# DYNAMIC DISCOURSE SPEECH TEMPO AND PHONOLOGICAL TIMING

*Chiu-yu Tseng*[a] *& Chao-yu Su*[b]

[a]Institute of Linguistics, Academia Sinica, Taipei, Taiwan;
[b]Taiwan International Graduate Program, Academia Sinica, Taipei, Taiwan
cytling@sinica.edu.tw

## ABSTRACT

The present study shows how the dynamics of output speech rate can be attributed to systematic contributions from layers of discourse units as well as phonological timing. Quantitative analysis of three modes of Mandarin continuous speech reveals tempo cadence by discourse units, thus demonstrating multiple layers of discourse tempo and boundary lengthening contribute to modulations of output speech tempo in addition to segmental duration. The surface dynamics at face value is thus predictable; both prosodic timing and phonological timing from natural speech data can be accounted for.

**Keywords:** discourse tempo cadence, dynamic speech rate, prosodic timing, phonological timing

## 1. INTRODUCTION

There exists a large body of literature on speech timing and rhythmic patterns from the phonological viewpoint of ischrony. Perception studies focused on temporal perception of small-scale speech units like syllables and words (see [1, 7] for syllable- vs. stress-timed studies example). Production studies focused on measurements of segmental duration at face value. Almost all of the data used were elicited minuscule speech units rather than continuous speech. Almost all of the production analysis measures segmental duration at face value. For example, the well-known PVI index compares vowel ratios at the word level (for more recent accounts [4, 6, 16]; another measures within-syllable consonant-vowel distribution ratio at various phrasal or sentential positions (for example [2, 9]). Both approaches lifted syllables or words from small fragments of elicited speech and used fine-grained measurements of segmental duration as references. Studies on Mandarin, a well-known syllable-timed language, is no exception (for two recent examples see how Mandarin uses syllables as proximate phonological unit by [3, 8]).

However, we are interested to know whether the syllable or segmental duration, by itself or averaged, could sufficiently represent the surface dynamics in output continuous speech. If not, how phonological isochronic timing and surface dynamics could be related. Using Mandarin to illustrate, we propose that while the smallest Mandarin timing unit is the syllable, discourse tempo units also contribute to tempo modulations, thus prosodic timing and surface dynamics should in fact be attributed to collective contributions instead of modulations of segmental durations alone.

In the following studies, we will show that the contributing factors to output discourse tempo consist of (1) segmental duration, (2) discourse temporal patterns and (3) discourse boundary patterns. While consonant and vowel durations contribute to fine grained temporal modulations, discourse units contribute, in a superimposed manner, to modulations by larger units

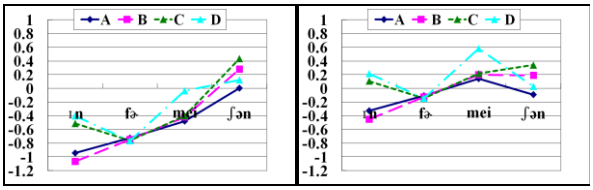
In the following study, we will present (1) why measuring output segmental duration can be misleading (Sec. 2), (2) why each discourse prosody unit has its cadence patterns and contributes systematically to overall discourse tempo (Sec. 5.1) and (3) why pre-boundary lengthening is systematic by larger discourse units (Sec. 5.2). The analyses will also illustrate how discourse constrained global modulations can be derived from quantitative analysis of speech corpora.

## 2. THE ROLE OF SEGMENTAL DURATION

To illustrate how segmental durations may contribute to output duration modulations, but are no part of prosodic timing, a 4-syllable English word *INFORMATION* embedded in carrier sentence *I said X five times* is used. Speech samples produced by four L1 American English speakers are used. The ratio of syllable sequences after normalizing possible speaker effects are shown in the upper

panel of Figure 1). We note that the longest syllable is the last unstressed syllable *–tion* instead of the stressed syllable *–ma*. However, we then normalized the number of phones in each syllable, and the derived pattern (Figure 1, lower panel) shows the longest syllable is in fact the stressed syllable *–ma*.

**Figure1:** Normalized relative duration pattern of syllable sequences of the word INFORMATION. The upper and lower panels indicate the normalization by speakers and the number of phones, respectively. The X-axis represents the syllable index; the Y-axis represents the normalized duration. A, B C and D represent speakers.




