Beyond Sentence Prosody

Chiu-yu Tseng

Institute of Linguistics, Academia Sinica, Taipei, Taiwan

cytling@sinica.edu.tw

Abstract

The prosody of a sentence (utterance) when it appears in a discourse context differs substantially from when it is uttered in isolation. This paper addresses why paragraph is a discourse unit and discourse prosody is an intrinsic part of naturally occurring speech. Higher level discourse information treats sentences, phrases and their lower level units as sub-units and layers over them; and realized in patterns of global prosody. A perception based multi-phrase discourse prosody hierarchy and a parallel multi-phrase associative template were proposed to test discourse prosodic modulations. Results from quantitative modeling of speech data show that output discourse prosody can be derived through multiple layers of higher level modulations. The seemingly random occurrence of lower level prosodic units such as intonation variations is, in fact, systematic. In summary, abundant traces of global prosody can be recovered from the speech signal and accounted for; their patterns could help facilitate better understanding of spoken language processing.

Index Terms: global prosody, discourse organizations, prosodic modulations, tones, intonation, speaking rate, pause, boundary.

10.21437/Interspeech.2010-3

1. Introduction

This paper examines fluent continuous speech and addresses the role of global paragraph prosody in relation to higher level discourse information. We argue that in addition to prosody from segmental, lexical, phonological and syntactic levels; discourse prosody is also an intrinsic part of naturally occurring speech which the human ear is sensitive to, and which cannot be pinned down from analysis of sentence prosody, nor entirely by corresponding text transcription. Higher level discourse information is viewed as additional information above the sentence whose association is not included in syntactic analysis, but whose manifestation through global discourse prosody is loud and clear. As sentence prosody reflects syntactic structure through overall declination, mid-sentence continuation rise and terminal fall [1, 2, 3]; discourse prosody must reflect phrase and sentence association through global patterns of topical resets, chunking, phrasing and terminating echo. In the following sections we will show how to discover their traces from speech data analysis and present some of the major features of global discourse prosody. We will show how discourse prosodic modulations are layered over lower units, including sentence intonation, in order to derive paragraph prosody, thus demonstrating that output discourse prosody is systematic and predictable.

The paper is arranged as follows. Sec. 2 is a brief introduction of a discourse prosody framework which allows quantitative account of layered contributions by linear regression. Sec. 3 presents the speech data used and annotation rationale. Sec. 4 discusses the nature of discourse boundary breaks and the flexibility of discourse unit size. Sec. 5 discusses discourse prosodic modulations in the pitch and tempo domain, with emphasis on global patterns and about boundary properties and pause duration. Sec. 6 discusses perceived prosodic highlighting and its implication.

2. Paragraph and Discourse Organization

It is well accepted that utterances are phrased into constituents and hierarchically organized into various domains at different levels of the prosodic organization [4, 5, 6]. The planning scale, cognitive and psycholinguistic functions of phrase groups have been well researched through pauses and prosodic timing structure [7, 8] while templates and heuristic segmentation has also been addressed [9]. To test whether global prosody could be tapped, we constructed a perception-based hierarchical discourse prosody framework called the HPG (**H**ierarchy of **P**rosodic Phrase **G**roup) that includes the multi- phrase speech paragraphs as a discourse unit [10, 11, 12]. The framework 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. Figure 1 is a schematic representation of the hierarchy. The layered HPG 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) which corresponds to a speech paragraph. A physio-linguistic unit BG correlating to an audible and complete change of breath is included [13, 14] to accommodate breathing during continuous speech production. Corresponding to the HPG units but not shown in Figure 1 are the 5 discourse boundary breaks B1/SYL, B2/PG, B3/PPh, B4/BG and B5/PG. The relationship of these prosodic units and boundary breaks are paragraph and discourse specified, which can be expressed as SYL<PW<PPh<BG<PG and B1<B2<B3<B4<B5. Additional units are discourse marker (DM) and prosodic filler (PF) [15]. The top-down perspective also suggests that discourse prosody context is more than single-unit neighborhood concatenation.

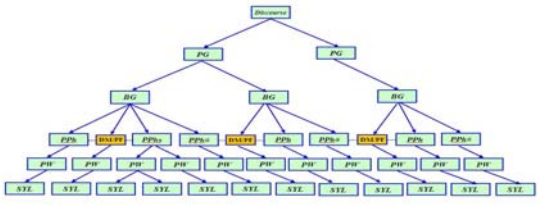


Figure 1 *A schematic representation of HPG (Hierarchy of Prosodic Phrase Group). The prosodic units from the lowest level are the syllable (SYL), the prosodic word (PW), the prosodic phrase (PPh), the breath group (BG) and the multiple phrase group (PG) or paragraph. DM (Discourse Marker) and PF (Prosodic 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).*

Figure 2 illustrates how a multi-phrase paragraph is superimposed by three paragraph specifications to indicate the size and span of a speech paragraph. They are three paragraph positions PG-Initial, PG-Medial and PG-Final. The minimal size of a PG is thus 3 phrases/sentences. [12] While the PG-Initial and PG–final units are single phrases, the number of PG–Medial phrases/sentences is not restricted. The top two layers BG and PG are collapsed the PG layer when a paragraph does not require more than one change of breath. The immediate lower level unit of the PG is PPh, corresponding to an intonation phrase. These perceptually based prosodic units are strictly heard prosodic events that purposely made no reference to other linguistic levels of information such as lexical, phonological or syntactic; so that discrepancies could be examined. [16].

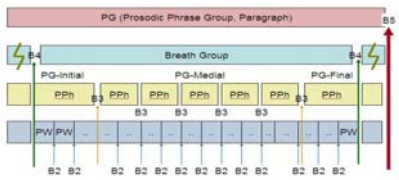


Figure 2 *A schematic representation of a prosodic phrase grouping by PG positions –Initial, -Medial and –Final, superimposed onto PPhs. Note that both the PG-Initial and –Final PPhs are single ones while –Medial PPh can be plural. Not shown here are the syllable layer, and the additional units DM (Discourse Marker) and PF (Prosodic Filler).*

The superimposition from the PG layer to the PPh layer can be viewed as a 3-position template. Each position specifies patterned global prosodic modulation of the superimposed PPh(s). Patterns of global modulation are illustrated in Figure 3 [11, 17].

Figure 3 *A schematic illustrations of global trajectories of pitch contours of a 5-PPh PG*

The perceived template suggests that down-stepping occurs at the PG-Initial and PG-Final phrases, but both their resets and declination slopes are not uniform. The medial phrases are flattened out without distinct patterns. In addition, there is global down-stepping across the paragraph that suggests paragraph association. This illustration shows that listeners perceive a paragraph as a unit, associating different degrees of F0 reset, tilting, flattening and declination to new/given topical information, continuation and terminating echo. It also implies that the planning of discourse units exceeds well above complex sentences. In Sec. 5 we will present modeling of F0 and duration patterns by PG specified positions to show that the perceived overall pattern is supported.

