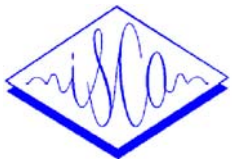

Speech Prosody 2004
Nara, Japan
March 23-26, 2004

Speech Rate and Prosody Units: Evidence of Interaction from Mandarin Chinese

Chiu-yu Tseng and Yeh-lin Lee

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

Abstract

This paper discusses evidence of interaction found between speech rate and prosody units in Mandarin Chinese speech. Mandarin speech data of 2 different speech rates that had been previously labeled for perceived boundaries and prosody units were further analyzed for duration patterns at each prosodic level. Each prosody level demonstrated patterns of duration adjustment for both speech rates that could be accounted for by the model used. These patterns of duration adjustments are clearly systematic, suggesting how each prosody levels may interact and to an extent govern the temporal distribution of units within. Our findings demonstrate that though speech rate may appear to be a global phenomenon across speech flow on the surface, it in fact is very much an in integrated part of prosody organization constrained by each prosody level. To put simply, duration adjustment is being made systematically at each prosody level during speech production instead of just an across-the-board phenomenon. As a result, interactions between prosody units and temporal distribution are predictable. We believe these findings are a step forward in understanding temporal organization and distribution of speech flow as well as speech prosody in general, and should be directly applicable to predicting speech prosody of unlimited TTS in particular.

10.21437/SpeechProsody.2004-58

1. Introduction

How to predict prosody from text and improve output naturalness remains a major bottleneck for unlimited TTS after decades of research efforts. The naturalness issues boils down to insufficient understanding and consequently still somewhat crude implementation of prosodic properties to synthetic speech output. More knowledge of prosody organization with respect to units, boundaries and domain is still lacking in general; more understanding of speech rate with respect to temporal distribution still unclear in particular. Previous researches have examined duration characteristics for both syllable-timing and stress-timing languages [1] [2] [3] [4], making it a necessary distinction in investigating timing related issues. Mandarin Chinese is a syllable-time language whereas temporal variations of the syllable level require more explicit understanding. From the phonetic perspective, segmental durations in connected speech need to be examined with respect to syllables first before moving on to higher and/or larger phonetic, phonological or prosodic units. In this study, we analyzed durational modifications of two speech rates at each prosody level, namely, prosodic words, prosodic phrase, and breath group to see if systematic patterns of temporal adjustment could be found, especially with regard to final lengthening.. A modified multi-layer linear regression model of Keller and Zellner [1] was used to clarify possible influences from the prosodic hierarchy [5] [6], and to test variation tendencies on syllable durations with respect to speech rate. However, other prosodic phenomena such as stress patterns, and phrasal prominence are not included for the time being.

2. Methodology

2.1. Speech Data

Speech corpora of two different reading rates were used, i.e., slower vs. faster speech. The slower speech is from 1 male untrained subject (hence SMS for Slower Male Speech) reading 595 paragraphs ranging from 2 to 180 syllables; the faster speech from 1 female radio announcer's relative faster reading (hence FFS for Faster Female Speech) of 26 long paragraphs ranging from 85 to 981 syllables. A total of 22350 syllables of SMS and 11592 syllables of FFS were analyzed. Average syllable duration is 304.7ms for SMS and 199.75ms for FFS. We performed analyses to (1.) compare duration variations with respect to different speech rate, and (2.) look for how speech rate may interact with prosody units. Both sets of speech data were first labeled automatically using the HTK toolkit and SAMPA-T notations [6]; then labeled for perceived prosodic boundaries by 3 trained transcribers. The HTK labeling was manually spot-checked; the manual perceptual labeling cross checked for intra-transcriber consistency. An in-house ToBI-based system developed for Mandarin [5] [6] was used for prosodic units and boundaries, with emphases on the phrase-grouping characteristic of Mandarin Chinese speech

2.2 Basic Features of Analyses

Using a step-wise regression technique, a linear model with four layers [1] was modified and developed to predict speakers' timing behavior with respect to different speech rate. A hierarchical and hence layered organization of prosody on the basis of boundaries and units was used to classify prosody units at levels of the syllable, prosodic word, prosodic phrase, breath group and prosody group [6]. Moving from the syllable layer in the prosody hierarchy upward to each of the higher prosodic unit and level, we examined each higher layer independently to see if residuals can be explained, and if so, at which level. All of the data was analyzed using DataDesk™ from Data Description, INC. Two benchmark values were used in this study to evaluate the closeness of the predicted value and the original speech data, namely, residual error (R.E.) and correlation coefficient (r). The residual error was defined as the percentage of the sum squared residue (difference between prediction and original value) over the sum squared original value.