The results suggest that though segmental duration contributes to output tempo, a duration pattern independent from segmental contribution can be derived from the speech data and better represents stress defined prosodic timing. The same rationale is then applied to analyze Mandarin speech data where the normalizations are further refined to derive discourse tempo cadence patterns (Sec.5).

## 3. DISCOURSE AS PROSODY FRAMEWORK AND UNIT

To account for discourse contribution to speech tempo and discourse timing, a hierarchical discourse prosody framework the HPG (Hierarchy of Prosodic Phrase Group) [14] is used. The framework specifies levels of perceived chunking and phrasing units located inside each level of perceived boundary breaks (B). The layered HPG prosodic units and corresponding boundary breaks are the syllable (SYL)/B1, the prosodic word (PW)/B2, the prosodic phrase (PPh)/B3, the breath group (BG)/B4 and the multiple phrase paragraph (PG)/B5.

## 4. SPEECH DATA

The speech data are read L1 Taiwan Mandarin microphone speech recorded in sound proof chambers differing by two speech genres: (1) plain text of 26 discourse pieces from Sinica COSPRO 1 coded as CNA (approximately 6700 syllables, produced by 1 male and 1 female radio announcers) and (2) three types of Chinese Classics varying in degrees of rhyme regularity from regular, semi-regular to irregular and coded as CL (approximately 3,500 syllables, produced by 1 male and 1 female untrained speakers).

Preprocessing of speech data includes (1) automatically labeled segments followed by manual spot checking for forced alignments and (2) manual tagging of perceived boundary breaks by the HPG specifications (Sec. 3) using the Sinica COSPRO Toolkit [12]. Only consistently tagged data (over 83%) across transcribers are selected for analysis.

## 5. METHODOLOGY

Multi-layered normalization methods are developed to remove attributes that may contribute and affect duration modulations in output speech and to derive duration patterns by discourse units. The general function is listed below.

$$x_i = \mu_i + factor_1 + factor_2 + .... + \varepsilon_i \quad (1)$$

At the word level (as illustrated in Sec. 2), *Factor1* represents segmental information at the Syllable level, *Factor2* represents respective syllable position by word, thus taking into account word-final boundary lengthening, while $\varepsilon_i$ represents all other unpredictable values. The same rationale is applied for further extracting discourse tempo where *factor1*, *factor2*…etc. in turn represents the attributes of prosodic boundary in each layer by the HPG protocol thus removing the effects of boundary lengthening by discourse units [17]. The attributes are removed layer by layer till all possibilities are considered. Residuals at each layer are regarded as contributions from its immediate higher layers and included in the next round of predictions. Predictions of sequential ratio of syllable duration sequences by perceived discourse units the PW and PPh are derived and presented in Sec. 6.

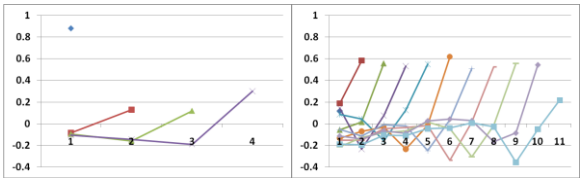
## 6. RESULTS

### 6.1. Duration predictions by PW and PPh without segmental contributions

Predictions by linear regression derived from discourse units the SYL, PW and PPh layers are derived by normalizing the contribution of segments. The predictions at the SYL layer is not presented here due to large amount of classes defined [13]. Figure 2 shows plotting of sequential ratio of syllable duration of PW (1-4 syllables, left panel) and PPh (1-11 syllables, right panel). The

overall PW cadence pattern features pre-B2-boundary lengthening of the final syllable, while the overall PPh cadence pattern features different degree of pre-B3-boundary lengthening of the last two syllables, and shortening of the antepenultimate syllable.

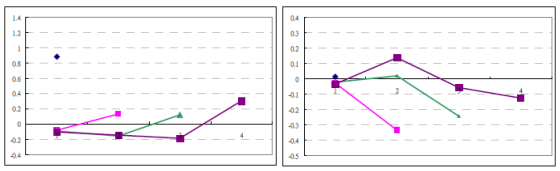
**Figure 2:** Regression coefficients of syllable durations obtained for PW model (left) and PPh model (right). The X-axis represents the position of each syllable within a PW/PPh; the Y-axis represents the coefficient values.