This base form of a multi-phrase paragraph template is a simple one. We propose that only a simple but flexible template is needed for discourse planning, reasoned on the basis of cognitive load. By flexible we mean that paragraph size is an elastic one. Quantitative account of layered contribution using a step-wise linear regression technique was adopted for the HPG framework [10, 11, 12]. From lower levels upward, a linear model of 5 layers is developed to predict the prosodic output over time. Prediction accuracy from the current layer is accepted as contribution from that particular prosodic level while residuals were regarded as contributions from higher levels. Residuals were included in the next round of prediction at the immediate higher level. The same prediction procedure was repeated until the highest level of the hierarchy is reached. Contributions obtained from each layer are added up to derive the ultimate prediction accuracy; thus prediction outcome is cumulative [10, 11, 12]. The regression analysis allows us to test multiple layers of prosodic contributions, and tease apart contributions from higher level information at and by different prosodic levels at the same time.

3. Speech Data and Annotation Rational

We believe that the most suitable type of speech data to examine discourse prosody is narrative instead of short phrases produced in isolation. Our speech data consists of read and spontaneous L1 Mandarin speech, as well as L1 and L2 English speech. Most of the data analyzed are L1 Mandarin. Read speech includes three types of Mandarin L1 speech recorded in sound proof chambers: (1) plain text of 26 discourse pieces from Sinica COSPRO [18] (approximately 6700 syllables, produced by 1 male and 1 female radio announcers), coded as CNA, (2) three types of Chinese Classics (CL) varying in degrees of rhyme regularity (approximately 3,500 syllables, produced by 1 male and 1 female untrained speakers) and (3) simulation of weather broadcast (WB) (approximately 7,000 syllables, produced by1 male and 1 female untrained speakers). All of the text was designed to illustrate discourse speech prosody. Spontaneous speech (Spnl) is university classroom lectures (approximately 90 min or 41,000 syllables, produced by one L1 Mandarin male speaker). English speech is reading of the IPA released version of Aesop's Fable "The North Wind and the Sun" (144 words) recorded in quiet rooms. Speech data from 10 L1 English (E) speakers and 9 Taiwan Mandarin L1 speakers are used [19].

The annotation is designed in accordance with the underlying principle of the framework, and is therefore human perception based rather than text based. Segmental identities are automatically labeled, followed by manual spot checking of alignments. Trained transcribers then listen to the speech data from headsets and manually tag 5 levels of perceived boundary breaks using the Sinica COSPRO Toolkit [18]. Cross-transcriber consistency is checked, and only consistently transcribed data are used for analysis.

4. How Many Boundary Breaks and How Flexible Is the Paragraph?

Instead of the most accepted two boundary breaks, namely, minor phrase break and major phrase break, the HPG discourse framework defines 5 levels of boundary breaks. Among them 3 boundary breaks B3, B4 and B5 are discourse boundary breaks, defined as phrase break, change-of-breath break, and paragraph break, respectively (Sec. 2). Why at least 3 discourse boundaries? In the following presentation, we will show that an extra boundary can help go a long way.

We have previously stated that it is imperative that a speech paragraph is a necessary discourse unit beyond the sentences, but only a simple by flexible base form is needed to accommodate speech paragraphs of different sizes. To show how much the size of paragraph varies, we present in Table 1 the results of discourse units PPh, BG and PG in number of syllables across 3 genres, 2 languages and 24 speakers.

Mandarin RS (read speech), namely, CNA (prose, 2 speakers), CL (rhymes, 2 speakers) and WB (weather broadcast, 1 speaker); spontaneous speech (Spnl, 1 speaker), and RS of L1 E (English, 10 speakers) and L2 E (9 speakers) are used. The results show that the average size by syllable number of a phrase is approximately the same across speaker, speech genre and language (μ/σ from 8/2 to 11/3); the size of a change of change of breath already varied considerably (μ/σ from 16/4 to 69/31); while the speech paragraph show the widest flexibility (μ/σ from 37/4 to 466/258). Spontaneous lecture speech proves to be the most interesting case because the paragraph unit is stretched to 446 syllables. Cross-genre comparison shows that a L1 Mandarin spontaneous paragraph could be 5 to 6 folds of prose reading, 6 to 10 folds of reading rhymed classics and 5 folds of weather broadcast. Cross-linguistic comparison shows that the same genre of L1 paragraph can be 11 times larger than L2 read speech. In order to complete such a large speech unit, 6.4 changes of breath is required while other speech genres only required anywhere from 2 to 3 changes of breath.
These results illustrate that paragraph size is indeed highly flexible, reflecting cognitive threshold of speech production planning by speech genre and language; and the elasticity of the planning template. However, if only two boundary breaks are used, the sentence will be the largest discourse unit, the paragraph unit may not emerge, and none of the above results would surface. In that case much of the prosody beyond the sentence will remain unsolved.

Table 1 *Size of discourse unit PPh, BG and PG by number of syllables across speech genre, speaker and language, namely, Mandarin (M) Read Speech (RS) CNA, CL, and weather broadcast WB; spontaneous speech (Spnl) ; and RS of L1 and L2 English (E)*

Genre/Lg	RS / M					SpnL / M	RS / E	
Speaker	CNAM051P	CAN F051P	CLM056	CLF054	WBM054	LSL	L1	L2
PPh (μ / σ)	10 / 4	10 / 4	9 / 2	8 / 2	12 / 5	8 / 6	11 / 3	10 / 4
BG (μ / σ)	27 / 11	28 / 14	20 / 8	16 / 4	45 / 23	69 / 32	17 / 2	21 / 3
PG (μ / σ)	84 / 54	79 / 44	68 / 37	45 / 17	88 / 69	446 / 258	37 / 4	39 / 7

5. Discourse Prosodic Modulations

In this section, we will discuss some of the major acoustic characteristics of global prosodic modulations using Mandarin tones, phrase intonation and discourse prosody to illustrate. Mandarin tones are probably among the most studied prosodic units in linguistic and speech research. The common consensus is to nail down tone segments and their identities in the speech flow, usually aided by the immediate context as context. However, recognition of tones has proven to be a difficult task after decades of much research. One of the misconceptions is that the tone patterns of tokens extracted from continuous speech somehow remain largely intact. This may be true for only to a limited numbers of candidates under investigation. Since both tones and intonation patterns are F0 perceived in unison, it is important to separate their respective patterns in the signal. We adopt the Command-Response Model [20] which makes possible extraction by parameters of different magnitude into three components indicating respective magnitude of global contour of larger domain Ap, local humps of smaller domain Aa and base frequency. The nature of the model is layering over from a higher level large unit to multiple lower level units; output prediction is therefore cumulative. By defining a higher and larger unit and the lower and smaller units, the model can be used for multiple predictions as well. When applied to the Mandarin, Ap and Aa values have long been used to predict tones and intonation [21, 22]. We use the Command-Response model to the HPG hierarchy and predict the F0 output by each layer and the ultimate and cumulative predictions. The model also enables us to model tones independently from intonation by layers of higher level information and other related factors such as boundary effects. In turn, phrase intonation can be modeled and examined independently. In Section 5.1., we will show what tone and intonation modeling reveals and how by including layers of discourse and boundary information the prediction is improved layer after layer.

