analyses on Mandarin tones and intonation made it clear why they vary in the output of continuous speech and at the same time why phonological processing is still possible. Tone Modeling in Aa, or modeling the ripples only, show F0 contribution by HPG layers at the syllable layer that correct prediction of Aa by tone identities amounts to only 40~45%; while the contribution from PW is 15~20%. This means less than half of syllable tones can be predicted correctly from surface values; the ripples are hardly distinguishable. Furthermore,

by including contextual information of the next higher layer PW, cumulative prediction accuracy at its best (65%) is still less than satisfactory (§5.4). In other words, tone adjacency and limited higher information is not sufficient 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 information from higher paragraph layers. Their contributions to output F0 should not be overlooked.

However, the results from §§5.5-5.6. PPh Intonation & Global Intonation Modeling in Ap of PPh, BG and PG, or modeling layering of waves (PPh), tide (BG) and beyond (PG) show that cumulative accuracy of Ap prediction is indeed improved. The ultimate accuracy of prediction by Aa, Ap (73.66% to 88.20%, 56.25% to 73.80%) and average of Aa and Ap (70.28% to 75.23%) show how the ripples, waves and tides are combined and how the predictions approach the F0 output.

Conversely, the results of Mandarin Tones by Aa Patterns (§5.3, Figure 3) show how when all levels of higher level information are removed from the speech output the tones remain distinct from each other. This is really interesting. We therefore reason these contrasts are sufficient to represent the necessary tone distinctions while distinct tones do not require fully realized canonical forms. Moreover, the same results also imply how the relative distinctive pattern may facilitate phonological processing. In this sense, the relative distinction can also been seen as a representation of linguistic abstraction.

In order to show why tones are not constrained by sentence intonation alone and how there are multiple levels of higher level constraints to the lower level units, we show in Mandarin Tones by Aa Patterns and PW positions (§5.4) how tones are affected by their respective positions in the immediate higher layer (the PW) and how the PW is further conditioned by its respective positions in PG (Figure 4) and PPh (Figure 5). As expected, there are some speaker differences, but the overall patterns are nonetheless similar and the effects from higher level units to lower level ones evident. In summary, results from §§5.3-5.6 make clear how many layers of modulations are required for the ripples in output speech to adjust accordingly and why the adjustments are systematic, productive and predictable.

From modeling intonation patterns at higher levels (§§5.5-5.6 Mandarin Phrase Intonation by Ap Patterns and BG & PG positions), we are able to show how global contour F0 patterns by larger units must also go through systematic adjustments specified by discourse positions, and how patterns across speakers are more similar at the higher levels but more varied at the lower levels. The results suggest that higher-level effects are stable across speakers, as shown in the similar patterns in Figures 6 and 7, reflecting that large-scale global planning of semantic cohesion is a more general strategy for prosody production. The same patterns in Figures 6 and 7 also demonstrate how a global high to low pitch pattern is characteristic of what is said in one breath and one paragraph, rather than a phrase-to-phrase down stepping. Furthermore, how a shift from one paragraph to the next one is marked is by adjacent sharp low to high pitch pattern. Therefore, it is feasible to assume that both small- and large-scale templates are employed by listeners during on-line speech processing; phonological processing is an

important part of it. Large unit look-ahead may require little information well ahead of production time. Simultaneously, it can also been seen as productive rules used to generate speech output and to predict speech outcome. The bidirectional and predictable relationship is the main reason why phonological processing is possible.

Lastly, we tested the productivity of the HPG discourse prosody framework and further tested it with speech data of different genres. From the results of distribution of higher-level information across speech genres, we confirmed our hypothesis that stylistic variations could be accounted for by contributions of distribution from higher level information (Figure 8). The HPG as a prosody base form is indeed a productive one; various output genres can be predicted and generated with one base form. Higher level information contributes to output prosody across prosody genres and speakers while the significance of contributions from the BG layer can NOT be ignored. Cross-genre comparisons also revealed systematic genre-specific layer-dependent patterns of contribution distribution from the PPh and BG layers, respectively. The HPG framework quantitatively accounts for the contribution patterns by prosodic layer and prosody genre. In short, the more regular the prosodic format is, the more contribution comes from the upper layer BG; and vice versa. In summary, output prosody is the cumulative outcome of contributions from the layers involved, as the ripple-wave analogy states. But there are more units than the ripple and wave in the making of discourse prosody.