## 6.2. Duration predictions by PW and PPh without boundary effects

The same predictions by linear regression were repeated by normalizing boundary contributions whereby PWs located at pre- and post-B3 boundary positions in a PPh were separated from those without boundary information [11]. Figure 3 shows plotting of sequential ratio of syllable duration of PW (1-4 syllables) with boundary effects (left panel) and without boundary effects (right panel). The derivation after removing boundary effects shows a completely different PW cadence. The results suggest that final syllable lengthening is a boundary effect and should be separate from the PW cadence pattern.

**Figure 3:** The left panel is the derived PW patterns after normalizing the number of phones in each syllable; the right panel is the derived PW pattern after normalizing pre- and post-boundary effects.



The same predictions by linear regression were repeated at the PPh layer. Figure 4 shows plotting of normalized sequential ratio of adjacent PPhs ranging from 6-11 syllables. Instead of considering only one immediate neighboring syllable of each B3, i.e., only one pre- and post-B3 syllable, we define the immediate between-PPh neighborhood as the last 4 syllables of a preceding PPh and the first 3 syllables of the following PPh. With this definition, the PPh neighborhood is defined by units that include the boundary immediate PW rather than single syllables, a definition that better reflects the rationale of the HPG framework. The overall PPh cadence pattern features a wider range of pre-boundary lengthening and post-boundary shortening, illustrating how PPh boundary effects are consistent and systematic regardless of the number of syllables in the PPh.

**Figure 4:** Regression coefficients of syllable durations obtained for PPh model where boundary effects were defined as the last 4 syllables of a preceding PPh and the first 3 syllables of the following PPh.

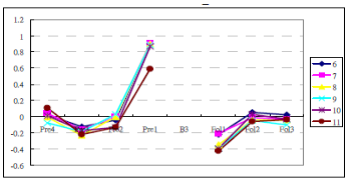
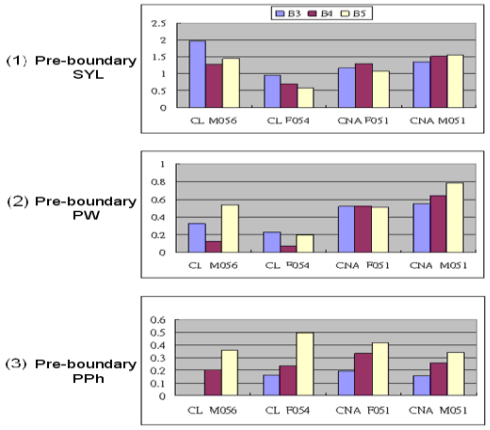


**Figure 5:** Cross boundary comparison of duration patterns by prosodic units syllable (SYL), PW and PPh. The horizontal axis represents indexes of the speech data and speaker. The vertical axis denotes normalized average duration of prosodic units.




## 6.3. Boundary lengthening by PPh

In this section, we will discuss discourse tempo from the perspective of pre-boundary lengthening [15]. We have shown (Sec. 6.2.) that the PPh cadence pattern is relative across the PPh than by the syllable; surely the so-called pre-phrase final syllable lengthening [5] is not sufficient to account for phrase or discourse tempo adjustments. The question is: by what unit is pre-boundary lengthening applied and how systematic? A simple derivation of mean syllable duration by discourse units the Syl, PW and PPh across the same speech data are presented in Figure 5. The results show that both pre-boundary lengthening by the SYL or PW (Figure 5, (1) and (2)) are not consistent while lengthening by the PPh is systematic across the speaker and speech genre (Figure 5).