5.1. Tone, Intonation and Global Modulations

5.1.1. Tone Modeling

Read L1 Mandarin speech from 4 speakers of two speech genres, prose CNA and varied rhymes CL were analyzed. Table 2 shows the accuracy of tone prediction (Aa predictions) from SYL, PW and boundary effects above PPh, i.e., contributions from the lower HPG layers. Table 1 shows the results that the cumulative accuracy of Aa prediction ranges from 56.25% to 73.80%. To start with, a tone model was constructed to predict individual tone identities from the speech data. The result shows that accuracy of cross-speaker prediction ranged from 38% to 46% only, indicating less than half of the tones were correctly identified. Next we added information from a current tones immediate neighborhood tone as context, and the prediction of accuracy was increased to a range of 45% to 55%. Subsequently, the layering over of PW information is added by two factors: One is PW boundary information that separates pre-boundary tokens from others, and the prediction accuracy was increased to a range of 48% to 61%. Another factor is PW position sequence that specifies the exact location of the current syllable inside a PW, which increased the prediction accuracy to a range of 51% to 67%. Additional boundary information above the phrase unit PPh was further added. The prediction accuracy was increased to a range of 54% to 73% when PPh boundary information was considered; and to a 56% to 74% when PG boundary information is considered. The contribution from different discourse boundaries was also tallied, ranging from 5% to 7%. In short, prediction accuracy from single tone 5 layers of prosodic information was consistent across speech genre, speaker and gender. For female speaker F054 of rhymed classics speech CL, accuracy of tone prediction was improved from 46.21% (by single tone) to 73.80% (after 5 layers of modulations); for male speaker M56 prediction accuracy was improved from 39.12% to 66.89%; for female speaker F51 of prose reading CNA 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 [23]. The above results demonstrate that tones are in fact hardly identifiable in continuous speech; their immediate neighborhood helps little. It is the cumulative contributions from higher level information that jointly help their identities to emerge.

Table 2 *Cumulative accuracy of Aa prediction 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
		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%

In turn, the same prediction procedures were applied to the intonation and discourse levels for accuracy and contribution analysis. Table 2 shows the accuracy of intonation prediction (Ap prediction) from the phrases unit PPh, the multi-phrase sub-paragraph at the change of breath BG, and the highest paragraph unit PG. A phrase model was constructed to predict the magnitude of individual phrase from the speech data. The result shows 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 of rhymed classics speech CL, accuracy of intonation prediction was improved from 72.98% (by single phrase) to 73.80% (after 5 layers of modulations); for male speaker M56 prediction accuracy was improved from 64.13% to 66.89%; for female speaker F51 of prose reading CAN from 54.41% to 56.25%; and for male speaker M051 from 57.43% to 59.31%.

Table 3 *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%

Table 4 shows cumulative tone prediction Aa (73.66% to 88.20%), cumulative intonation prediction Ap (56.25% to 73.80%) and average of combined predictions of cumulative Aa and Ap predictions (70.28% to 75.23%).

Table 4 *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%

The interacting relationship has long been acknowledged and described in Chinese linguistics by analogy of small ripples riding on large waves [24]. This analogy suggests layering over from the higher level unit (intonation) to the lower level one (tone). The question then is: Is there only two layers involved? The above results demonstrate that in continuous speech the output of tones and phrase intonation units have to undergo multiple layers of superimposition and modulations, and are in fact derived outcome. Their deviations from the canonical counterparts are systematic by constraints from discourse information in order to deliver discourse prosody.

5.1.2. Aa and Ap Patterns by HPG

In this section, we will present derived patterns of predicted Aa values at the syllable layer to show predicted tone patterns, and derived patterns of predicted Ap at the PG layer to show how predicted intonation patterns by three PG positions are distinct [25]. Patterns of phrase down-stepping within and across paragraph units will also be presented to show how phrase down-stepping is better understood as related discourse units rather in isolation.

5.1.3. Aa (Tone) Patterns at the Syllable Layer

Figure 4 shows Aa predictions at the SYL level; the patterns of 4 Mandarin lexical tones 1 to 4 and the neutral tone 5 are distinct. The Aa patterns of each tone are similar across the 4 speakers and two speech genres. In spite of the distinct patterns, correct prediction of Aa by tone identities is only 38% to 46% (Sec. 5.1) whereby residuals are treated as contributions from higher levels and included in the predictions at the immediate higher level the PW layer, then up to the PPh layer [23, 26].

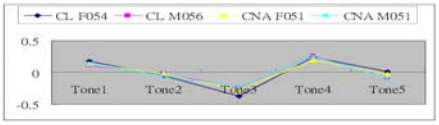


Figure 4 *Tone model of Aa. The horizontal and vertical-axis indicate the tone index and average Aa value, respectively.*

5.1.4. Modeling PPh F0 with and without Higher Level Information

To demonstrate the how predicted F0 contours with and without PG specifications –Initial, -Medial and –Final differ, the F0 contour patterns from one 3-phrase speech paragraph was extracted. Ap values of the same paragraph was also derived, first without PG effects, then with added PG effects. Extracted and predicted F0 contours are shown in Figure 5 for comparison. The upper panel shows the extracted F0 contours (blue) from the speech sample and predicted PPh contours made without PG effects (pink). Note the three predicted PPh contours are almost identical, but each contour differs from their extracted counterparts, except the medial PPh. The lower panels show the predicted PPh contours with added PG effects (pink) and the same extracted F0 contours (blue) from the speech sample. Note now the predicted PPh contours are closer to their extracted counterparts, thus illustrating the significant role of PG effects in paragraph prosody. The predictions also gives reason to why F0 resets of individual phrases in a paragraph context is not uniform.

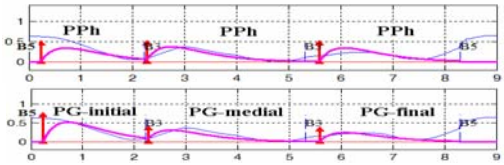


Figure 5 *F0 predictions with PG effect. The upper panel denotes intonation prediction with PG effect (red). The lower panel denotes overall F0 prediction with PG effect (red). X-axis denotes time; Y-axis normalized log F0.*

5.1.5. Global F0 Down-Stepping

Declination resets and down-stepping has been examined by the hierarchical organization of utterances, and most evident within a phrase [27]. Averaged Ap values were derived by PG positions from the same speech data to see if patterns of global down-stepping can be derived. Figure 6 shows plotting of derived patterns of two adjacent PG's. Global down-stepping by PG positions is evident and consistent; the overall high-to-low pattern can be expressed as PG-Initial>PG-Medial>PG-Final [25]. Down-stepping occurs both within and across phrase boundaries; resets of phrasal F0 are not uniform but systematic. The global down-stepping signals cross-phrase paragraph association from higher level discourse information, and is thus characteristic of discourse prosody. The pattern further explains why the pitch contour patterns of individual phrases in a discourse context are not uniform, and why sentence or utterance intonation differs substantially when from when it is uttered in isolation. Adopting the top-down

perspective and from an even higher level of discourse information above paragraphs, the repeated global down-stepping between two successive paragraphs features a sharp F0 low-to-high contrast, suggesting that change of topical information is more evident across larger discourse context.