3. Results and Analyses

3.1. The Syllable Layer

At this layer, we examined how segmental duration may influence syllable duration, how influences contributed by preceding and following syllables may affect segmental duration, and whether tones may interact with duration as well. Factors considered included 21 consonants, 39 vowels (including diphthongs), and 5 tones (including 4 lexical tones and 1 neutral tone). Classifications of segments were established to help simplify analyses of the speech data. The classifications for the two speech rates varied. Such classification should be useful for future analyses. Tables 1 and 2 showed the results of analyses of FFS.

Type	Consonants	Mean(ms)	Coef Var	Count
C1	d,g,b	20.1065	0.41	2132
C2	l,dz`,f	48.354	0.41	1533
C3	Z`,n,dz,dj,m	66.5804	0.30	2147
C4	t,p,k,h	87.2127	0.31	1420
C5	s`,ts`,sj	106.524	0.23	1864
C6	s,ts,tj	116.968	0.23	830
C7	Zero Initial	0	0	1663

Table1. Types of Consonants of FFS

Type	Vowels	Mean(ms)	CoefVar	Count
V1	@,U`,U	99.5293	0.43	1990
V2	o,u	124.498	0.36	780
V3	i,a	129.767	0.37	1467
V4	yE,ei,y,@n,in,uo,iE	142.158	0.34	1904
V5	ai,ou,uei,@N,oN,iN	149.889	0.30	2253
V6	an,au,yn,iau,aN	157.925	0.27	1323
V7	ia,iou,u@n,@`,iEn,ua	169.913	0.30	1172
V8	uan,yEn,iaN	54.3786	0.30	513
V9	uaN,uai,yoN	58.124	0.30	187

Table2. Types of Vowels of FFS

A Syllable-Layer Model was subsequently postulated as follows:

Dur (ms) = constant + CTy +VTy + Ton
+ PCt + PVt + PrT + FCt + FVt + FlT
+ 2-way factors of each factors above
+ 3-way factors of each syllable +
+ Delta 1

CTy, VTy and Ton represent consonant type, vowel type and tone respectively. Prefix of P and F represent the corresponding factors of the preceding and following syllable. A total of 49 factors were considered. A linear model for discrete data was built using Data Desk with partial sums of squares (type 3). Factors with p-value smaller than 0.5 were excluded from consideration.

Table 3 shows benchmark values of the Syllable-Layer Model found in the two different speech rates. The residue error was 48.9% in SMS and 40.1% in FFS. In other words, the Model explained 51.1% of syllable duration of SMS and 59.6% of FFS at the syllable layer. The residue that cannot be explained at this layer was termed as Delta 1 and will be dealt with at the immediate higher layers.

Test	SMS	FFS
R.E.	48.9%	40.1%
r	0.715	0.768

Table3. Evaluation of Syllable Layer Predictions

3.2. Prosodic-Word (PW) Layer

In this layer immediately above the syllable layer, our aim was to see whether possible effect caused by PW structure on syllable duration could be found. Our hypothesis was that syllable duration is affected by its position within a PW. Therefore, the PW Layer Model can thus be written as follows:

Delta 1 = f(PW length, PW sequence) + Delta 2

Each syllable was labeled with a set of vector value, for example (3, 2) denotes the unit under consideration is the second syllable in a 3-syllable PW. Using identical linear regression techniques as of the preceding layer, the coefficient of each entry was calculated. Figures 1 and 2 illustrate the coefficients of different PW durations. PWs over 5 syllables were not considered due to scarcity of samples.

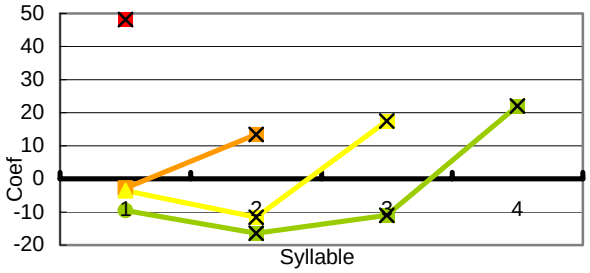

Figure1. Coefficients of SMS from the PW Model. The horizontal axis represents the position of each syllable within a PW; the vertical axis the coefficient values.

