Temporal characteristics of emphasis in continuous speech

Chiu-yu Tseng[1] *and Chao-yu Su*[1,2]

[1] *Phonetics Lab, Institute of Linguistics, Academia Sinica, Taipei, Taiwan*
[2] *Institute of Information System and Applications, NTHU, Taiwan*

cytling@sinica.edu.tw

Abstract

The present study examines how overall tempo adjustment can reflect the allocation of emphasis, whether emphasis is a local prosodic phenomenon, whether the degree of perceived emphasis corresponds systematically to speech signal, and whether temporal features can be derived from production analysis. Results from acoustic analysis showed positive correlations between emphasis in relation to both local and overall tempo modulations; higher degree of emphasis corresponds to overall tempo slowing while effects of phone duration adjustment is independent of segmental make-up. To demonstrate discourse effects on emphasis instead of the other way around, we normalized all possible effects of discourse layering-over and found sharper contrasts between emphasis and non-emphasis. Based on the results we feel it is reasonable to assume that interaction between discourse effects and emphasis is more significant than expected; emphasis is by no means a local prosodic phenomenon.

Index Terms: perceived emphasis, temporal features, overall tempo, discourse structure, normalization, continuous speech

1. Introduction

Speaker produced accentuation, focus or emphasis in speech, perceived as prominence, is one of the major features of expressive prosody. A commonly accepted definition of prominence refers to those words (syllables in Mandarin) that are perceived as standing out from their environment [1, 2]. This definition somehow suggests that emphasis is more of a local prosodic phenomenon. A survey of the literature reveals more reported studies perception studies while relatively less is known from production analysis. We are therefore interested to know if emphasis can be analyzed from production data and whether it is simply a local phenomenon that can be lifted from the speech string and examined in isolation, a regular practice in phonetic investigations. We noted that some previous studies of Mandarin continuous speech on prominence have established that 1) overall tempo adjustments are highly correlated to discourse structure, [3, 4, 5] and 2) identified emphases in the speech signal can by analyzed as an extra layer over discourse structure and triggers interaction [6, 7, 8]. We therefore hypothesize that the bearing prosodic unit is an interacting factor, and emphasis should be examined in relation to broader prosodic context.

The present study aims to examine temporal characteristics of emphasis in continuous Mandarin speech in relation to overall tempo. Specific to the present study are the following questions: 1) whether overall tempo adjustment reflects emphasis allocation in the bearing unit the prosodic phrase, 2) segmental adjustments, if any, are results of emphasis only or combined with discourse information, and 3) whether and how discourse structure interacts with emphasis state.

The paper is organized as follows: Sec. 2 describes speech materials used and annotation rationale. Sec. 3 describes methodology. Sec. 4 presents results including 4.1) relationship between tempo and prominence state and 4.2) discrimination of prominence state by duration distribution. Sec. 6 and 7 are discussion and conclusions.

2. Speech Data and Annotation

2.1. Speech data

We used both read and spontaneous microphone speech for the analyses. Read speech is 1 female's reading of 26 discourse pieces produced in sound proof chambers (45 min/11,600 syllables/85MB, coded CNA) [9]. Spontaneous speech is 1 male's lecture produced in a university classroom (approximately 26 min/7200 syllables/49 MB, coded LEC).

2.2. Preprocessing and annotation

The speech data were tagged in layers. The first layer of tagging is to force aligned segments by the HTK Toolkit; the tagged output was subsequently spot-checked manually by trained transcribers..

2.3. Tagging discourse units and discourse-specified syllable sequence by prosodic layer

Manual tagging of discourse prosodic units by the HPG discourse hierarchy. The perception-based hierarchy specifies the composition of discourse prosody by multiple layers of superimposing that cumulatively contributes to output prosody, whereby contributions could be quantified by layer [4, 5]. Figure 1 is a simplified schematic representation of HPG that shows 5 levels of perceived discourse prosodic boundaries B1 through B5. Prosodic units are defined by corresponding chunks located inside each level of boundary breaks. The HPG prosodic units are the syllable (SYL), the prosodic word (PW), the prosodic phrase (PPh), the breath group (BG, a physio-linguistic unit constrained by change of breath while speaking continuously) and the multiple phrase speech paragraph; SYL/B1<PW/B2<PPh/B3<BG/B4<PG/B5 [10, 11]. Inter-transcriber consistency for prosodic annotation was controlled.