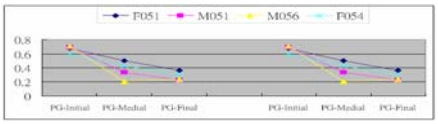


Figure 6 *Average Ap by PG-position of two adjacent PG's by speaker and by speech data type. The horizontal axis represents the PG-position index. The vertical axis represents the average Ap values.*

5.2. Duration and Tempo Adjustments

In this section, we discuss modulations of discourse prosodic units, speaking rate, boundary pause and boundary properties in relation to higher level discourse information. Duration modeling by discourse units is presented to show how each prosodic layer in the framework accounts for tempo pattern and cumulatively contributes to output tempo patterns. Syllable-cadence templates from each prosodic layer are derived and cross-phrase cadence patterns are also modeled to account for the output tempo structure associated with prosody organization.

5.2.1. PW Tempo

Tempo patterns of read L1 Mandarin speech CNA from 2 speakers were analyzed by the discourse hierarchy. Predictions at the SYL layer is not presented here due to large amount of classes defined [for detailed analysis see 28]. The fist layer of duration modeling in this paper is at the PW layer. Prediction by linear regression is derived and plotted in Figure 7, where each plotted point represents the corresponding regression coefficient of a syllable at a specific position in a PW, while plotted lines represent the overall PW duration pattern by syllable number. X-axis represents syllable index in PW; Y-axis represents the prediction of normalized values. Positive coefficients indicate that the duration of the syllable at the specific position is longer than the average value over the mean residue, while negative coefficients indicate shorter duration. The overall pattern of PW features lengthening of the last or pre-B2-boundary syllable.

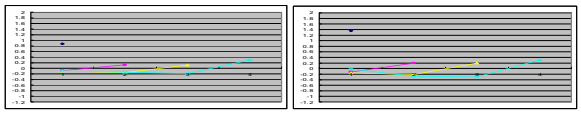


Figure 7 *Regression coefficients of syllable durations obtained for speakers F051 (left) and M051 (right) using the PW model. The X-axis represents the position of each syllable within a PW; the Y-axis represents the coefficient values.*

5.2.2. PPh Tempo

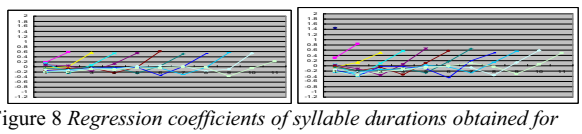


Figure 8 *Regression coefficients of syllable durations obtained for speakers F051 (left) and M051 (right) using the PPh model. The X-axis represents the position of each syllable within a PPh; the Y-axis represents the coefficient values.*

The same duration modeling was applied to the next prosodic layer the PPh layer. An adaptive threshold of 10 syllables is applied. Prediction by linear regression is derived and plotted in Figure 8. The overall cadence pattern of the phrase unit PPh is featured by different degree of pre-B3-boundary lengthening of the last two syllables, and shortening of the antepenultimate syllable.

5.2.3. Phrase and Paragraph Tempo

The same duration modeling is repeated at the paragraph layer. Predictions by PG positions –Initial, -Medial and –Final are plotted and plotted in Figures 9, 10 and 11, respectively. The overall cadence pattern of the PG-Initial PPh in Figure 9 shows slightly longer durations on the first syllable and pre-boundary lengthening by one syllable. Figure 10 shows similar pattern of pre-boundary syllable lengthening is found at the PG-Medial PPhs, but the first syllable of the PG-Medial PPh's is shortened. Note that duration adjustments for PG-Medial PPhs are not as distinct as PG-initial ones. Fig. 11 shows the coefficients of PG-Final PPh. Contrary to patterns the PG-Initial and –Medial PPhs, the paragraph final syllable of PPhs is shortened.

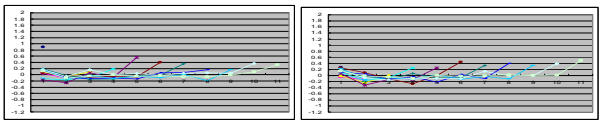


Figure 9 *Regression coefficients of syllable durations obtained for speakers F051 (left) and M051 (right) using the PG-Initial PPh model. The X-axis represents the position of each syllable within a PG-Initial PPh; the Y-axis represents the coefficient values.*

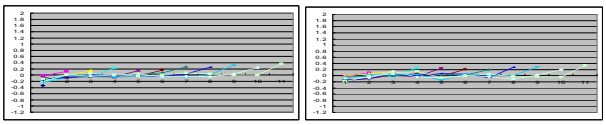


Figure 10 *Regression coefficients of syllable durations obtained for speakers F051 (left) and M051 (right) using the PG-Medial PPh model. The X-axis represents the position of each syllable within a PG-Medial PPh; the Y-axis represents the coefficient values.*

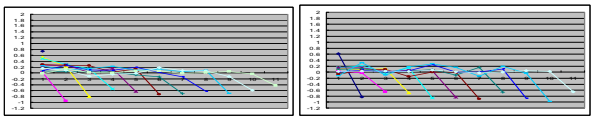


Figure 11 *Regression coefficients of syllable durations obtained for speakers F051 (left) and M051 (right) using the PG-Final PPh model. The X-axis represents the position of each syllable within a PG-Final PPh; the Y-axis represents the coefficient values.*

Figure 12 shows an example of duration predictions for 2 speakers at the SYL layer, PW layer, PPh layer, Paragraph (BG) layer, and the cumulative predictions next to the extracted tempo pattern from the speech paragraph. The positive coefficients at the PW and PPh layers reflect pre-boundary lengthening while the negative coefficients reflect pre-boundary shortening. A clear distinction between PG-initial and PG-final prosodic phrases is evident. However, cumulative overall predictions still reflect an effect of final-syllable lengthening.

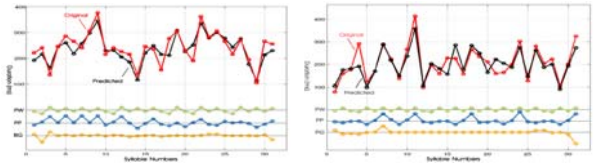


Figure 12 *shows an example of duration predictions for speakers F051P (left panel) and M051P (right panel) at the SYL layer (plotted in pink), PW layer (yellow), PPh layer (blue), Paragraph (BG) layer (purple), cumulative predictions (red) and extracted tempo pattern from the speech paragraph (black).*

5.2.4. Speaking Rate and Discourse Tempo

It has long been reported that particular tempo adjustment by larger discourse units reflects discourse organization [29], but tempo patterns for Mandarin has drawn little attention. The reason is simple: Mandarin is a syllable-timed language. It is usually accepted that syllable timing is the most important feature thus averaged syllable duration represents both speaking rate and tempo patterns. Interestingly, the much studied phrase-final lengthening, usually referred to as final lengthening [30, 31, 32, 33] is often examined phrase by phrase, and the unit often the last syllable, also. In the following discussion, we will demonstrate that paragraph tempo exists and how it is discourse constrained.