7. Conclusion

The goal of this paper is a short but comprehensive overview to show how prosodic modifications apply to not only tones and intonation but also to all prosodic units involved, and how output F0 is composed of layers of combined discourse information. An empirical account of why surface F0 variation is systematic and predictable is provided to show how to tease apart the acoustic correlate F0 to account for the melodic composition of speech output in a systematic manner, thus proving how the seemingly highly varied speech output is in fact predictable. Using corpus linguistic data, quantitative analysis and computational modeling, we prove why phonological processing is possible from the seemingly highly varied speech output. In other words, although it is well known that perceptually tolerated articulatory simplifications occur in phonological processes, such as segmental reduction, deletion and assimilation (e.g. Kohler 1990; Hura, Lindblom & Diehl 1992; Steriade 2001), our view is that prosodic modifications in continuous speech are obligatory to realize discourse association. Our other investigations on the temporal allocation patterns and duration alternations have produced similar evidence (Tseng & Lee 2004; Tseng et al. 2005a) to further substantiate the argument. Therefore, we emphasize here that traces of discourse information in continuous speech are abundant. These traces are crucial to speech production and processing, and cannot be overlooked in understanding speech.

We also wish to emphasize here the importance of data selection and methodological considerations. In terms of speech data, using continuous speech instead of single tones or syntactically simple sentences produced in isolation is the essential first step. But lifting tones and intonation units from continuous speech and treating these fragments as unrelated independent entities will not give room for the global picture to surface. In addition, since both tones and intonation are realized most notably through the acoustic correlate the F0, it is therefore critical to adopt an analysis method capable of separating their respective composition in the speech signal. In terms of research methodology, using quantitative analysis and computational modeling enabled us to extract features from large amounts of speech data and test the productivity of abstract nodes by predictability. In summary, we have shown how better understanding of the F0 composition from a discourse perspective helps facilitate a better understanding of speech prosody that goes beyond tones and intonation, and how in fact the actual result is indeed an algebraic sum (Chao 1968:39), but more than the two kinds of waves. In other words, there are more than ripples, waves and tides in the prosodic making of continuous speech prosody.

Chao also pointed out in the same paragraph "…occasionally the ripples may be 'larger' than the waves". In fact, our recent investigation (Tseng et al. 2010) has shown that these larger ripples can be attributed to accentuation, focus and prominence by linguistic structure and by speaker intension. We believe they are the direct reflection in the speech signal of the information structure in the linguistic content. Last but not least, we wish to point out that acoustic investigations of higher level information should not stop short at F0 modulations only. Careful investigations of temporal compositions and loudness patterns are also necessary to yield a more comprehensive picture of what is heard by the listeners. But that would be another chapter of the ripple wave and tide story.

References

Chao, Yuen Ren. 1968. *A Grammar of Spoken Chinese*. Berkeley: University of California Press.

Crystal, David. 1969. *Prosodic Systems and Intonation in English*. Cambridge: Cambridge University Press.

Fujisaki, Hiroya, & Keikichi Hirose. 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustical Society of Japan* (*E*) 5.4:233-242.

Halliday, M. A. K. 1967. *Intonation and Grammar in British English.* The Hague: Mouton.

Hura, Susan L., Björn Lindblom, & Randy L. Diehl. 1992. On the role of perception in shaping phonological assimilation rules. *Language and Speech* 35.1-2:59-72.

Keller, Eric, & Brigitte Zellner. 1996. A timing model for fast French. *York Papers in Linguistics* 17:53-75.

Kohler, Klaus J. 1990. Segmental reduction in connected speech: phonological facts and phonetic explanations. *Speech Production and Speech Modeling*, ed. by William J. Hardcastle & Alain Marchal, 69-92. Dordrecht & Boston: Kluwer Academic Publishers.

Ladefoged, Peter. 2006. *A Course in Phonetics* (5th edition). Boston: Wadsworth.

Lieberman, Philip. 1967. *Intonation*, *Perception*, *and Language*. Cambridge: MIT Press.

Mixdorff, Hansjörg. 2000. A novel approach to the fully automatic extraction of Fujisaki model parameters. *Proceedings of 2000 IEEE International Conference on Acoustics*, *Speech*, *and Signal Processing* (*ICASSP 2000*) 3:1281-1284. Istanbul, Turkey.