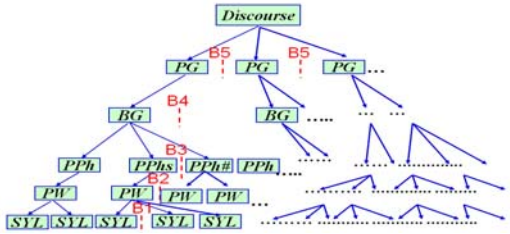

Figure 1: A schematic representation of HPG. The prosodic units from the lowest level are the syllable (SYL), the prosodic word (PW), the prosodic phrase (PPh), the breath group (BG) and the multiple phrase group (PG) or paragraph. The order of syllable strings by HPG layers can also be specified, and at the same time denote the size of respective

discourse units by number of syllables. Table 1 shows an example of tagging the order of syllable sequence by HPG layers PW, PPh and PG. The top panel shows a 8-syllable/character string. At the PW layer, the numbering indicate that there are three prosodic words in 2, 3 and 3 syllables, respectively. At the PPh layer, numbers 1 to 8 indicates that it is an eight-syllable phrase. At the PG layer, the sequence of eight 1's indicate that it is the first phrase of a given speech paragraph. Therefore, the last syllable of the first panel, shown as "菜", is the third syllable in a 3-syllable prosodic word which is the third and last of the three prosodic words in the same prosodic phrase. At the same time, it is also the last and eighth syllable of an 8-syllable prosodic phrase. In turn, the 8-syllalbe PPh is the first phrase in a speech paragraph. The second and third panels present more of the same paragraph. As a result, the numeric tagging by discourse units presents discourse-specified syllable sequence by order of prosodic unit; it makes possible examination of each syllable with reference to its respective position by prosodic layer.

Table 1: A tagging example of speech signal by HPG framework

SYL	'吃'	'了'	'集'	'一'	'口'	'苦'	'心'	'菜'
PW layer	1	2	1	2	3	1	2	3
PPh layer	1	2	3	4	5	6	7	8
PG layer	1	1	1	1	1	1	1	1
SYL	'使'	'我'	'敲'	'起'	'了'	'眉'	'頭'	
PW layer	1	2	1	2	3	1	2	
PPh layer	1	2	3	4	5	6	7	
PG layer	2	2	2	2	2	2	2	
SYL	'討'	'厭'	'的'	'苦'	'心'	'菜'		
PW layer	1	2	3	1	2	3		
PPh layer	1	2	3	4	5	6		
PG layer	3	3	3	3	3	3		

2.4. Manual tagging of perceived emphases

Following the spirit of the HPG framework, perceived emphases in continuous speech is defined by 4 degrees of perceived prominence and tagged manually. They are

- E0-- reduced pitch, lower volume and/or contracted segments
- E1--normal pitch, normal volume and clearly produced segments
- E2--higher pitch, louder volume irrespective of speaker's tone of voice
- E3--higher pitch or louder volume with speaker's change of tone of voice

In other words, E2 usually refers to syntactically defined focus (structural) whereas E3 refers to speaker intended focus (tone of voice). Speech data are manually tagged into a string of emphasis/non-emphasis tokens (ETS) by trained transcribers as an additional layer of prosodic tagging.