5.2.4.1 Dynamic Speaking Rate

The speaking rate of Mandarin is usually represented by averaged syllable duration. In the following section, we will show that this proves to be too crude a measure to represent the dynamics of speech tempo, and will demonstrate that tempo adjustment is in fact, global and systematic. First, we derived mean syllable duration from a wide range of speech genre, speaker and language to show the picture this kind of analysis portrays. Table 5 shows that the mean syllable duration for L1 Mandarin read speech ranged from 167 to 261ms, which by chance is by the same speaker who adjusts the overall speaking rate to read different text. That is, a slower rate to read rhyme classics and faster rate to read text of weather broadcast. Overall speaking rate of spontaneous lecture speech (167ms) is the same as weather broadcast. In short, not much difference is found for L1 mandarin. English read speech shows a slight difference between L1 and L2 speakers at 219ms and 258ms, respectively, that the L2 speakers are overall slower when reading English. Other than results reported above, we cannot tell whether distinct tempo patterns are associated with any of the parameters, or what whether the duration pattern changes over time.

Table 5 *Mean syllable duration (ms) across speech genre, speaker and language, namely, Mandarin (M) Read Speech (RS) CNA, CL, and weather broadcast WB; spontaneous speech (Spnl) ; and RS of L1 and L2 English (E)*

Genre/Lg	RS / M					SpnL /M	RS / E	
Speaker	CNA/M051P	CNA/F051P	CL/M056	CL/F054	WB/m054	Spnl	L1	L2
SR (ms)	191	200	197	261	167	166	219	258

We further derived syllable duration patterns of the same L1 Mandarin speech by speaker and genre. The results are plotted in Figure 13. Two successive paragraphs are plotted for each speech genre. Regardless of speaker and genre, read speech shows a steady slow-down pattern which can be characterized as PG-Initial<-Medial<-Final [25]. The PG-initial PPh is the fastest, the PG-medial the slower; and the PG-final the slowest. The patterns for the two successive paragraphs are identical, suggesting that adjustment by discourse position is consistent. These results are consistent with global F0 down-stepping (Sec. 5.1.5, Figure 6). However, the spontaneous lecture speech showed a distinctly different pattern which can be characterized as PG-Initial<-Medial>-Final, or fast-slow-faster [34]. The slowing down from PG-Initial position ends at the PG-Medial position and the speaker accelerate steadily to the end. Both of the two distinct patterns in Figure 13 reflect discourse planning and organization, differentiated by genre. It is clear that adjustment of paragraph speaking rate is systematic over time, whereby speech genres RS and Spnl are featured in different patterns. These results also present a finer picture of paragraph speaking rate and tempo than mean syllable duration.

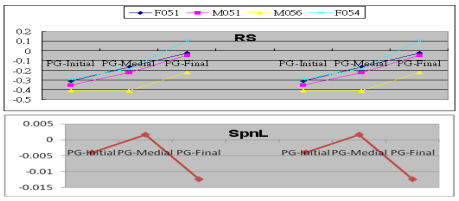


Figure 13 *Tempo allocation patterns by 3 paragraph positions and speech genres read speech (RS) shown in upper panel and spontaneous lecture speech (Spnl) shown in lower panel. Two successive paragraphs are plotted. The vertical axis represents normalized mean value of PPh tempo by PG position.*

5.2.4.2 Discourse Tempo Unit

In this section, we will discuss discourse tempo from the perspective of pre-boundary lengthening. Our tempo analysis by discourse organization (Sec. 5.2.1 to 5.2.3) shows that in read speech (RS), pre-boundary lengthening occurs at each prosodic level except the PG-Final position (See Figures 8 through 11 for respective patterns of PW, PPh, PG-Initial, PG-Medial and PG-Final), that ultimate output tempo is in fact cumulatively derived. Speech paragraph end still shows lengthening (see Figure 12). But note that the same results also demonstrate that phrase tempo is more accurately reflected by a cadence marked by the shortening of the antepenultimate syllable followed by lengthening of the last two syllables (Sec. 5.2 Figure 8), and with respect to duration of the phrase-initial syllable as well. At the same time, the results from averaged syllable duration from RS can also be interpreted as lengthening. The question then is: by what unit and how systematic? We suspect that the duration of the last syllable is not sufficient to account for phrase or discourse tempo adjustments, and hypothesize that lengthening patterns should be examined by discourse units and organization, not by the syllable. The results of pre-boundary duration patterns are derived by discourse units the Syllable, PW and PPh across speech data CNA and CL and 4 speakers are presented in Figure 14 [35].

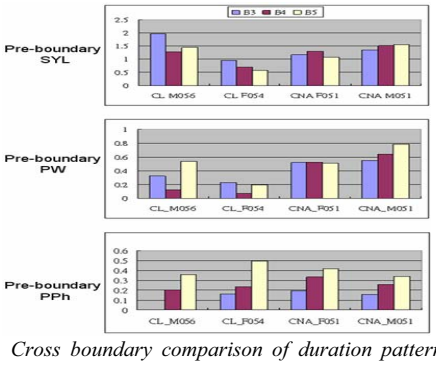


Figure 14 *Cross boundary comparison of duration patterns by prosodic units syllable (SYL), PW and PPh. The horizontal axis*

represents indexes of the speech data and speaker. The vertical axis denotes normalized average duration of prosodic units.

The duration patterns of the pre-boundary syllable are derived (Figure 14, upper panel). Though lengthening does occur, no consistent pattern is found by boundary type, speaker or genre. Next, the duration patterns of the pre-boundary PW are derived (Figure 14, middle panel). Again, lengthening occurs, but no consistent by boundary type, speaker or genre, either. Last, the duration patterns of pre-boundary PPh are derived (Figure 14, lower panel). Consistent lengthening patterns are found by discourse boundary across speaker and speech genre. Only lengthening by the larger PPh unit is consistent; the lengthening pattern can be expressed as pre-B5 PPh>pre-B4 PPh>pre-B3 PPh. The results suggest that pre-boundary lengthening is therefore a discourse phenomenon; global tempo modulations affect not the final syllable or prosodic word only, but the entire phrase as a whole. The results are also consistent with global speaking rate modulations (Sec. 5.4.2.1) regardless of the timing structure of the language. Thus, it is no surprise why the lengthening of lower level units is random.

5.2.5. Pause and Boundary

5.2.5.1 How Reliable Is Pause?

Boundary breaks, usually a period of silent pause, are considered to be the most important cue of boundary location across speech flow [36] and noted in the ToBI annotations. [37] Indeed both the L1 Mandarin, L1 English and L2 English speech data (Table 6) suggest a systematic pattern by boundary pause duration, i.e. B3<B4<B5. The larger the discourse unit is, the longer the pause duration regardless of speaker, speech genre and language. But the analysis also revealed great variations of pause duration at the same time, especially within-PG PPh boundary break B3. When we tried to develop automatic speech segmentation across speech flow, it was found that pause duration is an adequate cue to locate B4 and B5; but not for B3. In other words, only boundaries of speech paragraphs could be correctly identified, but not the multiple phrases within [38]. Figure 15 shows plotting of the distribution of pause duration by discourse boundary B2, B3 and B4 and speaker. The speech data is read speech of prose CNA. It is quite evident that the pause duration of B3 varied the most. Since the boundary breaks are manually annotated and checked for cross-transcriber consistency, the results suggest that pause duration may not be the primary cue for discourse boundary identification. The question then is how can PPh boundary B3 be perceived without pause?