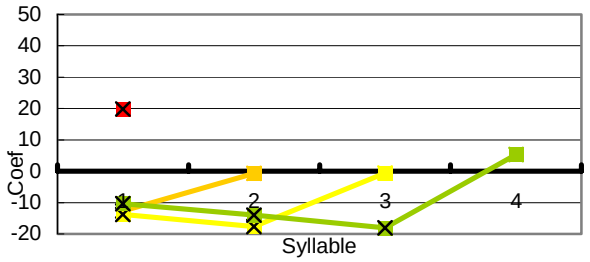

Figure2. Coefficients of FFS from the PW Model

Positive coefficients represent lengthened syllable durations at the PW layer; negative ones shortened syllable durations. Coefficients of p-value smaller than 0.1 were marked with the 'X' label in Figures 1 and 2. Note that several interesting phenomena could be observed: (1.) both speakers exhibit a pattern of PW-final syllable lengthening relative to other syllables considered; (2.) the longer the PW is the greater the duration of the final syllable becomes, and (3.) different speech rate contributed to different degrees of syllable variation. At the PW Layer, SMS showed within-layer syllable shortening but final-syllable lengthening in comparison with lengthening predictions made at the Syllable Layer. However, FFS showed the opposite: That is, while syllables of a PW were shortened as well, the final syllable maintained its prediction of the immediate lower layer. Table 4 shows benchmark values of the PW Model.

Test	SMS	FFS
R.E.	93.3%	96.45%
T.R.E	45.6%	38.76%
r	0.737	0.778

Table4. Evaluation of PW Layer Predictions

The PW Layer model explained 6.7% of Delta 1 of SMS and 3.55% of FFS. The overall prediction can be obtained by adding up the predicted value of both the syllable and PW layers. The Total Residual Error (TRE) is the percentage of sum squared residue over the sum square syllable duration. This result indicates that the residual error ratio cannot be

explained by either layers discussed so far, which we will deal with at the following higher layer(s).

3.3. Prosodic –Phrase (PPh) Layer

The same rationale was applied to this layer. The linear regression model is thus formulated as follows.

Delta 2 = f(PP length, PP sequence) + Delta 3

Figures 3 and 4 illustrate the derived coefficients. Only prosodic phrases with over 60 occurrences were considered for statistical validity. Each line represents a PPh with different durations.

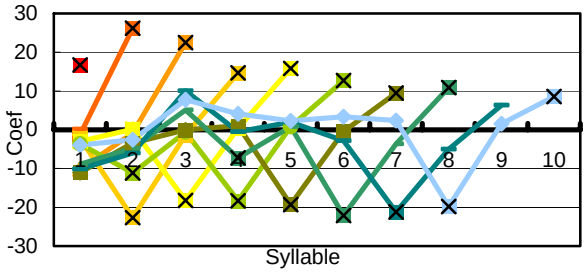

Figure3. Coefficients of SMS from the PPh Model.

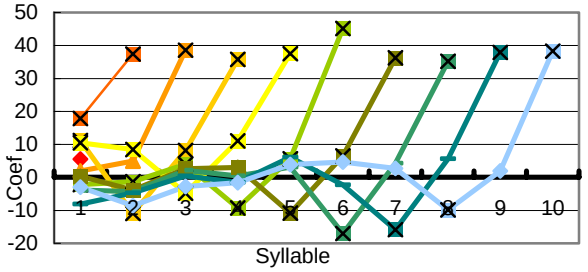

Figure4. Coefficients of FFS from the PPh Model

Figures 3 and 4 showed the following results: (1.) A clear cadence like phenomenon of PPh. (2.) Not only lengthening of the PPh-final syllable was found; shortening was also found at the third syllable counting backward. (3.) Final syllable lengthening at the PPh layer was found to be twice as long for FFS, demonstrating once again the contribution from speech rate and consequently a different pattern of rhythm. (4.) A complementary effect of final syllable lengthening was found between the PW Layer and the current PPh Layer. In other words, whenever the final syllable of a PW is lengthened, the same degree of final syllable lengthening could NOT be found at PPh level. Table 5 shows the evaluation of predictions at the PPh Layer.

Test	SMS	FFS
R.E.	93.0 %	86.5 %
T.R.E	42.4%	33.5%
r	0.760	0.814

Table5. Evaluation of Prosodic Phrase Layer Prediction.

Delta 2 of FFS could be explained only by 13.5% of the data at the current PPh layer, and the correlation coefficient r is 0.814. The remaining residue that cannot be explained is termed as Delta 3, and will be dealt with in the next higher layer.

3.4. Breath-Group (BG) Layer

In order to find how the syllable duration was affected by a BG effect due to breathing and hence longer pause, we further studied the residue from the PPh Layer, i.e., Delta 3 at the BG Layer. It was found that the difference occurred more often at the initial and the final portions of a PPh, while the influences on the initial, middle and final prosodic phrase within a breath group are also different. We postulate that BG poses duration effects on the initial and final portions of each PPh within, but not on the middle portion. Table 6 shows the results of our evaluations.