3. Methodology

3.2. 3.1.Tempo by ETs

Tempo and duration adjustment are analyzed to examine possible interaction between overall prosodic phrase tempo in relation to prominence allocation, tempo feature by ETs is defined as follows:

$$TP_j = \sum_{i=1}^{M_j} Dur_{ij} / M_j$$

while TP-Tempo, Dur-Duration by syllable
i -Order index of syllables within one ET
j -Order index of ET within one PPh
Mj-Number of syllable within one ET

In order to compare the overall tempo of different size of PPhs, tempo is further normalized by the following equation:

$$NorTP_{jk} = TP_{jk} / \sum_{k=1}^{Nk} TP_{jk}$$

while TP-Tempo, NorTP- Normalized tempo
j-Order index of ET within one PPh
k- Index of PPh
Nk-Number of ET within one PPh

3.3. Refinement of duration features by normalizing effects from multi-layering

In order to test whether the discrimination of emphasis can be improved by duration features irrespective of information from discourse structure, we refine the duration features to further remove possible effects from discourse layering over as defined bellow.

$$ReDur_{ij} = (Dur_{ij} - Mean_j) / STD_j$$

while ReDur-refined duration by phone,
STD- Standard deviation
i- Phone indices in whole speech flow
j- Class index by discourse structure

where j represents PW position, PPh position, PG position, position by multi-layer (PW+PPh+PG) and phone types, respectively. The values of duration features are then divided into 20 bins to represent distribution patterns. All of the refined duration features will be described by distribution/form patterns in the following sections

4. RESULTS

4.1. Relationship between overall tempo and emphasis state

To test whether overall tempo adjustment reflects the allocation of emphasis in the bearing prosodic phrase, tempo by ETs is compared with degrees of emphasis. Results are shown in Figure 2. The left panel shows averaged duration of 2ETs by 3 different clusters of emphasis allocation in read speech CNA, (E1, E2), (E2, E1) and (E3, E1), respectively. Ave is the averaged pattern of 2 ETs with all 3 types of emphasis allocation and represents the base form of PPhs containing 2 ETs. To further remove PPh effects, patterns are derived by subtracting the base form and shown in the right panel of Figure 2.

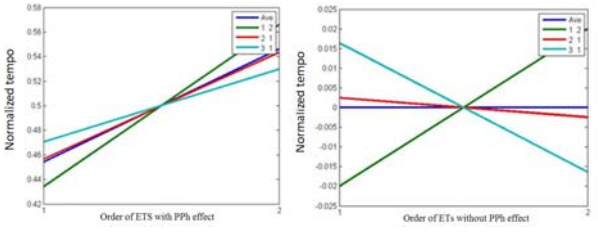

Figure 2: *Patterns of read speech (CAN) tempo by emphasis allocation within 2ETs in PPhs.*

The above results demonstrate that the higher the degree of emphasis the slower the overall tempo is, suggesting that emphasis degree is highly correlated with overall tempo (see figure in the right panel of Figure 2).

Figure 3 presents more detailed analysis of tempo by ETs and speech genre. The patterns of 2, 3 and 4 ETs with all types of emphasis allocation whichever appear more than ten times in spontaneous speech (LEC) and read speech (CNA) are presented in Figure 3.

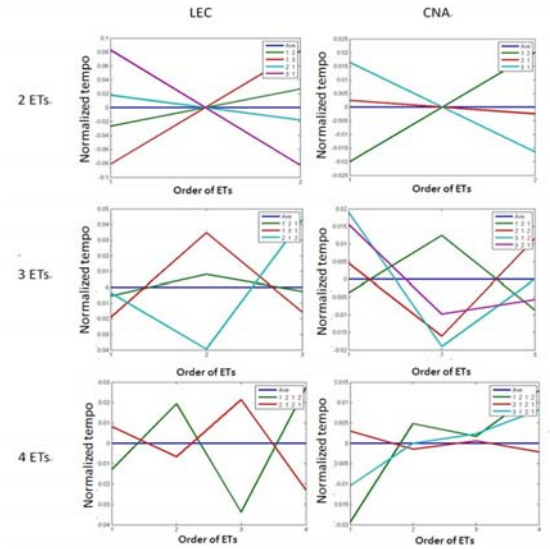

Figure 3: *Pattern of tempo without PPh effect by prominence allocation within 2-, 3- and 4-ETs PPhs in LEC (left) and CAN (right).*

The above results show the same correlation regardless of emphasis number the higher the degree of emphasis, the slower the tempo is. In other words, overall tempo by ETs does reflect the allocation of emphasis in a phrase. Similar patterns are found for both speech genres. The relationship for both spontaneous speech (LEC, left panel) and read speech (CAN) can be described as E1<base form<E2<E3, except the case of 4 collocating ETs in read speech (CAN, lower bottom right panel), E2 and E3 are not necessarily longer than the value of base form.