Table 6 *Pause duration (ms) by break (B3, B4 and B5), language Mandarin (M) and English (E) and genre Read Speech (RS) CAN, CL, weather broadcast WB; spontaneous speech (Spnl) ; and L1 and L2 E*

μ / σ	B3	B4	B5
RS_CNA/M051P	249 /207	520 /124	621/113
RS_CNA/F051P	229 /140	339 /172	394 /237
RS_CL/M056	267/105	486 /142	729 /321
RS_CL/F054	190 /117	497/155	782 /271
RS_WB/M054	165/145	490/123	555/166
SpnL_LSL	423/429	739/299	1153/498
RS_L1 (E)	197/135	515/196	762/173
RS_L2 (E)	355/252	543/180	725/267

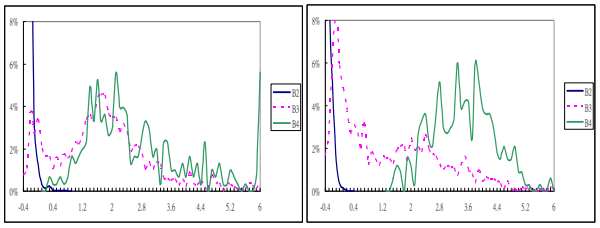


Figure 15 *Plotting of the distribution of pause duration of discourse boundary breaks B2 (blue solid), B3 (pink dotted) and B4 (green solid) in read speech (RS) CAN for speakers F051P (left) and M051P (right).*

5.2.5.2 Boundary without Pause

Since perception of within-PG phrase boundary B3 is consistent across transcribers, while the B3 pause duration proved insufficient B3 for automatic boundary identification, we hypothesized that crucial information may be found in the immediate boundary neighborhood that constitute an important part of the prosodic context. Instead of considering only one immediate neighboring syllable of annotated B3, i.e., one pre- and post-B3 syllable only, we defined immediate between-PPh neighborhood by the last 4 syllables of a preceding PPh and the first 3 syllables of the following PPh and compared their duration pattern. This definition is in accordance with our tempo analysis where the largest size of PW is 4 syllables (see Figure 7), and if proven suggests that (1) within-PG PPh neighborhood is not constituted by the smallest and lowest prosodic unit the syllable, and (2) immediate prosodic context is also constrained by at least some higher level information. Note that the cross-boundary contrast is more distinct in the revised model than that from the previous model. The results of sharp duration contrast are illustrated in Figure 16.

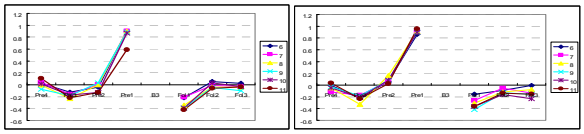


Figure 16 *Sequential position of each syllable between boundary break B3 and normalized duration pattern for speakers F051P and M051P. The X-axis denotes syllable sequence by boundary break B3; the Y-axis is the coefficient of normalized duration value.*

Accordingly, we have included factors of duration to fine-tune the linear regression model, and recalculated the predicted contributions from the PW layer to the final prosody output under the HPG framework. The TRE of duration improved by 10%, the overall prediction of the output prosody is consequently improved by 5%. In addition, the layered predictions are now more consistent with the actual break distributions in the speech data. Based on the above results, we believe that a detailed analysis of residual distributions of every prosodic layer (from syllable to PPh) can yield more stable and general patterns that lead to better prediction. The results also suggest that boundary decision is not by pause duration alone, but by contrastive neighborhood prosodic states as well. Evidence of boundary neighboring F0 contour patterns and intensity contrast also showed similar results. The results enable a better prediction of B3 and provide support to the idea that prosodic states relate more to higher level information [38]. Therefore, boundary CNA be signaled without pause when sufficient information of the immediate prosodic context is available while pause. In a later study, we have also shown that the most salient cue for boundary

identification is the combination of pre-boundary prosodic state and pause. The study also prompted subsequent studies of contrastive patterns by acoustic correlates, further supporting the significance of signal contrasts [35, 39].

6. Prosodic Highlighting

Another perceived feature of narrative prosody is focus and emphasis across the speech flow. Perceived emphases are manually tagged for RS (read speech) CNA and Spnl (spontaneous lecture speech). The distribution of these emphases of each genre is analyzed by units PPh and PG (phrase and paragraph) and position within [34]. The results are presented in Figure17.

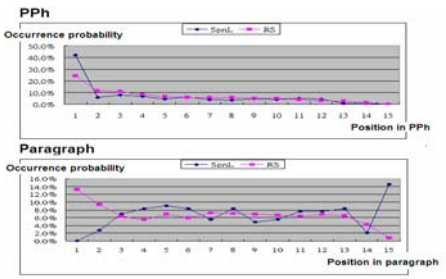
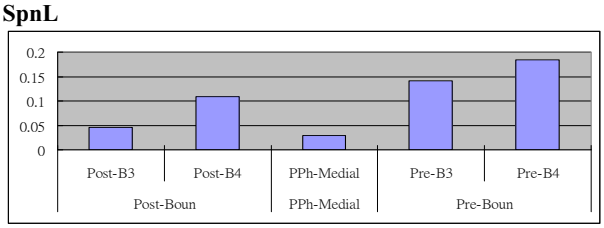


Figure 17 *the distribution of perceived emphasis by speech genre PPh position (upper panel), and BG position (lower panel).*

The distribution by phrase position (Figure 17, upper panel) shows similar overall tendency except the onset of the phrase. Tha is, in RS the distribution is PPh-Initial (50%)>-PPh-Medial (35%) >PPh–Final (15%), whereas in Spnl PG-Initial it is PPh-Initial<(31%)<-Medial (43%)>–Final (26%), respectively (Figure 17, upper panel). Or, emphases are more evenly distributed across spontaneous speech. We think that more distribution at the Ph-initial position in RS reflects structure/syntax information while more even distribution across the phrase of Spnl reflects both the speaker's intension and content highlights. Nevertheless, the distribution by paragraph position (Figure 17, lower panel) shows different results. The distribution of emphasis for RS is PG-Initial >PG-Medial>PG–Final; whereas for Spnl is the opposite PG-Initial<PG-Medial<PG–Final. Since the perceived emphasis are usually a PW, preliminary analysis reveals that these PWs are more in the nature of key terms, most of them nouns or compounds. Combined with the analyses of paragraph size (see Sec. 4 and Table1), we believe that these prosodic highlights are more information related and therefore adopt a view that the prosody of lecture speech is information structure in addition to discourse structure. By information structure we adopt a broad view to mean roughly structural and semantic properties of utterances relating to the discourse content, the actual and attributed attention states of the discourse participants, and the participants' attitudes. Thus notions like focus, presupposition, given vs. new, theme vs. rheme and the various dichotomies such as topic vs. comment or focus, ground or background vs. focus, etc. are subsumed.

The duration patterns of emphasized prosodic words are analyzed in relation to the overall tempo of their embedding PPh and by boundary type B3 and B4 (Figure 18). Results show that emphasized units are lengthened in SpnL than the tempo of the current PPh regardless of boundary type and positions. However, the duration pattern of emphasized units in RS is consistent with global tempo patterns where lengthening is a pre-boundary property. The overall global speaking rate remains the same. These results suggest that emphasis is more marked in SpnL than in RS [34].