At the BG layer, delta 3 could be explained by 2.2% in SMS and 5.2% in FFS. The overall prediction correlates with the original corpus at the correlation coefficient r = 0.766 in SMS and 0.825 in FFS, which is an encouraging outcome to the current investigations.

Test	SMS	FFS
R.E.	97.8 %	94.8%
T.R.E	41.52%	31.7%
r	0.766	0.825

Table6. Evaluation of Breath-Group Layer Predictions

The effect from the BG Layer on the immediate lower layer (the PPh) within is shown in Figures 5 and 6. Each figure illustrates the influences on the duration of the PPh under 6 syllables. Influences on the first and the last 3 syllables of PPh over 6 syllables were calculated and shown in purple. Both Figures 5 and 6 show lengthening by 10 to 20ms on the first and last syllable.

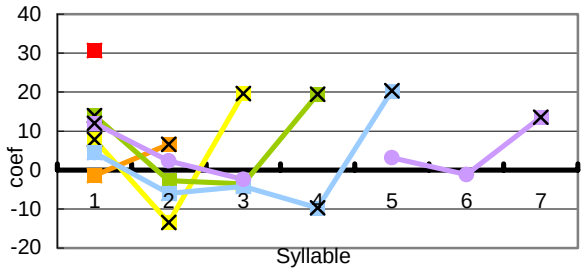

Figure5. Coefficients of SMS from Initial PPh of BG layer Mod**el**

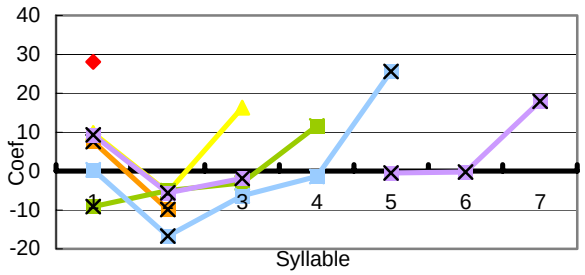

Figure6. Coefficients of FFS from Initial PPh of BG layer Model

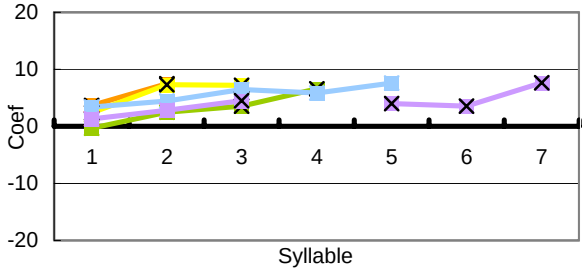

Figure7. Coefficients of SMS from Middle PPh of BG layer Model

Figure 7 and 8 show effects of the BG layer on PPhs that occurred in the middle of a BG. The first syllable was

shortened while the final one is lengthened for BG-middle PPhs considered; the influence is more pronounced in FFS than in SMS.

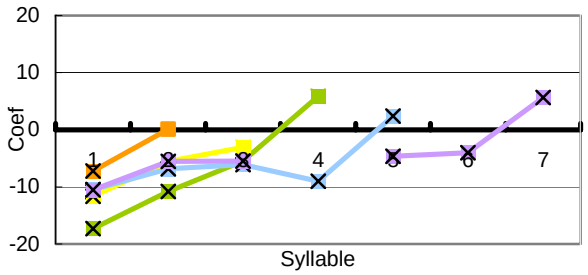

Figure8. Coefficients of FFS from Middle PPh of BG layer Model

Figures 9 and 10 illustrate the coefficients of final PPhs. Contrary to the initial PPhs, the final syllable of the final PPh is shortened. Note that the overall effect of final-syllable lengthening at the BG Layer is still found. The negative coefficients reflect a clear distinction between BG-initial and BG-final prosodic phrases. The observed temporal allocations provide evidence of prosody units and layers as constraints in speech flow.