4.2. Discrimination of emphasis state by duration distribution

In this section, we will discuss how to better examine emphasis state and whether adjustments of phone duration that signal emphasis state are independent of discourse structure. Following the same rationale above, we also refine duration features by removing effects of discourse contribution and intrinsic characteristics of phone types.
Figure 4 presents distribution of raw duration by consonant and vowel as reference for subsequent analysis.

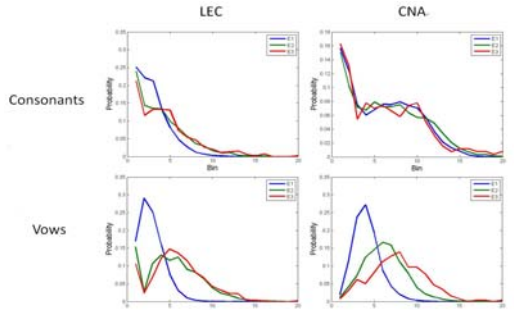

Figure 4: *Distribution of raw duration by consonants/vowels and speech genre.*

The distribution patterns demonstrate that consonant duration between emphasis and non-emphasis is not discriminative, but vowel duration is. In other words, emphasis is positively related to vowel duration. As a result, subsequent refinement of duration features will only apply to vowel duration and discussed below.

4.2.1 Emphasis state by duration distribution after normalizing discourse effect

To test whether emphasis is more of a local phenomenon, we examined the distribution of emphasis state by vowel duration, normalizing individual prosodic layer PW, PPh and PG independently. The results are presented in Figure 5.

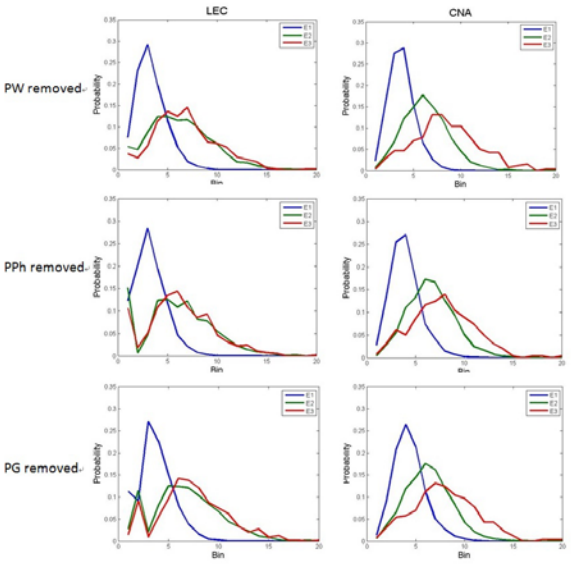

Figure 5: *Duration distribution by normalizing prosodic layer, PW, PPh and PG independently*

No significant improvement of discrimination between emphasis and non-emphasis is found when distribution from an individual prosodic layer is normalized. In fact, the distribution of vowel duration is similar to the distribution of raw duration (See Figure 4). One way to interpret the results would be that discourse structure poses no effect to emphasis state; emphasis is local. However, it has been reported that contributions from discourse layers are cumulative [4, 5, 6], thus we further normalized 3 prosodic layers the PW, PPh and PG all together to see if cumulative contributions from

discourse layers are in the signal. The results are presented in Figure 6.

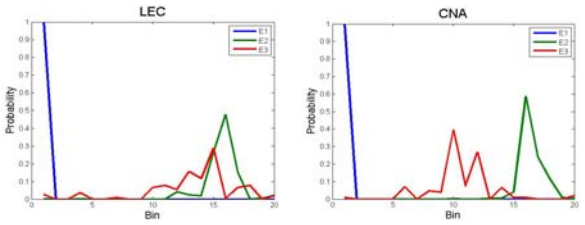

Figure 6: Duration distribution by normalizing cumulative effect of 3 prosodic layers, PW, PPh and PG.