## 7. DISCUSSION

The results of removing segmental information (Sec. 6.1.) revealed independent cadence patterns by discourse units the PW and PPh, suggesting that discourse prosodic timing can be derived from output speech. The same results also demonstrate that segmental duration at face value is neither prosodic nor phonological. In addition, the results of boundary effects (Sec. 6.2.) showed systematic and predictable cadence patterns by discourse units the PW and PPh also, thus demonstrating how the syllable alone is insufficient to account for output timing while discourse prosodic timing must be taken into account. The same rationale can also be applied to systematic boundary effects as part of discourse prosodic timing as well. In other words, the results of systematic pre-boundary lengthening by the larger and higher-level discourse units PPh (Sec. 6.3) rather than lower-level units the Syl and the PW further suggest that production planning of continuous speech timing include simultaneously scheming of higher-level chunking and phrasing and lower-level articulatory planning. The same results can also be interpreted as the existence of modulations of speaking rate by the words and phrase. In this light, the commonly adopted speaking rate analysis by averaged duration across the board, whether by the vowels, the syllable, or the number of words, is a rough reference [10] that prevents the dynamic prosodic discourse timing to surface. Discourse tempo modulations should be attributed to collective contributions from syllable-timing, discourse timing and boundary effects.

## 8. CONCLUSION

The obtained evidence from the present study demonstrates why the surface dynamic tempo of continuous speech can not be sufficiently accounted for either by fine-grained measurement of segmental duration patterns, word level rhythmic patterns, or averaged duration of one single unit. In addition to segmental duration due to the physical composition, discourse contributions, as seen in derived cadence patterns by discourse units, as well as systematic boundary modulations, must also be attributed in order to account for the surface dynamics. We believe the systematic nature of discourse attributes is why phonological timing from continuous speech could be retrieved. We also suspect that these higher level timing cadence may be cross-linguistic, and merits future studies of languages other than English and Mandarin.

## 9. REFERENCES


[1] Abercrombie, D. 1967. *Elements of General Phonetics.* Edinburgh: Edinburgh University Press.

[2] Cao, J. 2004. Restudy of segmental lengthening in Mandarin Chinese. *Proc. of Speech Prosody* Nara, 231-234.

[3] Chen, J.Y., Chen, T.M., Dell, G.S. 2002. Word-form encoding Mandarin Chinese as assessed by the implicit priming task. *J. Memory and Language* 46(4), 751-781.

[4] Dellwo, V., Wanger, P. 2003. Relations between language rhythm and speech rate. *Proc. ICPhS* Barcelona, 471-474.

[5] Edwards, J., Beckman, M.E. 1987. Perception of final lengthening. *Proc. Annual Meeting of the Linguistic Society of America*.

[6] Gibbon, D., Gut, U. 2001. Measuring speech rhythm. *Proc. Eurospeech* Scandinavia, 95-98.

[7] Lehiste, I. 1971. The timing of utterances and linguistic boundaries. *J. Acoust. Soc. Am.* 51(6B), 2018-2024.

[8] O'Seaghdha, P.G., Chen, J.Y., Chen, T.M. 2010. Proximate units in word production: Phonological encoding begins with syllables in Mandarin Chinese but with segments in English. *Cognition* 115(2), 282-302.

[9] Sagisaka, Y. 2003. Modeling and perception of temporal characteristics in speech. *Proc. ICPhS 2003* Barcelona, 1-6.

[10] Tseng, C. 2010. Beyond sentence prosody. *Proc. Interspeech* Makuhari.

[11] Tseng, C., Chang, C. 2008. Pause or no pause? Prosodic phrase boundaries revisited. *Tsinghua Science and Technology* 13(4), 500-509.

[12] Tseng, C., Cheng, Y., Chang, C. 2005. Sinica COSPRO and Toolkit—Corpora and platform of Mandarin Chinese fluent speech. *Proc. Oriental COCOSDA* Jakarata, 23-28.

[13] Tseng, C., Fu, B. 2005. Duration, intensity and pause predictions in relation to prosody organization. *Proc. Interspeech* Lisbon, 1405-1408.

[14] Tseng, C., Pin, S., Lee, Y., Wang, H., Chen, C. 2005. Fluent speech prosody: Framework and modelling. *Speech Communication* 46(3-4), 284-309.

[15] Tseng, C., Su, Z. 2008. Boundary and lengthening—On relative phonetic information. *Proc. 8th Phonetics Conference of China and the International Symposium on Phonetic Frontiers*, Beijing.

[16] White, L., Mattys, S.L., Series, L., Gage, S. 2007. Rhythm metrics predict rhythmic discrimination. *Proc. ICPhS* Saarbrücken, 1009-1012.

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