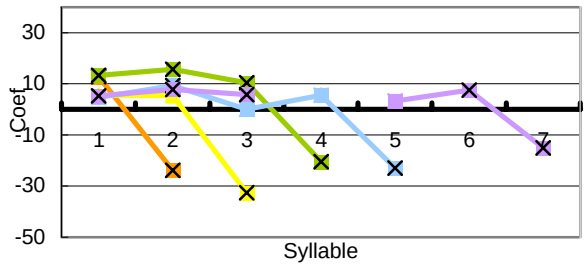

Figure9. Coefficients of SMS from Final PPh of BG layer Model

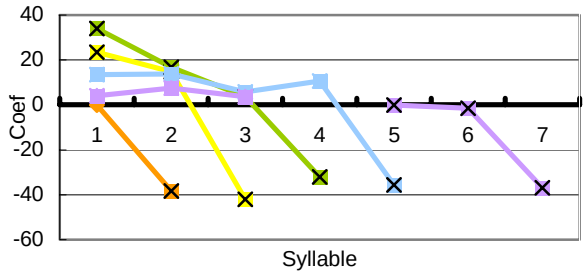

Figure10. Coefficients of FFSS from Final PPh of BG layer Model

To illustrate that the postulated models could predict temporal arrangement with respect to speech rate, we compared the prediction from each prosody layer to a BG of 31 syllables of FFS. Figure 11 show clear interactions of duration variation between prosodic layers; Table 7 evaluation of prediction at each proposed layer.

Layer / Test	Syllable		PW		PP		BG	
	SMS	FFS	SMS	FFS	SMS	FFS	SMS	FFS
R.E (%)	48.9	40.1	93.3	96.45	93.0	86.5	97.8	94.8
T.R.E(%)	N/A	N/A	45.6	38.8	42.4	33.5	41.5	31.7
r	0.715	0.768	0.737	0.778	0.760	0.814	0.766	0.825

Table7 Evaluation on Prediction of Each Layer

4. Discussion

A hierarchical prosody organization was postulated on the basis of prosody units, boundaries and domains with emphases on characterizing phrase-grouping as part of a top-down process. Analyses of speech rate were performed in a bottom-up fashion from syllables upward to various prosodic units to show correlations could be found. Duration adjustments that could not be explained at a lower prosody layer could find answers at higher layers, offering evidences in the following sense: (1.) temporal distributions should be viewed with respect to prosody organizations, (2.)different speech rate may interact with prosody differently, thereby characterizing what speech rate could mean in the physical sense, (3.) trade-off effects were found between prosody levels, and (4.) a hierarchical organization does function during speech production, indicating that a possible optimization schema may very well be in operation during speech production.

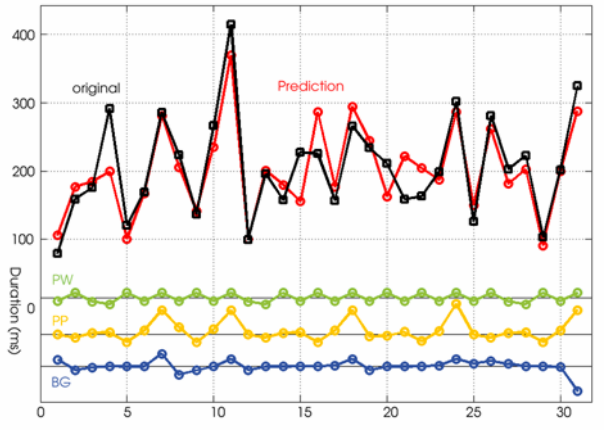

Figure 11.Comparision between speech data and predictions

5. Conclusion

We believe that examining speech rate in relation to prosody units is a significant first step to understanding temporal organization of speech flow, and fundamental to understanding of prosody of speech flow, especially with respect to phrase grouping in connected speech, a linguistic fact that is most prominent in Mandarin speech. Evidences found in the studies above offer possible explanations to prosodic constraints on temporal arrangement, which should also offer some insights to other syllable-based languages as well. Furthermore, we believe our results at this stage are already directly applicable to unlimited TTS of Mandarin Chinese, and should significantly improve output naturalness.

6. References

[1] Keller, E., Zellner Keller, B. “A Timing model for Fast French”, *York Papers in Linguistics*, 17, University of York. 53-75. (1996)

[2] W. N. Campbell, “Speech-Rate Variation and the Prediction of Duration”, Coling 88’, Vol1. (1988)

[3] Zellner Keller B, Keller E., “Representing Speech Rhythm” Improvements in Speech Synthesis. (pp. 154-164). Chichester: John Wiley. (2001)

[4] Chu, M. and Feng, Y., “Study on Factors Influencing Durations of Syllable in Mandarin”, *Proc. Eurospeech 2001*, Scandinavia.(2001)

[5] Tseng, C., “Prosodic Group: Suprasegmental Characteristics of Mandarin Connected Speech from a Speech Data Base”, ICCL-6, Leiden, the Netherlands (1997)

[6] Tseng, C. and F. Chou, “Machine Readable Phonetic Transcription System for Chinese Dialects Spoken in Taiwan”, *JASJp*, (E) 20.3. (1999)