The results in Figure 6 show that when all possible contribution from discourse information to vowel duration adjustments is removed, discrimination between emphasis and non-emphasis is significantly enhanced. That is, the duration distribution of E2 and E3 are clearly distinguished from E1. Therefore, it is reasonable to assume that emphasis is not a local phenomenon by itself, but rather, an interacting factor with discourse structure.

4.2.2. Emphasis state by duration distribution after normalizing phone effect

Another perspective to examine emphasis related duration pattern is to test the effects of duration contribution from phones in syllables. Figure 7 is the distribution of vowel duration by normalizing intrinsic characteristics by vowel types.

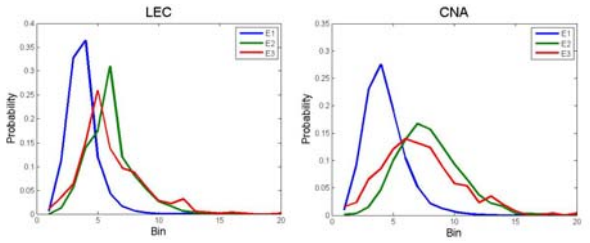

Figure 7: Duration distribution by normalizing intrinsic characteristics by phone types.

The results show no significant improvement for discrimination between prominence and non-prominence after removing effect of vowel types; the patterns are similar to the distribution of raw duration (see Figure 4). The results suggest that emphasis state has little to do with the vowel type in the syllable, and further support the fact that the nature of emphasizing is superimposed and suprasegmental in nature.

5. Discussion and Conclusions

The above analyses suggest that in both read and spontaneous speech overall tempo adjustment can reflect the allocation of emphasis, while emphasis is by no means a local prosodic phenomenon. The degree of emphasis is positively correlated to overall tempo modulations; higher degree of emphasis corresponds to overall tempo slowing (in average duration shown in Sec. 4.1). By normalizing different types and effects of discourse layering-over to refine duration representation, discrimination between emphasis and non-emphasis is enhanced, thus proving considerable contributions from discourse/global information to emphasized local words. The results further show that phone types have little to do with emphasis discrimination (Sec. 4.2.2), thus proving that emphasis caused duration adjustment is independent of segmental make-up of emphasized words. Therefore, it is reasonable to assume that interaction between discourse effects and emphasis is more significant than interaction with intrinsic (physical) characteristics of phones. We believe the results presented in the present study sheds new lights to further our understanding of emphasis in prosody analysis. Future work will be twofold. On the linguistic side, we will continue to see how degrees of emphasis may corresponds to information weighting and in relation to post-focus compression. On the application side, we will focus possible application of derived acoustic patterns to ASR or SDR(spoken document retrieval).

6. References

[1] Terken, J., 1991, "Fundamental frequency and perceived prominence of accented syllables", *Journal of the Acoustical Society of America* 89: 1768-1776,

[2] Ladd, D. J., 1996, *Intonational Phonology*, Cambridge University Press,

[3] Cao, J., 2004. "Restudy of segmental Lengthening in Mandarin Chinese." *Proc. of Speech Prosody 2004*, Nara, Japan.

[4] 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.

[5] 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.

[6] Tseng, C., Su, C., and Huang, C., 2011. "Prosodic Highlights in Mandarin Continuous Speech — Cross-Genre Attributes and Implications". *The 12th Annual Conference of the International Speech Communication Association*. Florence, Italy.

[7] Cambier-Langeveld, T., 1999. "The interaction between final lengthening and accentual lengthening: Dutch versus English." *ICPhS99`*, San Francisco, 467-470.

[8] Edwards, J. & Beckman, Mary E., 1988. "Articulatory timing and the prosodic interpretation of syllable duration." *Phonetica*, 45, 156-174.

[9] Tseng, C., Cheng, Y., and Chang, C., 2005. "Sinica COSPRO and Toolkit—Corpora and Platform of Mandarin Chinese Fluent Speech," Oriental COCOSDA 2005, Jakarata, Indonesia,

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

[11] 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.