SpnL



RS

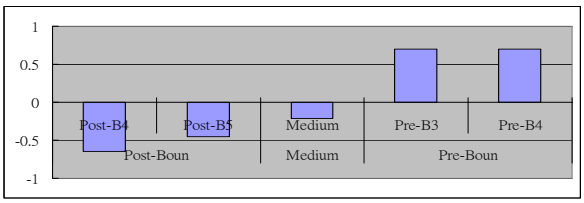


Figure 18 *Relative tempo of perceived emphasis, by position in PPh and discourse boundary type; and by speech genre. X-axiis represents position PPh and boundary type; Y-axis represents the relative tempo of emphasis. Zero means the tempo of emphasis is equal to current PPh tempo.*

We view these perceived emphasis as a form of prosodic highlighting, and believe there may very well be genre related strategies of highlighting that can be derived from the speech signal. In addition to tempo patterns, pitch patterns are also directly to discourse structure [40]. We therefore speculate that prosodic highlights also suggest that the weighting of information chunks may be pinned down from the acoustic signals, and will explore further in the future.

7. Discussion

We have demonstrated from the discussions and examples above that prosodic chunking and phrasing occur not only at the sentence level, but also at the discourse level. Our proposal is such that higher-level discourse information takes syntax, phonology and lexicon as sub-level units, and hierarchical contributions add higher units to lower ones, thereby triggers a series of systematic prosodic modulations in order to signal topical change, continuation and termination. The higher level chunking and phrasing is hierarchically specified, thereby includes not only in neighborhood smoothing but also cross-phrase phrase association. The size of the discourse unit may vary considerably, as evidenced by different speech genre; but it is precisely the associative prosody that holds the information chunks together and when necessary, all the way to the end. From a top-down perspective, it becomes clear how lower level units are subject to layers of higher level specifications and how each unit adjust differently layer by layer to cumulatively yield the final outcome. Our examples of tone and intonation modeling (Sec. 5.1.1 to 5.1.4) portray modulations in the pitch domain, show how these units are deviated from their canonical form for a reason; and how systematic global pitch adjustment include both within- and between-phrase down-stepping. The modeling of duration and tempo patterns portrayed, again, why and how adjustment must take into account contributions from all layers involved (Sec.5.2.1 to 5.2.3). The lengthening patterns by discourse unit the PPh, not by the syllable or words, offers yet another support to the dynamics of overall speaking rate modulations and how tempo units and cadence to beyond words (Sec.5.2.4).

Our perception based analysis also enabled us to find out why discourse boundary breaks can be signaled without a period of silent pause, thus proving that pause is not necessarily the most important boundary cue and why pause durations are random when it is only a secondary cue (Sec. 5.2.5). The perceived prosodic highlights provides a promising start for future attempts that hope to separate the speech segments with clear patterns from those that are greatly distorted, thereby suggesting prosody as a major cue to regulate information weighting. All of our examples argue that paragraph and discourse prosody is not composed by serial concatenation of individual sentence intonation [41]. Consequently, extracting individual phrases and sentences continuous speech and examining them as unrelated units would not explain why identical syntactic structure is often produced with greatly varied output intonation. We therefore propose that well known prosody features such as pre-boundary continuation rise, intonation declination and phrase final lengthening [1, 2, 3] be studied in and by larger chunks, and in association with discourse as well as information structure.

8. Conclusions

Our perception based prosody investigations started from a linguistic perspective, by which we assume that human language production is rule based, and human speech processing is essentially abstracting meaning from speech sounds, often fragmented and distorted. We presented evidence to show why more understanding of higher level information as in discourse effects to fluent speech is essential, and how cross-phrase templates of prosody-related melodic as well as rhythmic cadence, intensity and boundary patterns may together account for the necessary speech planning in text reading and spoken discourses. We believe that our perception motivated multi-phrase PG model offers at least in part a knowledge base and viable framework for formulating theories of higher prosodic organization manifested through speech prosody. We believe the same framework can be adopted to accommodate any discrete intonation model at the PPh level. Recent studies have shown that our idea is helpful to continuous speech segmentation [42]. In summary, in naturally occurring speech, topical and information structure in discourse prosody is an intrinsic part, overriding prosody at the segmental, lexical, syntactic levels. Better understanding of continuous speech prosody is necessary to facilitate more efficient spoken language processing.

9. Acknowledgements

The author would like to thank Professors Lin-shan Lee, Sin-Horng Chen, Hsiao-chuan Wang, Hsin-min Wang, Yih-Ru Wang and Yuan-Fu Liao for their long-term support and collaboration, and Dr. Chen-Yu Chiang for his participation of late. Special thanks go to Professor Lin-shan Lee for the invaluable data of his classroom lecture. Thanks to group members of the Phonetics Lab, Institute of Linguistics, Academia Sinica go to Fu-chiang Chou (1995-99), Shu-Hua Ren (1996-2000)Shou-De Lin (1998-99), Chia-Chuan Si (1998-20), Da-de Chen (1998-2000), Ru-Chun Tseng (1998-2000), Shao-huang Pin (2000~2004), Yun-Ching Cheng (2000 --present), Wei-Shan Lee (2000--present), Feng-Lan Huang (2002~2004), Yeh-lin Lee (2003~2004), Chi-ching Chen (2003~2004), Bau-Ling Fu (2004~2005), Chun-Hsiang Chang (2004, 2007), Zhao-yu Su (2005--present) and Chi-Feng Huang (2008--present).

10. References

[1] Crystal, D. 1969. *Prosodic Systems and Intonation in English*. Cambridge: Cambridge University Press.

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

[3] Ladefoged, P. 2006. *A Course in Phonetics.* Boston, MA: Wadsworth, 5th Edition.

[4] Shattuck-Hufnagel, S., Turk, A., 1996. A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguist Research*, 25(2): 193.

[5] Gussenhoven, C. 1997. Types of focus in English? In Daniel Buring, Matthew Gordon and Chungming Lee (eds.) *Topic and Focus: Intonation and Meaning: Theoretical and Crosslinguistic Perspectives*. Dordrecht: Kluwer.

[6] Selkirk, E. 2000. The interaction of constraints on prosodic phrasing. In Merle Horne (ed.) *Prosody: Theory and Experiment*, Dordrecht: Kluwer. pp. 231-262.

[7] Keller, E., Zellner, B., Werner, S., and Blanchoud, N. 1993. The prediction of prosodic timing: Rules for final syllable lengthening in French. *Proceedings of ESCA Workshop on Prosody*, Lund, Sweden, pp. 212-215.

[8] Zellner, B. 1994. Pauses and the temporal structure of speech."in E. Keller (ed.) *Fundamentals of Speech Synthesis and Speech Recognition*, pp. 41-62, Chichester: John Wiley.

[9] Cutler, A. & Butterfield, S. 1992. Rhythmic cues to speech segmentation: Evidence from Juncture misperception. *Journal of Memory and Language*, 31:218-236.

[10] Tseng, C., Pin, S., Lee, Y., 2004. Speech prosody: issues, approaches and implications. in Fant, G., H. Fujisaki, J. Cao and Y. Xu Eds. *From Traditional Phonology to Mandarin Speech Processing, Foreign Language Teaching and Research Process*, pp. 417-438.