Mixdorff, Hansjörg, Hiroya Fujisaki, Gao Peng Chen, & Yu Hu. 2003. Towards the automatic extraction of Fujisaki model parameters for Mandarin. *Proceedings of 8th European Conference on Speech Communication and Technology* (*EUROSPEECH 2003*), 873-876. Geneva, Switzerland.

Steriade, Donca. 2001. Directional asymmetries in place assimilation: a perceptual account. *The Role of Speech Perception in Phonology*, ed. by Elizabeth V. Hume & Keith Johnson, 219-250. San Diego: Academic Press.

Tseng, Chiu-yu. 2002. The prosodic status of breaks in running speech: examination and Evaluation. *Proceedings of the First International Conference on Speech Prosody 2002*, 667-670. Aix-en-Provence, France.

Tseng, Chiu-yu. 2006. Prosody analysis. *Advances in Chinese Spoken Language Processing*, ed. by Chin-Hui Lee, Haizhou Li, Lin-shan Lee, Ren-Hua Wang & Qiang Huo, 57-75. Singapore: World Scientific.

Tseng, Chiu-yu. 2010. Yupian de jipin gouzu yu yuliu yunlü tixian [An F0 analysis of discourse construction and global information in realized narrative prosody]. *Language and Linguistics* 11.2:183-218.

Tseng, Chiu-yu, Shao-huang Pin, & Yeh-lin Lee. 2004. Speech prosody: issues, approaches and implications. *From Traditional Phonology to Modern Speech Processing*, ed. by

Gunnar Fant, Hiroya Fujisaki, Jianfen Cao & Yi Xu, 417-437. Beijing: Foreign Language Teaching and Research Press.

Tseng, Chiu-yu, & Shao-huang Pin. 2004. Modeling prosody of Mandarin Chinese fluent speech via phrase grouping. *Proceedings of Speech and Language Systems for Human Communication* (*SPLASH-2004*/*Oriental-COCOSDA2004*), 53-57. New Delhi, India.

Tseng, Chiu-yu, & Yeh-lin Lee. 2004. Speech rate and prosody units: evidence of interaction from Mandarin Chinese. *Proceedings of the International Conference on Speech Prosody 2004*, 251-254. Nara, Japan.

Tseng, Chiu-yu, Shao-huang Pin, Yeh-lin Lee, Hsin-min Wang, & Yong-cheng Chen. 2005a. Fluent speech prosody: framework and modeling. *Speech Communication* 46.3-4:284-309.

Tseng, Chiu-yu, Yun-ching Cheng, & Chun-Hsiang Chang. 2005b. Sinica COSPRO and Toolkit—Corpora and Platform of Mandarin Chinese Fluent Speech. *Oriental COCOSDA 2005*. Jakarta, Indonesia.

Tseng, Chiu-yu, & Zhao-yu Su. 2007. From one base form to multiple output styles—predicting stylistic dynamics of discourse prosody. *Proceedings of the 8th Annual Conference of the International Speech Communication Association* (*INTERSPEECH 2007*), 110-113. Antwerp, Belgium.

Tseng, Chiu-yu, & Zhao-yu Su. 2008. Yi Fujisaki moxing yanzheng lianxu yuliu zhong zidiao ji yunlüci duiying yu jiecengxing yunlü jiagou HPG de yiyi [Analysis of Mandarin tones and prosodic words as discourse units using the Fujisaki model]. *Proceedings of the 20th Conference on Computational Linguistics and Speech Processing*, 53-65. Taipei, Taiwan.

Tseng, Chiu-yu, Zhao-yu Su, & Lin-shan Lee. 2010. Prosodic patterns of information structure in spoken discourse—a preliminary study of Mandarin spontaneous lecture vs. read speech. *Speech Prosody 2010*, 4 pages. Chicago, USA.

Wang, Changfu, Hiroya Fujisaki, Sumio Ohno, & Tomohiro Kodama. 1999. Analysis and synthesis of the four tones in connected speech of the standard Chinese based on a command-response model. *Proceedings of the 6th European Conference on Speech Communication and Technology* (*EUROSPEECH '99*), 1655-1658. Budapest, Hungary.

Zellner Keller, Brigitte, & Eric Keller. 2001. Representing speech rhythm. *Improvements in Speech Synthesis*, ed. by Eric Keller, Gérard Bailly, Alex Monaghan, Jacques Terken & Mark Huckvale, 154-164. Chichester: John Wiley & Sons.