[11] Tseng, C., Pin, S., Lee, Y., Wang, H. and Chen, C. 2005. Fluent speech prosody: Framework and modeling, *Speech Communication* (Special Issue on Quantitative Prosody Modeling for Natural Speech Description and Generation), Vol. 46:3-4, pp. 284-309.

[12] Tseng, C. 2008. Corpus Phonetic Investigations of Discourse Prosody and Higher Level Information (in Chinese). *Language and Linguistics*, 9(3): 659-719.

[13] Lieberman, Philip. 1967. *Intonation, perception, and language.* Cambridge: M.I.T. Press.

[14] Tseng, C. 2002. The prosodic status of breaks in running speech: Examination and Evaluation. *Proceedings of the 1st International Conference on Speech Prosody 2002*, (Apr. 11-13, 2002), Aix-en-Provence, France, pp. 667-670.

[15] Tseng, C., Su, Zh., Chang, C. and Tai, C. 2006. Prosodic filers and discourse markers—Discourse prosody and text prediction. *TAL 2006 (The Second International Symposium on Tonal Aspects of Languages)*, (April 27-29, 2006), La Rochelle, France.

[16] Tseng, C. and Su, Zh. 2007. What do speakers do and why –The story of prosody-syntax non-overlap and higher level Discourse Information. *Oriental COCOSDA 2007*, (December 4-6, 2007), Hanoi, Vietnam, pp. 27-32.

[17] Tseng, C. 2006. Prosody analysis. In *Advances in Chinese Spoken Language Processing,* edited by Chin-Hui Lee, Haizhou Li, Lin-shan Lee, Ren-Hua Wang, Qiang Huo, World Scientific Publishing, pp. 57-76, Singapore.

[18] Tseng, C., Cheng, Y. and Chang, C. 2005. Sinica COSPRO and Toolkit—Corpora and platform of Mandarin Chinese fluent speech. *Proceedings of Oriental COCOSDA 2005*, (Dec. 6-8, 2005), Jakarata, Indonesia, pp. 23-28.

[19] Meng, H., Tseng, C., Kondo, M., Harrison, A. and Visceglia, T. 2009. Studying L2 suprasegmental features in Asian Englishes: A position paper. *Proceedings of Interspeech 2009*, (Sep. 6-10, 2009), Brighton, U.K. pp. 1715-1718.

[20] Fujisaki, H. and Hirose, K. 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustical Society of Japan (E),* 5(4): 233-241.

[21] Mixdorff, H., 2000. A novel approach to the fully automatic extraction of Fujisaki Model parameters. *Proceedings of the ICASSP 2000*, vol. 3, pp. 1281-1284.

[22] Mixdorff, H., Hu, Y. and Chen, G., 2003. Towards the automatic extraction of Fujisaki Model parameters for Mandarin. *Proceedings of Eurospeech 2003*.

[23] 鄭秋豫、蘇昭宇 2008. 以 Fujisaki 模型驗證連續語流中字調及韻律詞對應於階層性韻律架構 HPG 的意義. *第二十屆自然語言與語音處理研討會*, (Sep. 4-5, 2008), 台北, 台灣. pp. 53-65.

[24] Chao, Y. R., 1968. *A Grammar of Spoken Chinese*. Berkeley and Los Angeles, California: University of California Press.

[25] Tseng, C. and Su, Zh. 2008. Discourse prosody and context–Global F0 and tempo modulations. *Proceedings of the Interspeech 2008*, (Sep. 22-26, 2008), Brisbane, Australia, pp. 1200-1203.

[26] Tseng, C. 2010. An F0 analysis of discourse construction and global information in realized narrative prosody (in Chinese). *Language and Linguistics*, 11(2): 183-218.

[27] Ladd, D. R. 1988. Declination "reset" and the hierarchical organization of utterances. *Journal of the Acoustical Society of America* 84:530-544.

[28] Tseng, C. and Fu, B. 2005. Duration, intensity and pause predictions in relation to prosody organization. *Proceedings of the Interspeech 2005*, (September 4-8, 2005), Lisbon, Portugal, pp. 1405-1408.

[29] O'Shaughnessy, D. 1995. Timing patterns in fluent and disfluent spontaneous speech. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing,* pp. 600 – 603. Detroit, Michigan, U.S.A.

[30] Crystal, Thomas H. & House, Arthur S. 1982. Segmental durations in connected speech signals: Preliminary results. *The Journal of the Acoustical Society of America*, 72(3): 705-716.

[31] Crystal, Thomas H. & House, Arthur S. 1988. Segmental durations in connected-speech signals: Current results. *The Journal of the Acoustical Society of America*, 83(4): 1553-1573.

[32] Edwards, J. and Beckman, M. 1987. Perception of final lengthening. *The Annual Meeting of the Linguistic Society of America*, 13 pages.

[33] Beckman, M. & Edwards, J. 1990. Lengthenings and shortenings and the nature of prosodic constituency. In *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, edited by John Kington and Mary Beckman. Cambridge; New York : Cambridge University Press.

[34] Tseng, C., Su, Zh, and Lee, L. 2010. Prosodic patterns of information structure in spoken discourse—A preliminary study of Mandarin spontaneous lecture vs. read Speech. *Proceedings of the Speech Prosody 2010*, (May 11-14, 2010), Chicago, U.S.A.

[35] Tseng, C. and Su, Zh. 2008. Boundary and lengthening—On relative phonetic information. *The 8th Phonetics Conference of China and the International Symposium on Phonetic Frontiers*, (April 18-20, 2008), Beijing, China. 6 pages.

[36] Yang, L. 2004. Duration and pauses as cues to discourse boundaries in speech. *Proceedings of the Speech Prosodcy 2004*, (March 23-26, 2004), Nara, Japan.

[37] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P. Pierrehumbert, J. and Hirschberg, J. 1992. TOBI: A standard for Labeling English Prosody. In *Proceedings of the 1992 International Conference on Spoken Language Processing*, Vol. 2, pp. 867-870. Banff, Canada

[38] Tseng, C. and Chang, C. 2008. Pause or no pause? –Prosodic phrase boundaries revisited. *Tsinghua Science and Technology*, 13.4: 500-509.

[39] Tseng, C., Su, Zh. and Lee, L. 2009. "Mandarin spontaneous narrative planning—Prosodic evidence from National Taiwan University Lecture Corpus." *Proceedings of the Interspeech 2009*, (Sep. 6-10, 2009), Brighton, U.K. pp. 2943-2946.

[40] Menn, L. and Boyce, S. 1982. Fundamental frequency and discourse structure. Language and Speech 25(4): 341-379.

[41] Tseng, C. and Su, Zh. 2008. What's in the F0 of Mandarin Speech –Tones, Intonation and beyond. *ISCSLP 2008 (The 6th International Symposium on Chinese Spoken Language Processing)*, (December 16-19, 2008), Kunming, China, pp. 45-48.

[42] Chiang, C-Y. 2009. *Unsupervised Joint Prosody Labeling and Modeling for Mandarin Speech*. Hsinchu: National Chiao Tung University dissertation.